Predictions for Biomedical Decision Support

Xiaoqian Jiang

CMU-ISR-10-128

December 2010

Institute for Software Research

School of Computer Science

Carnegie Mellon University

Pittsburgh, PA 15213

Thesis Committee:

Dr. Latanya Sweeney, Chair

Dr. Kathleen M. Carley

Dr. Carolyn P. Rose

Dr. Lucila Ohno-Machado

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy.

This work is funded in part by the National Library of Medicine (R01LM009520), National Library of Medicine (R01LM009018), NHLBI (U54HL108460), the Komen Foundation (FAS0703850), and Harvard University Special Fund.

Keywords: discrimination, calibration, AUC, Hosmer-Lemeshow test, isotonic regression, Platt scaling, reliability diagram, adaptive learning, structured learning, maximum margin optimization, convex optimizaton, markov network, conditional random fields, time series regression, Hidden Markov Models

Abstract

Medications designed for a general population do not work the same for each individual. Similarly, patterns observed from naturally occurring disease outbreaks do not necessarily describe outbreaks of purposeful disease outbreaks (e.g. bioterrorism). To tackle challenges posed by individual differences, my thesis introduces data-driven paradigms that predict a particular case will have the outcome of interest. My insight is to accommodate individual differences by coherently leveraging information from complementary perspectives (e.g., temporal dependency, relational correlation, feature similarity, and estimation uncertainty) to provide more reliable predictions than possible with existing cohort-based approaches.

Specifically, I carefully investigated two representative problems, bioterrorism-related disease outbreak and personalized clinical decision support, for which previous research does not provide satisfactory solutions. I developed a Temporal Maximum Margin Markov Network framework to consider the temporal correlation concurrently with relational dependency in bioterrorism-related diseases' outbreaks. This framework reduces the ambiguity in estimating outcome variables from noisy manifestations by considering complementary information. It outperformed state-of the-art models with synthetic and real world datasets, and improved average state prediction accuracy in predicting simulated biohazards. Regarding personalized clinical decision support, I focused on an important but little-studied measurement "calibration," which stratifies how outcomes affect various genetic population groups within a patient-diagnosis population. I designed joint optimization framework to combine discrimination and calibration, and demonstrated models (DP-SVM, SIO and AC-LR) developed under this multi-targeted framework perform better on both metrics than single-targeted models. I conducted various real data experiments including Hospital Discharge Error, Myocardial Infarction and Breast Cancer Gene Expression Data to verify the efficacy of my joint optimization framework.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Modern biology and medicine are strongly influenced by advancements in computer science, which have significantly increased computational power and improved the feasibility of collecting data [158, 167, 169]. The merging of computer science and biomedicine makes biomedical informatics a complex interdisciplinary research field [102, 122, 126, 144, 167, 191], which often involves analyzing complex, mutually coupled and intensive health-related data.

Along with these advancements, many challenges have arisen [20, 95, 102, 105, 119, 126, 191]. For example in many biomedical applications, it is difficult to access the absolute risk of adverse events in a timely manner. This is largely because of the discrepancy between the huge amount of information and human experts' limited time to review it [29, 127, 143, 199]. Thus, automated tools that reliably predict outcomes are highly appreciated [74, 85, 115, 173, 193, 196, 201], as foreseeing potential emergencies obviously benefit decision makers for public policy and patient care services from early estimation of the risk of adverse events to reach informed decisions. For example, if primary care providers have enough awareness, they could reduce prescription errors occurring during hospital discharge, and saving patients from temporary harm or hospitalization [2, 28, 86]. Similarly, policy makers can respond to emergencies, such as a bioterrorism attack, more effectively with the help of prediction systems that forecast the attack and estimate its impact [131, 154, 190].

However, predictions for biomedical decision support are non-trivial tasks because they have to accommodate individual differences in an abundance of observations. For example, medications designed for a general population do not work the same for each individual. Similarly, patterns observed from nat-

urally occurring disease outbreaks do not necessarily describe outbreaks of purposeful disease outbreaks (e.g. bioterrorism). To tackle these challenges, my thesis introduces data-driven paradigms that predict a particular case will have the outcome of interest. My insight is to accommodate individual differences by coherently leveraging information from complementary perspectives (e.g., temporal dependency, relational correlation, feature similarity, and estimation uncertainty) to provide more reliable predictions than possible with existing cohort-based approaches.

Specifically, I carefully investigated bioterrorism-related disease outbreaks and personalized clinical decision support, for which previous research does not provide satisfactory solutions. Figure 1.1 illustrates both problems. Due to differences in their backgrounds, the manifestations of these prediction problems are not the same but both ones need to be examined in a fine granularity to accommodate individual differences for the best performance.

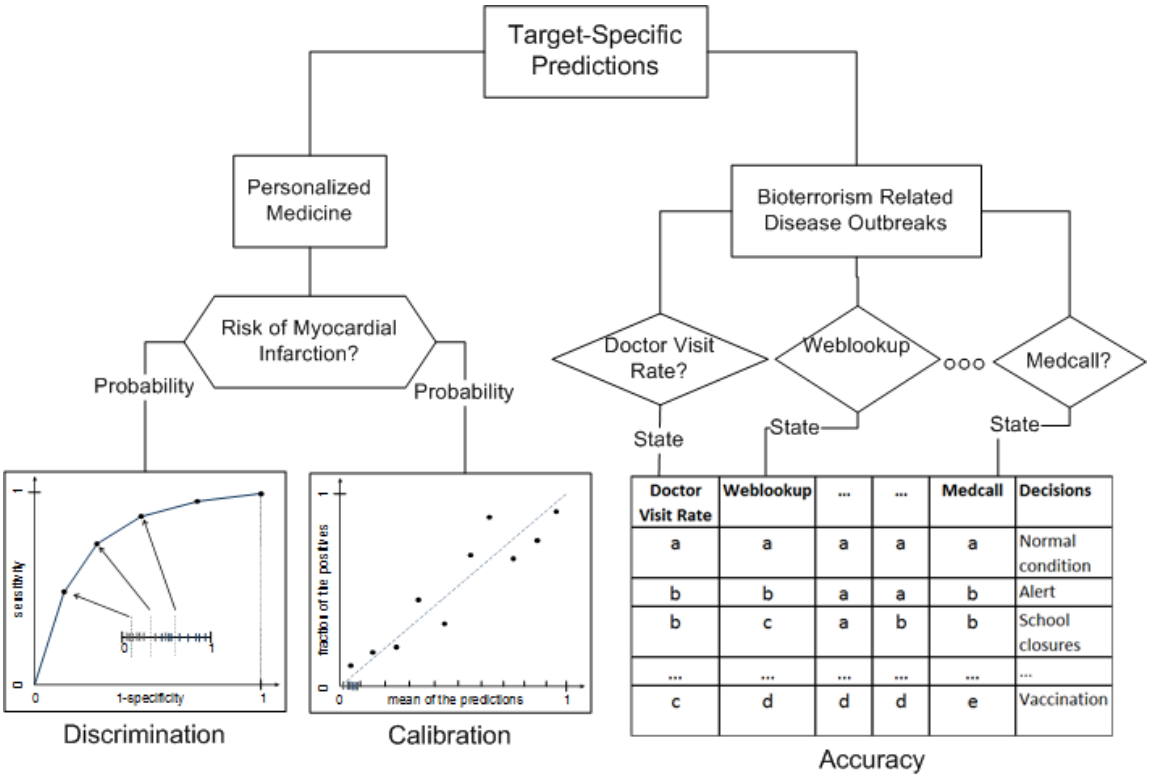


Figure 1.1: Biomedical decision support needs target-specific predictions.

I developed a structured predictive framework, the Temporal Maximum Margin Markov Network (TM3N), to co-estimate multiple correlated variables. As opposed to traditional approaches that predict outcome vari-

ables' states independently, my framework considers semantic correlations of heterogeneous variables and observations of individual variables globally. The global optimization strategy reduces the ambiguity in estimating outcome variables, which is otherwise difficult to handle if outcome variables are considered independently. Thus, the TM3N framework can learn from temporal and relational correlated noisy data, automatically adapt to new conditions, and learn correlated hidden states simultaneously with tractable inference.

Regarding personalized clinical decision support, I focused on *calibrating* biomedical decision systems to identify the parameters of unique genetic populations. Previous theories were based on a diagnosis population; that is, if a patient has a disease, the predictive model is constructed for various treatments of patients with the same diagnosis. We know now, however, that medication that is right for one member of this diagnosis population might not work the same for all members. I believe that predictions should be tailored and targeted to benefit patients in specific genetic groups based on detailed patient information. Toward this end, I studied an important but little-studied quality measurement for probabilistic predictive models, called "*calibration*," which stratifies how outcomes affect various genetic population groups within a patient-diagnosis population. I designed a unified framework that combines two families of model quality measurements, *discrimination* and *calibration*. I demonstrated that models developed under this multi-targeted framework perform better on both quality measurements than single-targeted models.

To demonstrate the effectiveness of the methods developed in this thesis, I conducted experiments using public data from UCI's Machine Learning Repository and clinically related datasets. These clinical datasets are: 1. BioWar Simulation data, which were obtained from the BioWar simulation engine developed by the CASOS lab at Carnegie Mellon University, and which contain aggregated demographics and medical-related information at city level; 2. breast cancer gene expression data, which contain the gene expressions of 541 patients, each with a total feature size of 247,965; 3. myocardial infarction data, which correspond to patients' records collected from emergency rooms in Sheffield, England and Edinburgh, Scotland (the total number of patients' records is 1,853 and their feature size is 48); and 4. hospital discharge error data, which contain 3,833 patients' records with demographic and cultural information collected in a retrospective study of a teaching hospital in Boston, MA.

Given the focus of the thesis, I expect readers from both machine learning and biomedical informatics backgrounds will find it useful. In order to accommodate different readers, I decided to utilize both mathematical models and their intuition. I explain why the problems are important, how my models are

developed, and the evaluation criteria used. To this end, I inserted "interpretation" below the equations to assist biomedical readers in understanding the formulations aimed at machine learning readers.

1.1 Overview

Predictive models are important for risk assessment in biomedical informatics [102, 104, 122, 126, 144, 194]. In many cases, policy makers or care providers can think carefully about the consequences of their decisions, and thus act effectively, when they are better informed. For example, the detection of early stage disease could make a big difference to the treatment of patients [13, 29, 85, 126, 143, 167, 209]; similarly, the prediction of major catastrophes may save thousands of lives in a bioterrorism attack [24, 46, 47, 114, 131, 137, 154].

One type of the diagnosis error, the failure to follow up on test results in a timely manner, can lead to significant delays in diagnosis and treatment, resulting in patient morbidity and mortality [2, 28, 86]. In healthcare settings with well-developed computerized information systems, a post-discharge test result follow-up process could be automated and tracked. However, many hospitals and clinics in the U.S. have not adopted electronic health records. While laboratory systems can often identify culture results that show the growth of an organism, these results are usually linked to the ordering provider. By the time the results return, the provider responsible for the patient may have changed, or the patient may have left the hospital. In these scenarios, the test results pending at the time of discharge from the hospital often need to be followed up manually in order to reduce errors and improve patient care. However, it is not possible to conduct follow-ups with everyone due to insufficient staff time and resources. Thus, a practical approach to reduce hospital discharge error is to track patients based on their risks from high to low, which can be estimated by dedicated predictive models.

Similar systems could also assist in another clinical problem, that of providing preventative treatment to myocardial infarction patients [53, 157, 183, 188, 209]. It is important for caregivers to estimate the risk of myocardial infarction in patients because the risk factor assessment requires a substantial amount of staff time and expensive lab tests. Indeed, not every patient would be appropriate for the preventive treatment of the condition. Thus, to better utilize the limited resources, it would be beneficial for caregivers to screen and prioritize the assessment of patients most likely to benefit from the treatment. The estimation is therefore critical, and allows caregivers to rank their parents and select those for whom the preventative treatment will

be suitable.

The predictive power of state-of-the-art machine learners could also address the problems related to a potential bioterrorism response [19, 24, 46, 114, 154]. In the event of bioterrorism attacks, hospital and clinical services might receive unreliable local claims, so their collaboration with public safety organizations on information aggregation and interpretation is crucial to the success of their response [131, 137]. However, due to the gap between explosive information and the little time available for responding to the situation, human experts cannot grasp the big picture in a timely manner, and their responses are limited by possessing only partial knowledge of the situation. Obviously, these supports are not sufficient to make an informed decision. The success of hazard prevention and control hinges on the ability of synthesizing locally information quickly, and foreseeing the consequences of the decision [131]. Ideally, if an advanced predictive model reliably predicts the probability and trends of an outbreak, healthcare facilities can implement prevention and control measures rapidly, and policy makers can establish appropriate interventions such as resource allocation, vaccinations, quarantines, and school closures. Furthermore, the government can then activate the network of communication effectively. The implementation of these vital emergency response procedures relies on efficient and accurate predictive models.

The situations outlined above show that the predictive model could improve the decision-making process. Recent breakthroughs in clinical research have shown that applying machine learning techniques can assist physicians in diagnosing and treating conditions [15, 49, 64, 74, 85, 115, 173, 192, 193, 196]. However, developing predictive models for the biomedical decision-making process is different [13, 15] due to its focus on improving the prediction reliability of individuals, and difficulties associated with biased labeling [21, 138] and mutual coupling of outcome variables [133]. For example, it is very common in biomedical applications to use only a small number of positive labeled cases, and large numbers of unlabeled cases, to predict a future event. Multiple correlated factors must also be taken into account simultaneously over time to compensate for the ambiguity of the observed information. These differences mean that traditional machine learning techniques developed based on theories from cohort studies have been difficult to apply. To this end, I will present new ideas to close the gap and develop methods to address these real world challenges.

1.2 Outline of Thesis

My thesis is organized as follows:

- Chapter 2 describes the biomedical data that motivated my research and used to validate the predictive methods I developed in this thesis.
- Chapter 3 provides an background overview of relevant research in structured model and two major families of probabilistic model evaluation metrics.
- Chapter 4 focuses on bioterrorism related diseases outbreak prediction and introduces a novel approach, Temporal Maximum Margin Markov Network, to estimate temporal and relational correlated hidden states.
- Chapter 5 recaps relationship between *discrimination* and *calibration*, and develops a unified framework that enables global considerations of both objective. I also implemented a model called Doubly Penalized Support Vector Machine (DP-SVM) based on this framework to demonstrate the performance improvement due to joint optimization.
- Chapter 6 and 7 develops advanced methods which extend the framework in Chapter 5 to further improve the *discrimination* and *calibration* abilities of probabilistic models.
- Chapter 8 discusses the data scalability impact to model's performance. I compare multiple approaches including those developed in my thesis and methods that are popular. Using various clinical related data, I demonstrated that methods developed in this thesis adapted as well as, if not better than, state of the art approaches at various scales of the training set.
- Chapter 9 analyzes another common but important problem in biomedical informatics: biased labeling. I develop a Structured Biased Support Vector Machine (SB-SVM) approach to handle this issue and demonstrate its efficacy on synthetic and biomedical datasets.
- Chapter 10 demonstrates the cross-data applicability of developed methods. I also introduces a web platform, WEBCALIBSIS, that I developed as a first step to access probabilistic model ubiquitously.
- Chapter 11 reviews this thesis and discusses its contributions, limitations and future works.

1.3 High Level Results

This thesis aims to develop predictive models that would support biomedical decision-making process. Specifically, I investigated two representative biomedical problems: bioterrorism related outbreak prediction, and calibration models for personalized medicine. Although goals of, and motivations behind, these problems differ, they encounter similar difficulties in modeling more detailed information on an individual level. It is widely recognized that medication right for one member of a diagnosis population may not be generally appropriate. Similarly, disease patterns observed over the past decades might not be able to account for an unexpected bioterrorism related outbreak. Unfortunately, traditional approaches developed under cohort study theory tend to ignore individual characteristics. They are also insufficient for effective capture of temporal and relational dependencies from noisy observations.

To tackle these problems, I developed a range of data driven approaches based on "tailored" information of targeted factors. With more detailed global models, I can maximize the likelihood of observations and provide more reliable estimates of latent factors of interest.

Regarding my first motivational problem, bioterrorism related outbreak prediction, I designed an encoding technique (SAX+) that maps numerical measurements of different manifestations to human interpretable states, and have developed approaches to model temporal and relational factors concurrently, combining these techniques in a global optimization framework – Temporal Maximum Margin Markov Network (TM3N). In other words, instead of predicting individual factors like "death rate" in the next time tick, my approach predicts a network of outcome variables considering their mutual dependencies over time. The complementary nature of temporal and relational information helps TM3N to achieve better accuracy and reliability. To verify the efficacy of my approach, I used synthetic data generated from a linear dynamic system, where synthesized temporal and relational factors are controlled by a trade-off parameter α. TM3N demonstrated superior performance at various levels of the trade-off parameter from 0.1 - 0.8. The results confirmed that combining complementary information helps to reduce the ambiguity exists in any single perspective. I then applied TM3N models to predict multiple states of correlated outcome variables in the BioWar simulation data. TM3N led the accuracy (69%) followed by Hidden Markov Model (65%), Maximum Margin Markov Network (58%) and Conditional Random Fields (57%). The performance advantage of TM3N and its applicability are confirmed by the results of another real world experiment for building occupancy detection. Again, TM3N outperformed Hidden Markov Model, Maximum Margin Markov Network

and Conditional Random Fields and their average accuracies are 70%, 37%, 49% and 50%, respectively. Finally, I compared TM3N with Hidden Markov Model, Maximum Margin Markov Network and Conditional Random Fields using the BioWar-II data, which contains multiple five-year simulation periods of various sized agents ($10\% - 100\%$). The results showed TM3N scales well at increasing amount of training data and outperformed the other models.

For the second problem, personalized medicine relies on reliable prediction on individual's risk of getting sick [13, 16, 87, 112, 145, 148, 155, 194]. I focused on calibrating biomedical decision systems to identify the parameters of unique genetic populations. I studied an important but less-studied quality measurement for probabilistic predictive models, called "*calibration*", which stratifies how outcomes affect various genetic population groups within a patient-diagnosis population. I investigated the relationship between two major components of model quality (i.e. *discrimination* and *calibration*) and showed that considering *calibration* concurrently with *discrimination* can improve conventional single-target probabilistic models. Under a unified framework for considering both metrics, I implemented Smooth Isotonic Regression (SIO) and Adaptively Calibration for Logistic Regression (AC-LR). The SIO method introduced a smooth projection function to alleviate the problem of overfitting in Isotonic Regression, which is a state of the art *calibration* model. The AC-LR approach pushed the boundaries further with adaptive binning based on input-specific information. To verify the usefulness of both approaches, I compared them with a popular probabilistic model, Logistic Regression (LR) and existing *calibration* methods like Platt Scaling for Logistic Regression (LR-PS) and Isotonic Regression for Logistic Regression (LR-IS). The experiments using synthetic data showed that SIO has superior *calibration* ability without decreasing the *discrimination* power. The real data experiments for SIO used a set of eight different data including Breast Cancer Gene Expression, Hospital Discharge Error and Pima Indian Diabetes. In general, SIO model demonstrated better *calibration* performance comparing to LR, LR-PS and LR-IR under Hosmer-Lemeshow goodness-of-fit test.

The efficacy of AC-LR was verified in a similar way. I showed intuitively how AC-LR is superior to existing approaches using made up examples, in which I visualized 1D and 2D non-linear separable cases, which be handled by AC-LR but not the others. Then I conducted real data experiments using Hospital Discharge Error, Myocardial Infarction and Breast Cancer Gene Expression Data. In Hospital Discharge Experiment, AC-LR passed HL-test at 0.05 significance level with a p-value of 0.349 while all the other methods failed. In addition, AC-LR even improved AUC from 0.704 (the best of previous approaches) to 0.717 showing joint optimization of *calibration* and *discrimination* improved single-target models in both

perspectives. For the Myocardial Infarction dataset, AC-LR demonstrated its performance advantage over conventional methods again. AC-LR passed HL-test at 0.05 significance level with p-values of 0.645 and 0.246 for Sheffield data and Edinburgh data while LR, LR-PS and LR-IS failed. Improvements for *discrimination* are also prominent, AC-LR achieved an AUC of 0.880 and 0.863 comparing to 0.876 and 0.845 of LR, LR-PS and LR-IS for Sheffield data and Edinburgh data, respectively. Similarly, the performance of AC-LR led the competition of *discrimination* and *calibration* in the Breast Cancer Gene Expression data.

Besides these general evaluation, I also looked at data scalability impacts, biased labeling influences and model applicability issues. I used data from various sources and divided training and testing sets with different ratios to evaluate models developed in this thesis. Chapter 8, 9 and 10 compared different models under various situations and demonstrated performance advantages of methods developed in this thesis.

Chapter 2

Data Description

My research focused on real world biomedical challenges: bioterrorism related disease outbreaks and *calibration* for personalized clinical decision support. The following datasets corresponding to these problems essentially motivated my research in this area. I used several other datasets to verify methods developed in later chapters but they are not as relevant as the four described in this chapter [1]. Specifically, the first related to bioterrorism related disease outbreaks while the others are related to personalized clinical decision support. The following sections describe these data.

2.1 BioWar-I, II Data

These datasets, including BioWar-I and BioWar-II, were generated using the BioWar simulation engine, developed by CASOS lab at CMU (`www.casos.cs.cmu.edu/projects/biowar/`).

BioWar is a single integrated engine that simulates the impact of a bio-terrorist attack in a U. S. city. The model combines state-of-the-art computational models of social networks, communication media, and disease transmission with demographically resolved agent models, urban spatial models, weather models, and a diagnostic error model; unlike traditional models that look at hypothetical cities, BioWar is configured to represent real cities by incorporating census data, school district boundaries, and other publicly available information [31]. This engine focuses on bio-terrorist attacks, but its structure is applicable to emergent and familiar diseases as well.

[1] All the data for my thesis is accessible from the following URL: `http://scholar.privacy.cs.cmu.edu/thesis/data/`

Although BioWar-I and BioWar-II are both simulated data, their reliability and faithfulness are well acknowledged by a number of scientific publications [30, 31, 32, 36, 38]. Because no bio-terrorism related attacks have been reported in U.S., I decided to treat BioWar-I and BioWar-II as my pseudo-truth for the model construction. The simulated data consist of the following city-scale bio-attacks. 1) BioWar-I data consist of a one-year-period of observations in the city of Pittsburgh, PA, from 9/1/2002 to 8/31/2003. The total number of simulated agents are 306,181. There was one outbreak of airborne diseases (avian influenza) during the simulation period. 2) BioWar-II data contain multiple five-year-period observations from 9/1/2002 to 8/30/2007. The number of simulated agents are set to vary from 153,090 to 1,224,726 at an approximately equal scale ($150k$); namely, the number of simulated agents varies from 10% (153,090) to 100% (1,224,726). The city of simulation is Norfolk, VA. There was one outbreak of airborne diseases for every year during the simulated period.

For both data, the simulated agents interact and transmit airborne diseases (avian influenza) over time. The statistics of activity are aggregated in a window of every 4 hours and thus there are six time ticks everyday, that is, $365 * 6$ time ticks for each year. For BioWar II, I requested that simulations be performed on ten five-year periods rather than one single 50-year period to avoid factors like birth and aging that would interfere with the disease impact on the mortality, work absence and doctor visit rates. Table 2.1 summarizes BioWar-I inputs (one-year-period of Pittsburgh), because BioWar II inputs are similar, they are not shown.

Table 2.1: Summary of the "BioWar-I" data. Data were generated by the BioWar simulation engine.

File	Summary
Activity.csv	The activities of 79,497 kids and 226,684 adults within a period of 2,189 time ticks, each time tick corresponds to 4 hours.
Actual_incidence.csv	The incidence of 52 diseases is the rate at which new cases occur in a population during a specified period.
Actual_symptom_incidence.csv	The prevalence of 52 diseases is the proportion of a population that are cases at a point in time.
Deaths.csv	Actual death of the population at time ticks, reported in every 4 hours.
Deaths_day.csv	Actual death of the population aggregated over days.
Infected_agents_sample.csv	Ids of the infected agents and their corresponding ailments over time.
Social_network_sample.tsv	A snapshot of the social network indicating the agent social relationships.

An important mechanism in BioWar simulation is utilizing social network information to estimate agents' activities and their disease spreading pattern. This mechanism provides "autonomy" to simulated agents and

allows more flexibility in using the model. The mortality and demographics of the infected population are obtained by simulating the behaviors of agents that interact and transmit diseases.

Table 2.2: Outcome variables in BioWar-I and BioWar-II datasets.

adults_at_home	is-restaurant	gender	is-home
kids_at_home	is-doctor	death rate	is-theater
at-work	is-university	insurance	demographics
Weblookup	is-pharmacy	emergency visits	in-hospital
medcalls	is-stadium	Doctor visit	is-work
adults-at-home	is-store	kids-at-home	num-exchanges

Table 2.3: Summary of the co-variates and output variables for BioWar-I data.

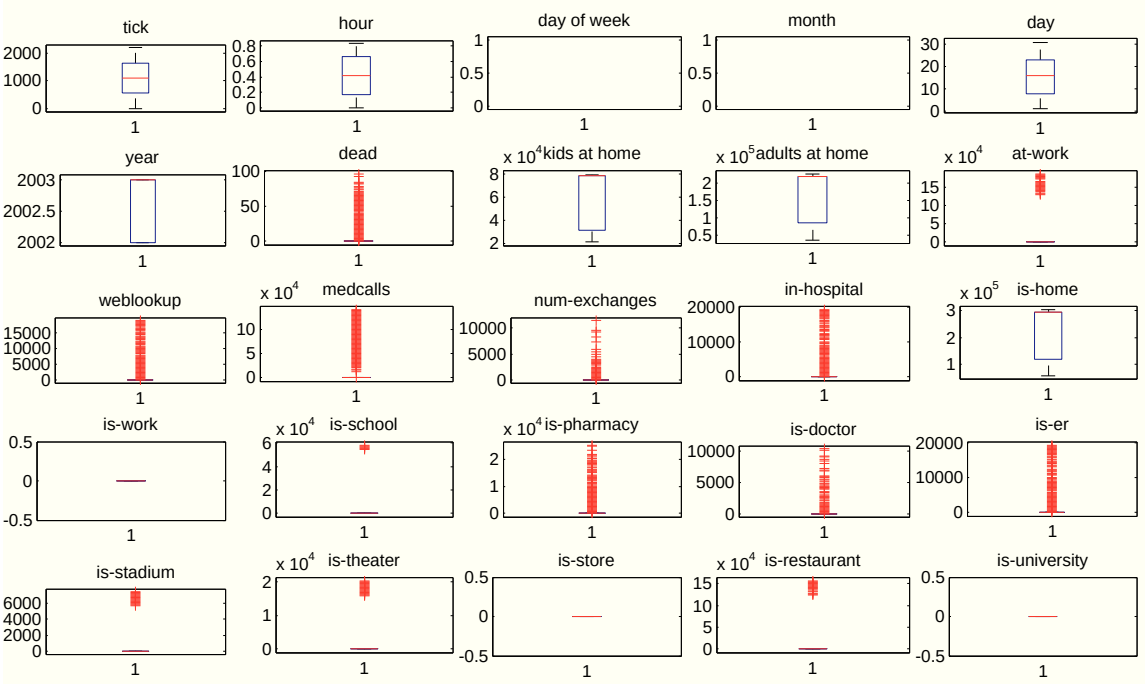
tick	dayOfWeek	month	day	dead	is.er
Min. : 0.0	Fri:312	Aug : 186	Min. : 1.00	Min. : 0.000	Min. : 0
1st Qu.: 547.2	Mon:312	Dec : 186	1st Qu.: 8.00	1st Qu.: 0.000	1st Qu.: 0
Median :1094.5	Sat:312	Jan : 186	Median :16.00	Median : 0.000	Median : 7
Mean :1094.5	Sun:318	Jul : 186	Mean :15.72	Mean : 4.338	Mean : 696
3rd Qu.:1641.8	Thu:312	Mar : 186	3rd Qu.:23.00	3rd Qu.: 0.000	3rd Qu.: 13
Max. :2189.0	Tue:312	May : 186	Max. :31.00	Max. :97.000	Max. :19401
	Wed:312	(Other):1074			
kidsAtHome	adultsAtHome	at.work	weblookup	medcalls	num.exchanges
Min. :20959	Min. : 37447	Min. : 0	Min. : 0.0	Min. : 0	Min. : 0.0
1st Qu.:31566	1st Qu.: 86910	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0	1st Qu.: 0.0
Median :78695	Median :217960	Median : 0	Median : 8.0	Median : 0	Median : 0.0
Mean :59889	Mean :151609	Mean : 41487	Mean : 695.9	Mean : 18867	Mean : 101.9
3rd Qu.:78701	3rd Qu.:217985	3rd Qu.: 0	3rd Qu.: 15.0	3rd Qu.: 0	3rd Qu.: 0.0
Max. :79497	Max. :226684	Max. :187787	Max. :19043.0	Max. :141408	Max. :11438.0
in.hospital	is.home	is.work	is.school	is.pharmacy	is.doctor
Min. : 0	Min. : 58563	Min. :0	Min. : 0	Min. : 0.0	Min. : 0.0
1st Qu.: 0	1st Qu.:118408	1st Qu.:0	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0.0
Median : 7	Median :296655	Median :0	Median : 0	Median : 0.0	Median : 0.0
Mean : 696	Mean :211498	Mean :0	Mean : 9277	Mean : 557.6	Mean : 125.9
3rd Qu.: 13	3rd Qu.:296683	3rd Qu.:0	3rd Qu.: 0	3rd Qu.: 13.0	3rd Qu.: 3.0
Max. :19401	Max. :306181	Max. :0	Max. :58370	Max. :25178.0	Max. :10315.0
is.stadium	is.theater	is.store	is.restaurant	is.university	is.military
Min. : 0	Min. : 0	Min. :0	Min. : 0	Min. :0	Min. :0
1st Qu.: 0	1st Qu.: 0	1st Qu.:0	1st Qu.: 0	1st Qu.:0	1st Qu.:0
Median : 0	Median : 0	Median :0	Median : 0	Median :0	Median :0
Mean :1437	Mean : 3997	Mean :0	Mean : 31342	Mean :0	Mean :0
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.:0
Max. :7408	Max. :20286	Max. :0	Max. :157115	Max. :0	Max. :0

Table 2.2 lists the outcome variables of the simulated disease outbreaks. The names are mostly self-explanatory.

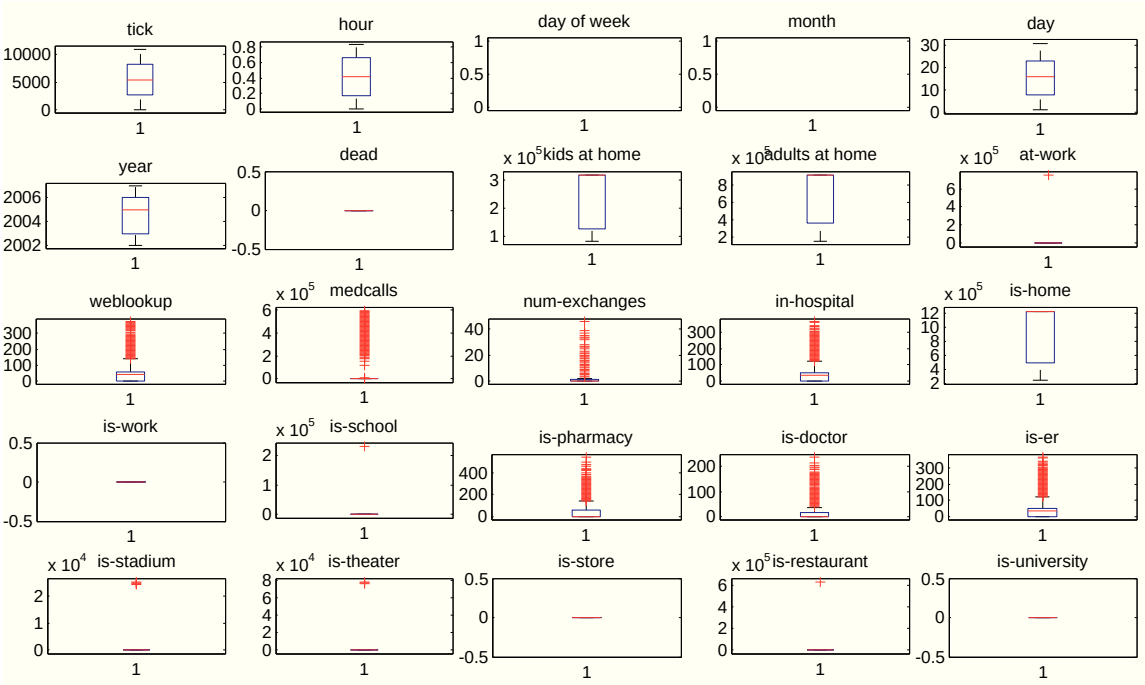
Table 2.4: Summary of the co-variates and output variables for BioWar-II data.

tick	dayOfWeek	month	day	dead	is.er
Min. : 0	Fri:1560	Dec : 930	Min. : 1.00	Min. :0	Min. : 0.00
1st Qu.: 2737	Mon:1566	Jan : 930	1st Qu.: 8.00	1st Qu.:0	1st Qu.: 0.00
Median : 5474	Sat:1560	Jul : 930	Median :16.00	Median :0	Median : 36.00
Mean : 5474	Sun:1566	Mar : 930	Mean :15.72	Mean :0	Mean : 38.94
3rd Qu.: 8212	Thu:1566	May : 930	3rd Qu.:23.00	3rd Qu.:0	3rd Qu.: 49.00
Max. :10949	Tue:1566	Oct : 930	Max. :31.00	Max. :0	Max. :368.00
	Wed:1566	(Other):5370			
kidsAtHome	adultsAtHome	at.work	weblookup	medcalls	num.exchanges
Min. : 85069	Min. :154198	Min. : 0	Min. : 0.00	Min. : 0	Min. : 0.0000
1st Qu.:126708	1st Qu.:362493	1st Qu.: 0	1st Qu.: 0.00	1st Qu.: 0	1st Qu.: 0.0000
Median :316864	Median :907737	Median : 0	Median : 42.00	Median : 0	Median : 0.0000
Mean :240999	Mean :622832	Mean :175130	Mean : 45.76	Mean :102464	Mean : 0.8463
3rd Qu.:316867	3rd Qu.:907859	3rd Qu.: 0	3rd Qu.: 57.00	3rd Qu.: 0	3rd Qu.: 1.0000
Max. :316867	Max. :907859	Max. :753630	Max. :375.00	Max. :595796	Max. :46.0000
in.hospital	is.home	is.work	is.school	is.pharmacy	is.doctor
Min. : 0.00	Min. : 239387	Min. :0	Min. : 0	Min. : 0.00	Min. : 0.000
1st Qu.: 0.00	1st Qu.: 489130	1st Qu.:0	1st Qu.: 0	1st Qu.: 0.00	1st Qu.: 0.000
Median : 36.00	Median :1224600	Median :0	Median : 0	Median : 0.00	Median : 0.000
Mean : 38.94	Mean : 863832	Mean :0	Mean : 37529	Mean : 37.51	Mean : 9.815
3rd Qu.: 49.00	3rd Qu.:1224726	3rd Qu.:0	3rd Qu.: 0	3rd Qu.: 56.00	3rd Qu.: 15.000
Max. :368.00	Max. :1224726	Max. :0	Max. :231797	Max. :542.00	Max. :237.000
is.stadium	is.theater	is.store	is.restaurant	is.university	is.military
Min. : 0	Min. : 0	Min. :0	Min. : 0	Min. :0	Min. :0
1st Qu.: 0	1st Qu.: 0	1st Qu.:0	1st Qu.: 0	1st Qu.:0	1st Qu.:0
Median : 0	Median : 0	Median :0	Median : 0	Median :0	Median :0
Mean : 4988	Mean :15666	Mean :0	Mean :127495	Mean :0	Mean :0
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.:0
Max. :25269	Max. :78581	Max. :0	Max. :634041	Max. :0	Max. :0

Table 2.3 and Table 2.4 summarize the basic statistics (e.g., mean, min, max, and median) of outcome variables in BioWar-I and BioWar-II, respectively. I used boxplots to illustrate the basic statistics of these outcome variables. I plotted one subfigure for each outcome variable because if outcome variables are plotted together, most outcome would be overwhelmed by a few outcome variables with high values. Please refer to Figure 2.1. Note that outcome variables "day of week" and "month" cannot be plotted because their contents are strings rather than numerical values.



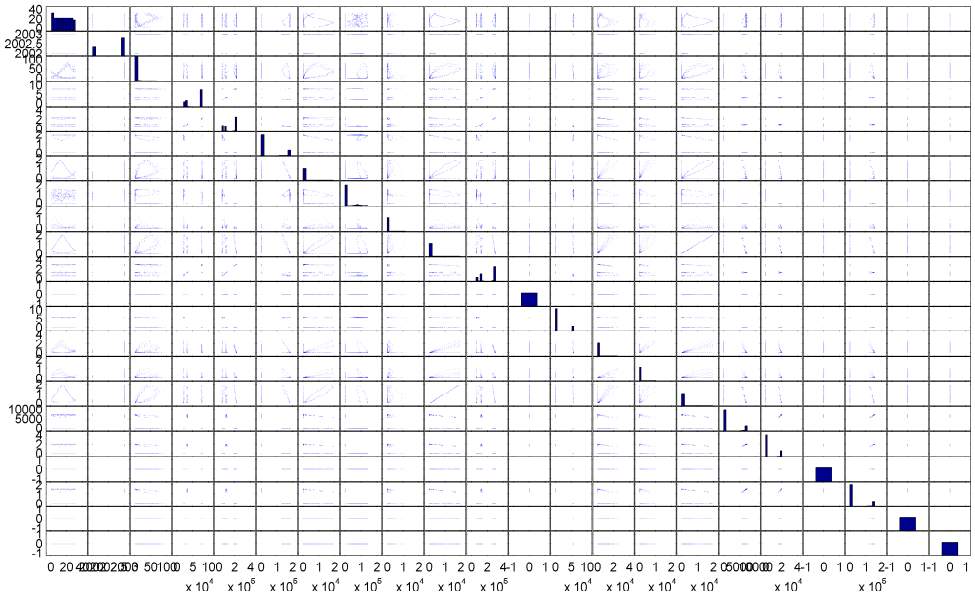
(a) BioWar I data.



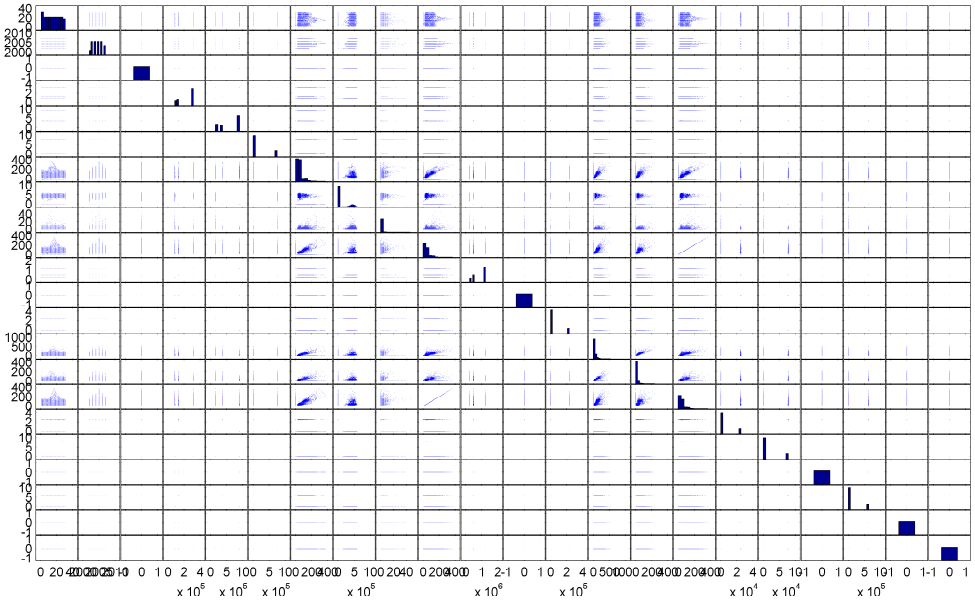
(b) BioWar II data.

Figure 2.1: Variable boxplots of BioWar-I and BioWar-II data.

The boxplots describe basic statistics of these outcome variables but cannot illustrate the distribution of these variables. To see more details, I used matrix plots to illustrate the distribution of individual outcome variables as well as their pairwise co-occurrence, as indicated in Figure 2.2. Although the subplots are tiny, Figure 2.2 shows that a strong relational dependency exists between various outcome variables.



(a) BioWar I data.



(b) BioWar II data.

Figure 2.2: Matrix plots for BioWar-I and BioWar-II data.

In addition to relational dependency, the data demonstrate interest in temporal correlations. For example, I plotted the outcome variable "death" against "days in a year" of BioWar-I in Figure 2.3. A single peak of death outbreak due to avian influenza is observed. The unimodal pattern of "death" is consistent with previous observations showing a strong time dependency of the outbreak.

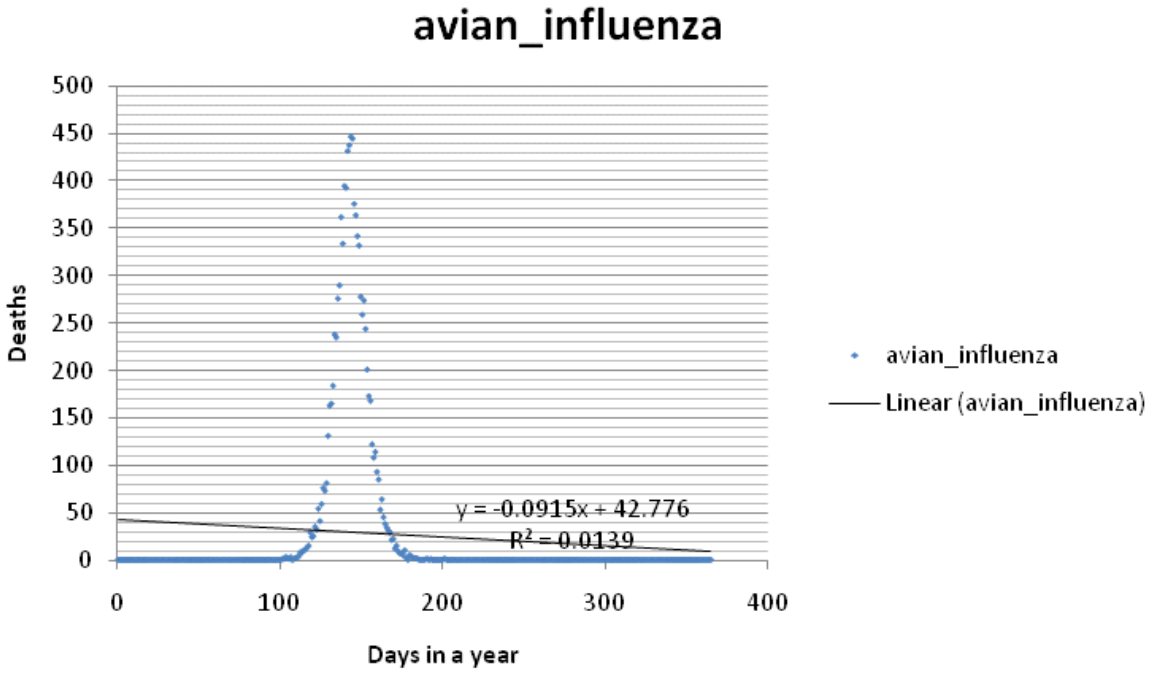


Figure 2.3: Daily aggregated measurements of co-variable deaths. There is an outbreak of the avian influenza between day 100 and day 200, when a large amount of mortality occurs. It can be observed that the death rate is highly correlated to the readings in the previous time tick. The data show a unimodal but non-stationary pattern because the mean of the distribution is not “locally constant”.

Although most outcome variables demonstrate time correlation, many of them show how patterns that are more complex than the outcome variable "death". The following figure shows the outcome variable "adult_at_home" in BioWar-I has three distinct patterns of working hours, off-work hours, and weekends.

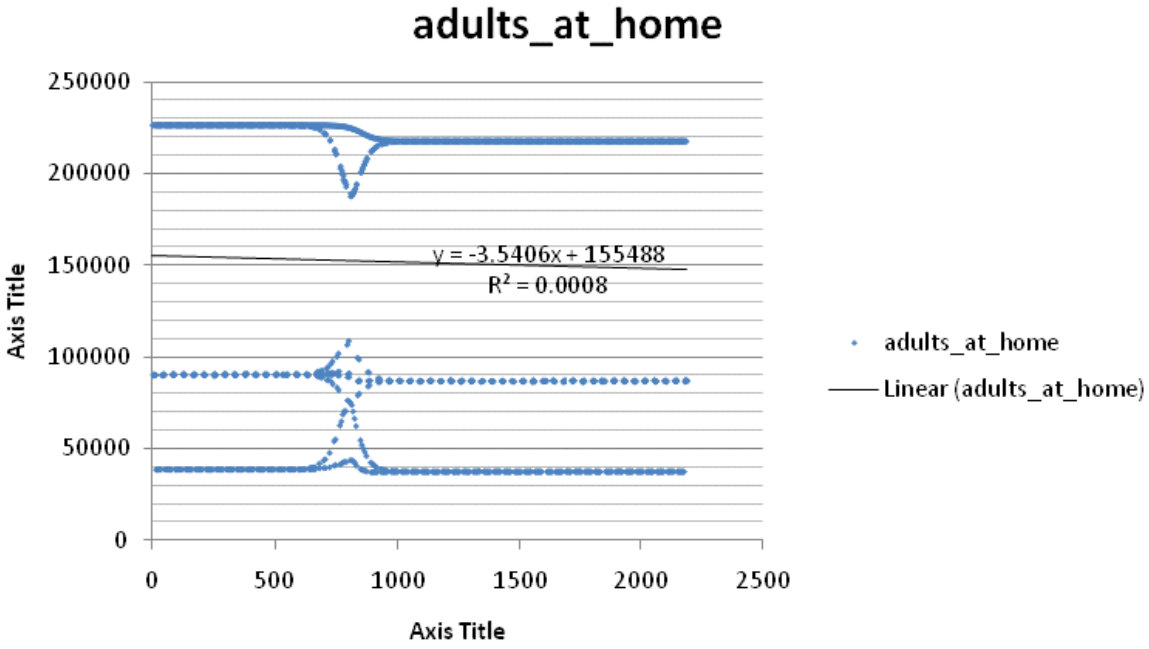


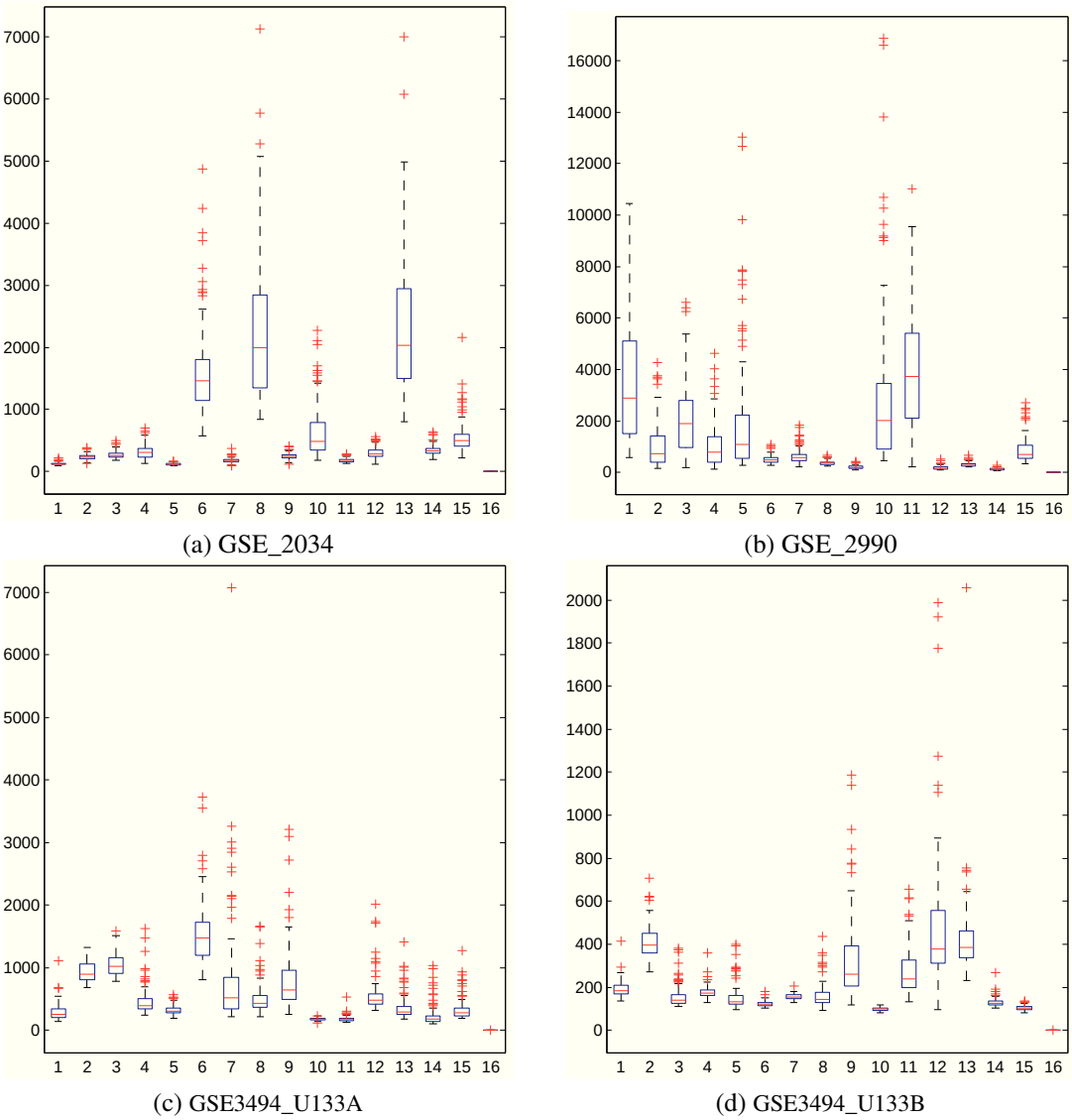
Figure 2.4: Daily aggregated measurements of co-variable "Adults_at_home". The X axis corresponds to the time tick and the Y axis corresponds to the number of adults at home. Obviously, there are three different patterns, which corresponds to the number of adults at home on working hours, weekends and nights, respectively.

2.2 Breast Cancer Gene Expression Data

The second data set, denoted as Breast Cancer, is obtained from the NCBI Gene Expression Omnibus (GEO). The three individual data downloaded were previously studied by Wang et al. (GSE2034) [187], Sotiriou et al. (GSE2990) [166], and Miller et al. (GSE3494) [128], respectively. I studied this dataset because gene expression is a good resource for personal medicine. New opportunities are available to treat patients differently for better performance based on more detailed information (Gene Expression) on individuals.

To make my data comparable with previous studies, I followed the criteria in [140] to select patients, who did not receive any treatment and had negative lymph node status. Among these pre-selected candidates, only patients with extreme outcomes, either poor outcomes (recurrence or metastasis within five years) or good outcomes (neither recurrence nor metastasis within eight years) were selected. The number of samples after filtering were: 209 for GSE2034 (114 good/95 poor), 90 for GSE2990 (60 good/30 poor), and 242 for GSE2034 (224 good/18 poor).

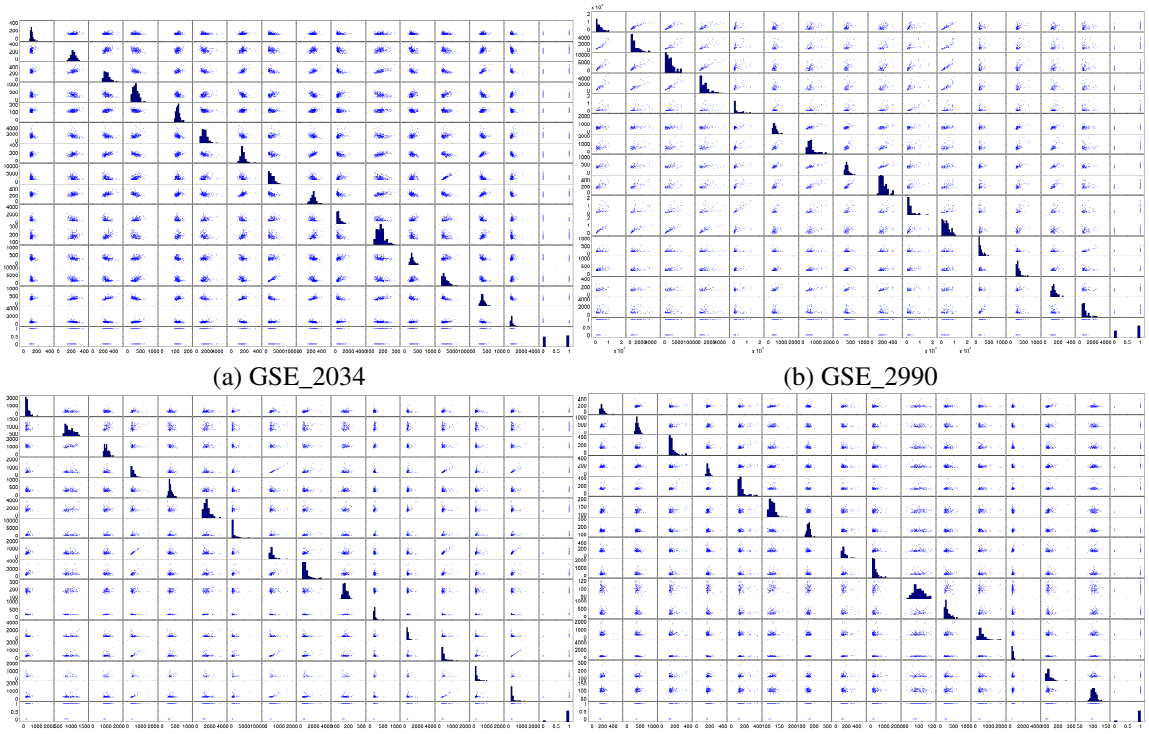
I also apply a split to divide GSE3494 into two groups, as suggested by [140], GSE3494-A and GSE3494-B, according to the sample's Affymetrix platform. Thus, the breast cancer data-set has four separate data. All of these data have a feature size of 247,965, which corresponds to the gene expression results obtained from micro-array experiments. They were preprocessed to keep only the top 15 features ranked using a t-test (see [140] for details). Figure 2.5 shows the boxplot of these selected gene features. It can be observed in the figures below that effective gene features are different from each other in different population groups.



(a) GSE_2034 (b) GSE_2990

(c) GSE3494_U133A (d) GSE3494_U133B

Figure 2.5: Boxplots of Breast Cancer Gene Expression data. Each column corresponds to one feature vector, and the last column indicates the outcome variable.



(a) GSE_2034 (b) GSE_2990

(c) GSE3494_U133A (d) GSE3494_U133B

Figure 2.6: Matrix plots of Breast Cancer Gene Expression data. Each subfigure corresponds to a matrix plot of one data set.

I also plotted the co-variable occurrence of breast cancer data in Figure 2.6 to investigate feature correlations visually.

2.3 Myocardial Infarction Data

The third dataset, Myocardial Infarction, corresponds to clinical Myocardial infarction (MI) patient records. I obtained the data from the authors of [98]. The goal of this study was to determine which, and how many data items are required to construct a decision support algorithm for early diagnosis of acute myocardial infarction using clinical and electrocardiographic data available at presentation [98].

These data were collected from patients admitted and discharged on a regimen. The data contains patient records of two medical centers in the Great Britain; among these, 500 patients admitted to the emergency department with chest pain were observed in Sheffield, England, and 1,353 patients with the same symptoms were observed in Edinburgh, Scotland.

Table 2.5: Description for the Myocardial Infarction variables.

ID	Abbreviation	Clinical Explainations
1-7	age	Age in years (under 30, 30-39, 40-49, 50-59, 60-69, 70-79, 80 and over)
8	Smokes	Smoker
9	Exsmoker	Ex-smoker
10	Fhistory	Family history of ischaemic heart disease
11	Diebetes	Diabetes mellitus
12	BP	Hypertension
13	Lipids	Hyperlipidaemia
14	CPmajorSymp	Is chest pain the major symptom?
15	Restrostern	Central chest pain
16	Lchest	Pain in left side of chest
17	Rchest	Pain in right side of chest
18	Back	Pain radiates to back
19	Larm	Pain radiates to left arm, neck or jaw
20	Rarm	Pain radiates to right arm
21	breath	Worse on inspiration
22	postural	Pain related to posture
23	Cwtender	Chest wall tenderness
24	Sharp	Pain described as sharp or stabbing
25	Tight	Pain described as tight, heavy, gripping or crushing
26	Sweating	Sweating
27	SOB	Short of breath
28	Nausea	Nausea
29	Vomiting	Vomiting
30	Syncope	Syncope
31	Episodic	Episodic pain
32-36	Worsening	Hours since 1st symptom (0-5, 6-10, 11-20, 21-40, over 40)
37-42	Duration	Hours of pain at presentation (0-5, 6-10, 11-20, 21-40, 41-80, over 80)
43	prev-ang	History of angina
44	Prev-MI	Previous myocardial infarction
45	Worse	Worse than usual angina/similar to previous acute myocardial infarction
46	Crackles	Fine crackles suggestive of pulmonary oedema
47	Added-HS	Added heart sounds
48	Hypoperfusion	Signs of hypoperfusion
49	Stelve	New ST-segment elevation
50	NewQ	New pathological Q waves
51	STorT-abnorm	ST segment or T-wave changes suggestive of ischaemia
52	LBBBorRBBB	Bundle branch block
53	Old-MI	Old electrocardiogram features of myocardial infarction
54	Old-isch	Electrocardiogram signs of ischaemia known to be old

The total number of patients is 1,853, the feature size is 54 and the target is a binary variable indicating whether a patient has myocardial infarction (MI). Table 2.5 summarizes the feature variables and their clinical meanings. Note that the last six features $(49 - 54)$ correspond to electrocardiograph readings that are

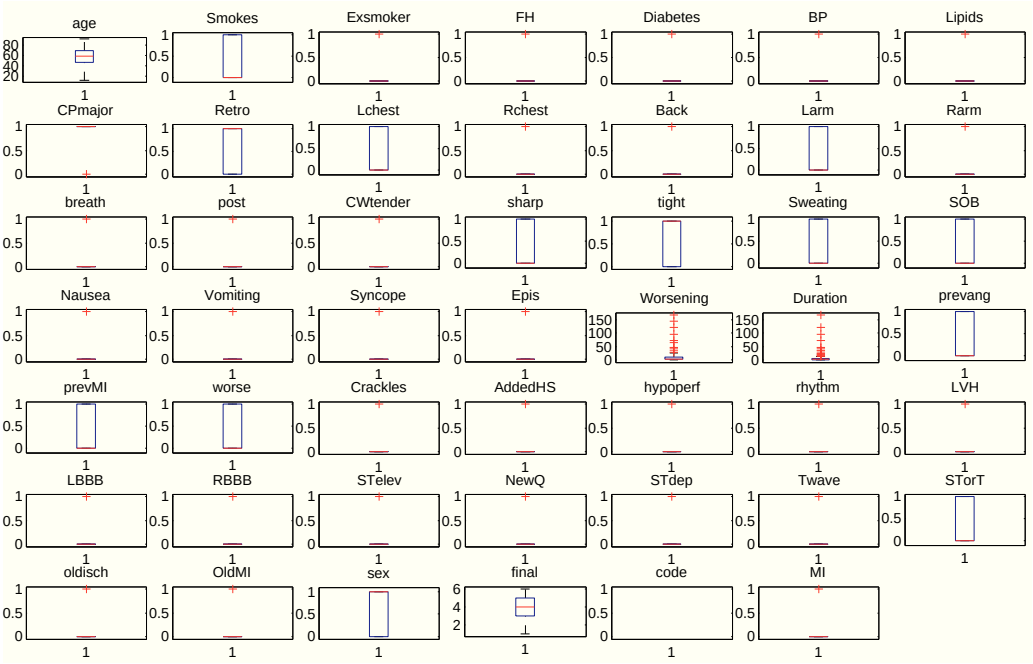
Table 2.6: Statistics of variables for the Edinburgh data.

Abbreviation				
age	min: 13.0	median:59	mean:57.6	max: 92
Smoker	0: 785	1: 468		
Exsmoker	0: 959	1: 294		
Fhistory	0: 967	1: 286		
Diabetes	0: 1165	1: 88		
BP	0: 1053	1: 200		
Lipids	0: 1215	1: 38		
CPmajorSymp	0: 62	1: 1191		
Restrostern	0: 331	1: 922		
Lchest	0: 907	1: 346		
Rchest	0: 1109	1: 144		
Back	0: 1122	1: 131		
Larm	0: 670	1: 583		
Rarm	0: 1042	1: 211		
breath	0: 1031	1: 222		
postural	0: 1017	1: 236		
Cwtender	0: 1201	1: 52		
Sharp	0: 1208	1: 45		
Tight	0: 572	1: 681		
Sweating	0: 739	1: 514		
SOB	0: 731	1: 522		
Nausea	0: 1124	1: 129		
Vomiting	0: 1124	1: 129		
Syncope	0: 1208	1: 45		
Episodic	0: 1161	1: 92		
Worsening	min: 0.0	median: 4.0	mean: 17.4	max: 168
Duration	min: 0.0	median: 3.0	mean: 8.84	max: 168
prev-ang	0: 699	1: 554		
prev-MI	0: 836	1: 361		
Worse	0: 892	1: 361		
Crackles	0: 1106	1: 147		
Added-HS	0: 1247	1: 6		
Hypoperfusion	0: 1203	1: 50		
Stelve	0: 1199	1: 54		
NewQ	0: 1240	1: 13		
STorT-abnorm	0: 1240	1: 13		
LBBBorRBBB	0: 1203	1: 50		
Old-MI	0: 1101	0: 152		
Old-isch	0: 1141	1: 112		
MI	0: 979	1: 274		

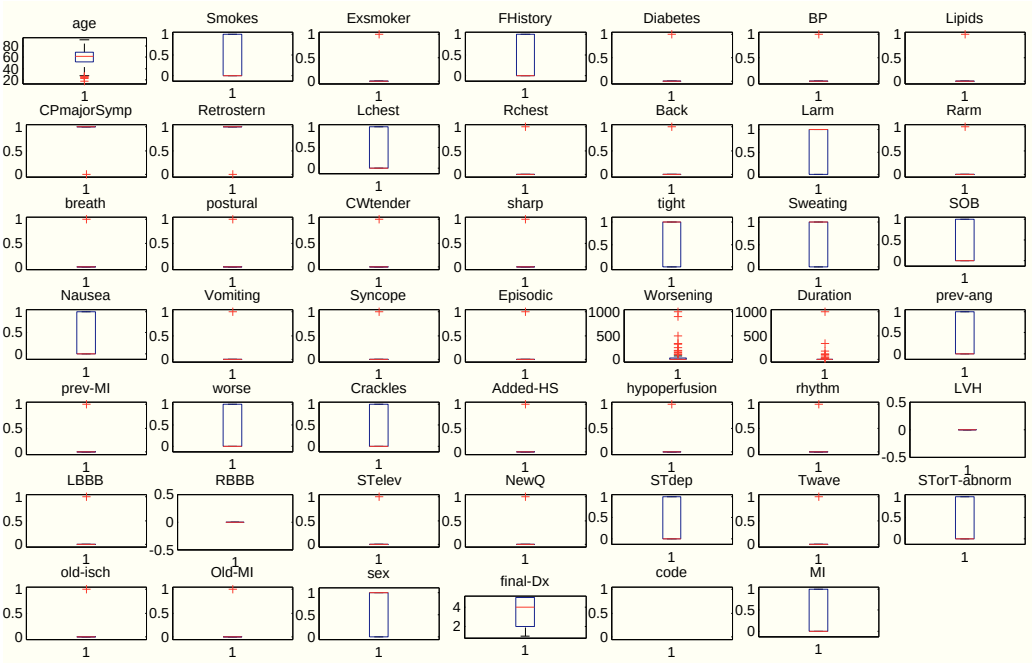
Table 2.7: Statistics of variables for the Sheffield data.

Abbreviation				
age	min: 17.0	median:61	mean:59.9	max: 91
Smoker	0: 318	1: 182		
Exsmoker	0: 388	1: 112		
Fhistory	0: 373	1: 127		
Diabetes	0: 451	1: 49		
BP	0: 403	1: 97		
Lipids	0: 482	1: 18		
CPmajorSymp	0: 37	1: 463		
Restrostern	0: 110	1: 390		
Lchest	0: 373	1: 127		
Rchest	0: 438	1: 62		
Back	0: 426	1: 74		
Larm	0: 237	1: 263		
Rarm	0: 418	1: 82		
breath	0: 422	1: 78		
postural	0: 455	1: 45		
Cwtender	0: 491	1: 9		
Sharp	0: 400	1: 100		
Tight	0: 246	1: 254		
Sweating	0: 235	1: 265		
SOB	0: 281	1: 219		
Nausea	0: 341	1: 159		
Vomiting	0: 449	1: 51		
Syncope	0: 467	1: 33		
Episodic	0: 417	1: 83		
Worsening	min: 0.0	median: 6.0	mean: 50.37	max: 1000
Duration	min: 0.0	median: 4.0	mean: 12.34	max: 1000
prev-ang	0: 281	1: 219		
prev-MI	0: 377	1: 123		
Worse	0: 338	1: 162		
Crackles	0: 373	1: 127		
Added-HS	0: 476	1: 24		
Hypoperfusion	0: 441	1: 59		
Stelve	0: 403	1: 97		
NewQ	0: 470	1: 30		
STorT-abnorm	0: 403	1: 97		
LBBBorRBBB	0: 474	1: 26		
Old-MI	0: 454	1: 46		
Old-isch	0: 473	1: 27		
MI	0: 346	1: 154		

highly correlated to the target, so should not be included for prediction. I represent every categorical feature by a set of binary features in order for it to be applicable to learning algorithms.



(a) Edinburgh MI data.



(b) Sheffield MI data.

Figure 2.7: Boxplots of Myocardial Infarction data.

2.4 Hospital Discharge Error Data

The last dataset, Hospital Discharge Error data, involves 77,348 patient records related to a real world Microbiology Culture Follow-up Errors study. The dataset was created through a retrospective analysis of all microbiology cultures preformed at an academic hospital in Boston, MA in 2007.

Figure 2.8 overviews the data, where the number of patients involved in every stage of the clinical decision are listed. Of $77,348$ inpatient culture results, 4819 (6%) are returned post-discharge.

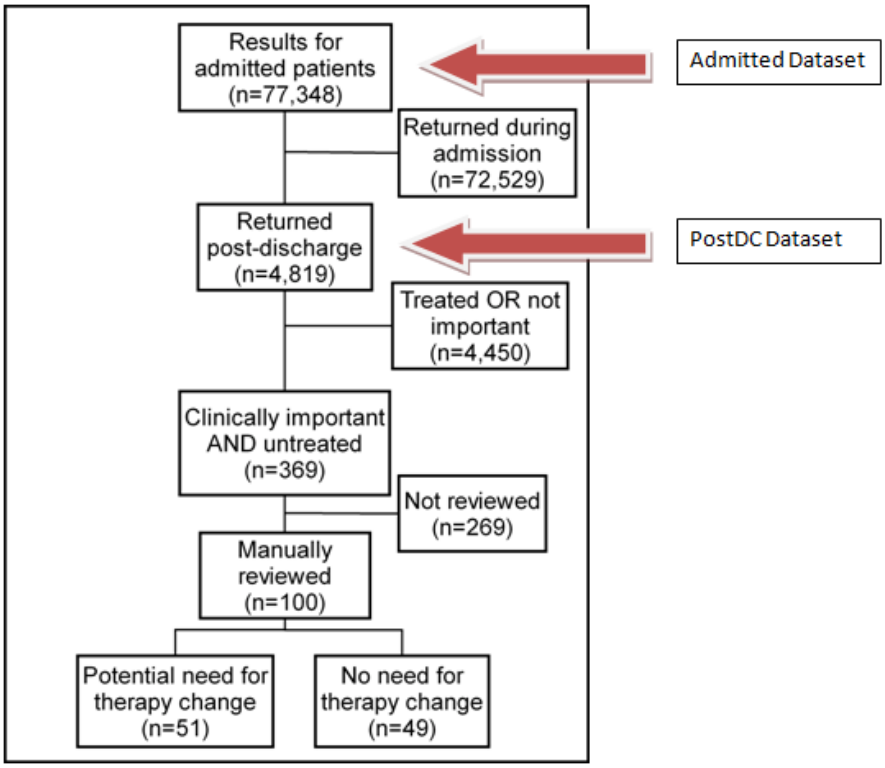


Figure 2.8: Overview of the hospital discharge error data.

Of all post-discharged patients, 369 were clinically important and untreated at discharge; among 100 manually-reviewed cases, 51% potentially required a change in therapy [59]. Urine cultures were more likely to potentially require change in therapy than non-urine cultures (Odds Ratio: 2.9, 95% Confidence Interval: $1.2 - 7.2$; p=0.02); in addition, 73% of 26 results from surgical services potentially required a therapy change, compared to 57% of 30 results from general medicine, 38% of 16 results from oncology, and 33% of 27 results from medical sub-specialties. Overall, 3.9% of post-discharge cultures potentially

necessitated an antibiotic change [59].

Regarding patient demographics, this data contains age, gender, race and insurance. Regarding the hospital encounter, the dataset contains the visit type (admission, emergency room, procedure or outpatient) and admitting service, if applicable. Related to the microbiology result, the dataset contains the specimen type (blood, urine, sputum and cerebral spinal fluid), the hospital day number that the specimen was collected, whether the result was pending at the time of discharge from the hospital, whether the specimen was collected on a weekend, whether the preliminary results (for blood cultures) were reported on a weekend, and whether the final results were reported on a weekend. In addition to the data pulled directly from the hospital computer system, this dataset contains an additional outcome variable, which indicates whether the case represents a potential post-discharge follow-up error using experts' knowledge. This variable is true if the following three criteria are met: (1) the result is considered clinically relevant; (2) the results return after the patient is discharged from the hospital; and (3) there is no antibiotic on the discharge medication list to which the organism is sensitive based on the microbiology results. The features thus consisted of are thus consisted of eight categorical variables and two numerical variables. The target is a Boolean variable (Pot_error) indicating the potential error.

Table 2.8: Description of variables in hospital discharge error data. Eight out of ten explanatory variables are categorical and two variables are numerical.

Name	**Details**
Features	
Specimens:	0=blood, 1=urine, 2=sputum, 3=csf
Spec_days:	Number of days between admission date and specimen collection date.
Collect_week:	0=specimen collected on Weekday, 1=specimen collected on Weekend
Final_week:	0=final result on Weekday, 1=final result on Weekend
Vistyp:	1=admission, 0=non-admission
Svc:	0=<blank> (patient not admitted), 1=ONC, 2=MED, 3=Medical Sub-specialties, 4=Surgery and Surgical Sub-specialties, 5=Other
Age:	Age in years
Female:	0=male, 1=female
Race:	0=white, 1=black, 2=Asian, 3=Hispanic, 4=other, 5=unknown/declined
Insurance:	0=medicare, 1=medicaid, 2=commercial, 3=other
Target Variable	
Pot_error:	0=not a potential follow-up error, 1=a potential follow-up error

Table 2.8 summarizes both feature variables and the target variable. To handle the categorical features, I explicitly express each categorical variable as a set of Boolean variables. For example, specimens are replaced by three Boolean variables. The fully expanded feature set thus has 20 dimensions.

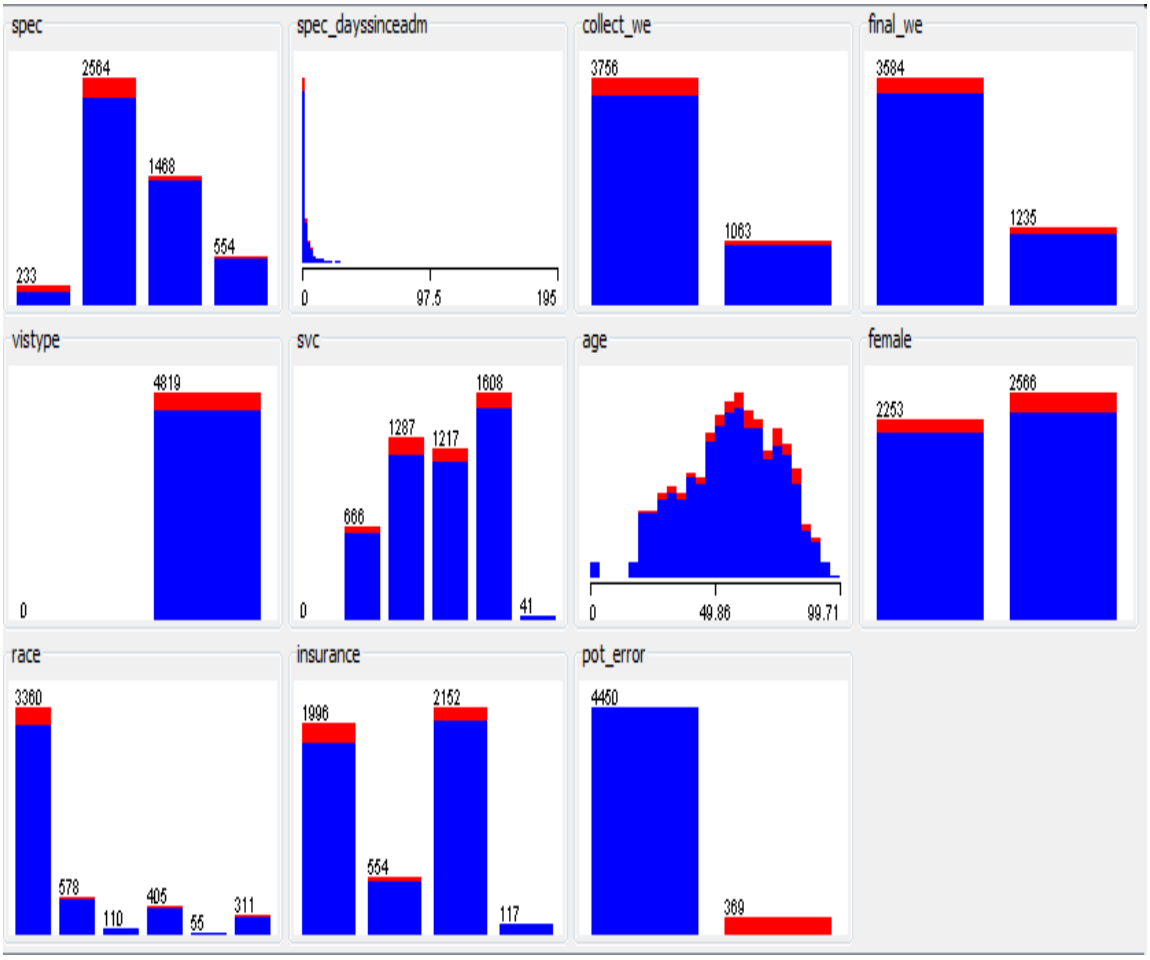


Figure 2.9: Histograms of variables in hospital discharge error data.

In the thesis, I focus on the prediction of 369 clinically important and highly suspicious observations out of 4,819 returned post-discharge observations. Figure 2.9 illustrates the distribution of these variables. It is easy to observe that there is a significant data unbalance between suspicious observations and unlabeled observations.

I also illustrated the co-variate occurrence pattern in Figure 2.10 using matrix plots.

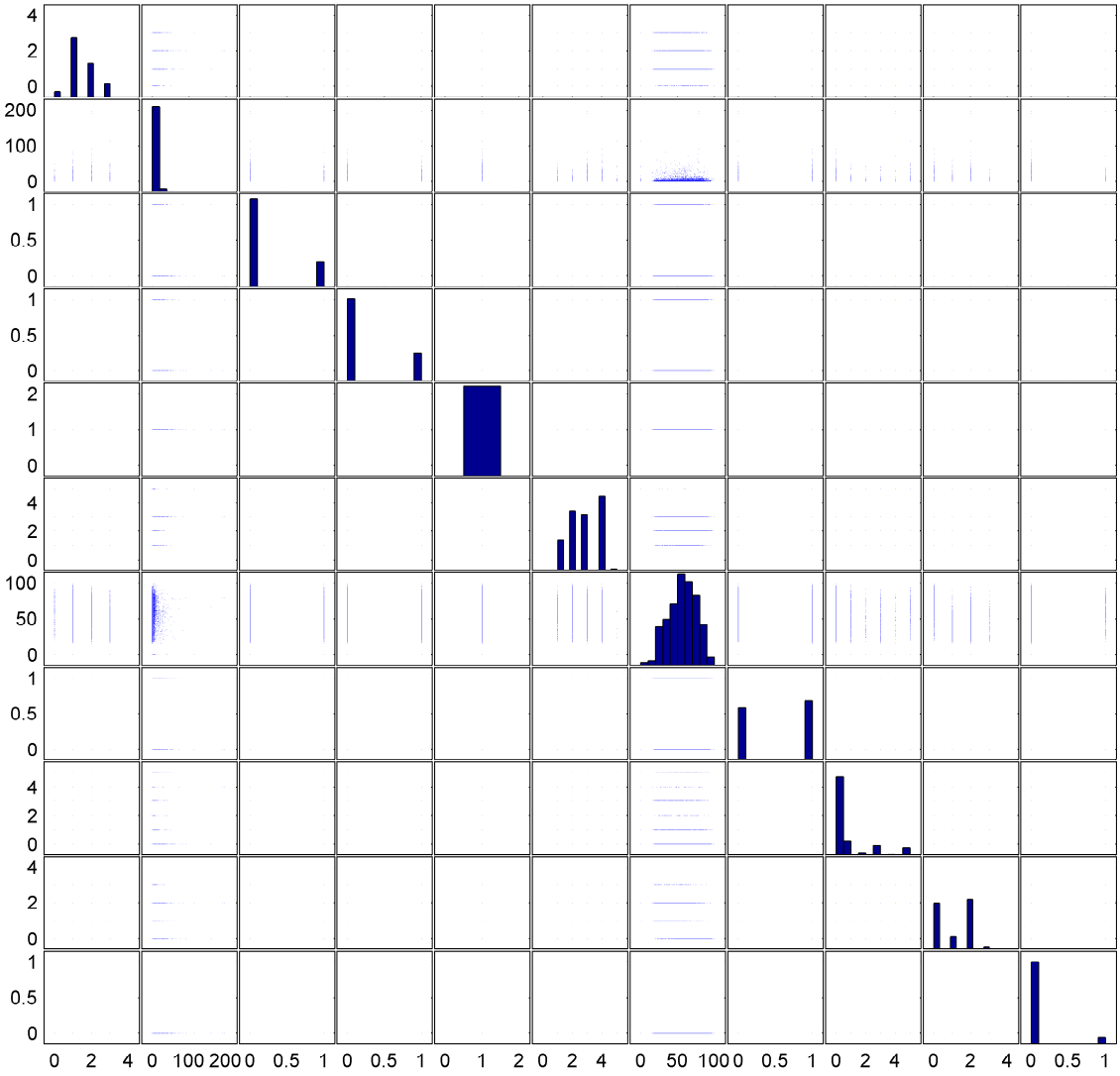


Figure 2.10: Matrix plots of hospital discharge error data.

In this chapter, I described four datasets, which directly motivated this thesis. These datasets represent two typical types of of biomedical problems: the large scale co-estimation problem and the calibration for personalized medicine problem. I preserved the description of other datasets used to verify models in separate chapters.

2.5 Comparison

A summary comparison of the basics of the four datasets, e.g., feature size, outcome variable type and number of subjects, is shown in Table 2.9. BioWar-I, II data have a large amount of subjects (~1e5) and a relatively smaller number of co-variables. In contrast, Breast Cancer Gene Expression data have a small number of subjects but a large number of co-variables (~1e5). Other datasets like Myocardial Infarction and Hospital Discharge Error have intermediate sizes of both the number of subjects and the number of co-variables.

The main differences between these datasets are: 1) the type of their outcome variables; 2) the number of outcome variables of interest. For example, regarding BioWar-I, II data, I am interested in predicting multiple outcome variables of discrete values while with the other data, my aim is to predict a single binary outcome. These differences are due to the nature of the problem of interest.

Table 2.9: Comparison of four datasets.

Dataset		Number of Subjects	Number of Co-Variables	Number of Outcome Variables	Type of Outcome Variable
BioWar	BioWar-I	306,181	25	25	Discrete value
	BioWar-II	1,224,726	25	25	Discrete value
Breast Cancer	GSE2034	209	247,965	1	Binary
	GSE2990	90	247,965	1	Binary
	GSE3494-A	242	247,965	1	Binary
	GSE34394-B	242	247,965	1	Binary
Myocardial Infarction	Sheffield	1,353	54	1	Binary
	Edinburgh	500	54	1	Binary
Hospital Discharge Error		4819	20	1	Binary

The BioWar-I and II data encode relational dependencies between outcome variables over time. These data are compounded by manifestations from different sources related to the same disease outbreak event. The decision of whether or not to intervene requires more comprehensive information about the current situation. Thus, a useful prediction model has to jointly consider multiple latent factors (states) implied by these manifestations. To this end, BioWar simulation data are treated as a dynamic system with multiple correlated outcome variables that evolves over time. Because decision makers are interested in interpretable results, my goal is to predict the hidden states (discrete values) of continuous measurements for manifestations. For this purpose, prediction accuracy is the most important metric for evaluating the performance.

The datasets, Breast Cancer, Myocardial Infarction and Hospital Discharge Error are different from

BioWar-I, II. These data correspond to patient records with dichotomous outcomes, e.g. having a breast cancer or not. The motivation of this study is not to improve the accuracy of existing models but increase the reliability of predictions. Concretely, reliability (alternative known as the "*calibration*") is reflected in how the outputs represent the true probability of the class membership [33]. For this reason, the data of interest are those with a single binary outcome variable, which is essential for calculating the class membership.

Because these data are significantly different, I developed a specific model to handle each of them. Otherwise, models that provide reliable estimates of binary class memberships cannot reveal the multiclass hidden states of a disease outbreak, and models designed to estimate temporal and relational correlated latent factors cannot offer a *calibrated* inference of class memberships of novel co-variates.

2.6 Limitation

I investigated four different datasets in this chapter. These datasets correspond to two problems: a) bioterrorism related disease outbreak and b) *calibration* for personalized medicine, which motivated my research in biomedical informatics.

The data were collected from different sources and cover a range of typical biomedical applications, including bioterrorism related disease outbreak, Breast Cancer Gene Expression, Myocardial Infarction and Hospital Discharge Error. The first one is about public health and the rest correspond to clinically relevant personalized medicine. Despite the variety of motivations, these data have a number of common characteristics, e.g., the type of their outcome variable and the co-variable patterns. Both BioWar-I and BioWar-II have multiple outcome variables, while Breast Cancer Gene Expression, Myocardial Infraction and Hospital Discharge Error data shared a common bond: a single dichotomous outcome variable. The commonalities offered by these data provide compelling support for a deep exploration of a generalized prediction model for decision making support.

The characteristics of these data also limit the scope of my thesis. Because datasets are significantly different, I have to develop specific models to handle each of them, e.g., one model deals with multiple relational correlated hidden states over time, the other model focuses on providing reliable predictions of class membership. The BioWar-I and BioWar-II data imply a dynamic system with multiple correlated outcome variables that evolves over time. Due to computational complexity and noisy observation, I had to consider the mutual coupling effect of various factors over time concurrently, thus narrowing the search

space for feasible solutions in a discriminative manner. Regarding the other data related to *calibration* for personalized medicine, they all have a single binary outcome variable, by which the class memberships are estimated. Such data characteristics limited my exploration for *calibration* to be in accordance with single outcome probabilistic models. Despite these limitations, these data serve well to illustrate prominent biomedical problems, motivate meaningful research, and provide compelling support for evaluation.

Chapter 3

Background

This thesis aims to develop prediction models to support biomedical decision making. To understand the techniques described in the following chapters, it is necessary to introduce the basics about models and evaluation metrics that are relevant. One particular type of models that need to be discussed is called "structured learning model", which involve optimizing dependent states in addition to modeling individual manifestations. These models include a broad range of popular methods, including Hidden Markov Model, Kalman filter, Conditional Random Fields and Maximum Margin Markov Networks. They are also basics of the framework I developed for modeling bioterrorism-related disease outbreaks. Thus, I decided to use the first half of this chapter to provide necessary backgrounds of structured models.

The second half of this chapter discusses two major components of the model performance evaluation. They are the basis for accessing the probabilistic model's validity and reliability, and for understanding the insufficiency of existing approaches. The first metric is *discrimination* ability, which indicates the model's ability to rank various patients correctly. This metric is the goal of most cohort studies; however, it is not sufficient to support decision making at a personal level. Thus, I introduced another important but less studied metric called *calibration*, which stratifies how outcomes affect various genetic population groups within a patient-diagnosis population. Specifically, this metric indicates how well predicted values represent observed outcomes.

3.1 Structured Learning Frameworks

This section provides an overview of the popular structured learning frameworks for the temporal model and the relational model. Specifically, I cover models including the Hidden Markov Model (HMM) [152], Kalman Filter (KF) [96], Conditional Random fields (CRFs) [107] and Maximum Margin Markov Networks (M3N) [172]. These models are closely related to my models developed in later chapters.

3.1.1 Temporal Models

The basic temporal model is Dynamic Systems (LDS) with Gaussian noise [72]. I denote states as $Y = [y_1, \ldots, y_p]$ as a k-dimension hidden state vectors; $X = [x_1, \ldots, x_p]$ as p-dimension feature vectors.

In most scenarios, the states Y cannot be observed directly. However, the models assume that hidden states can be summarized by k-dimension state variables ($k \ll p$). At each time step, an observable feature set, the p-dimension X generated by the system use underlying states, is non-observable.

The state Y evolves with a simple first-order Markov dynamic. Note such evolution is hidden and the observations are corrupted by an additive Gaussian noise. For both continuous valued and discrete valued state variable Y, the basic generative model can be written as:

$$Y_{t+1} = AY_t + w, \quad , w \sim \mathcal{N}(0, Q), \tag{3.1}$$

$$X_t = CY_t + v, \quad , v \sim \mathcal{N}(0, R), \tag{3.2}$$

where A is the $k \times k$ state transition matrix and C is the $p \times k$ observation measurement matrix.

> **Interpretation:** The LDS model specifies a linear relationship between state consequences using a Markov chain. The observations X are sampled from a distribution of states Y.

The k-dimension vector w and p-dimension vector v correspond to random variables representing state evolutions and observation noises, respectively. The noises are time independent and Gaussian distributed with zero mean and covariance matrices, Q and R. These nosies are essential to the system. Without the noise w, the state Y_t would always either shrink exponentially to zero or blow up exponentially. Similarly, in the absence of the observation noise v the state would no longer be hidden. Figure 3.1 illustrates this basic

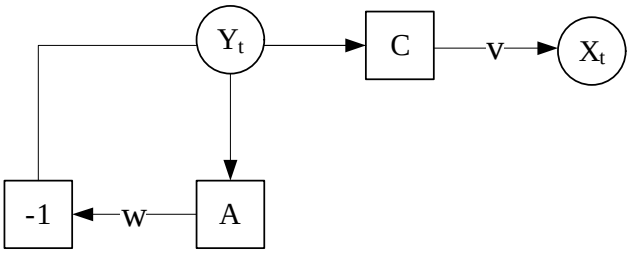


Figure 3.1: Generative model of the linear dynamical system. The -1 block is a unit delay. The covariance matrix of the input noise w is Q and the covariance matrix of the output noise v is R.

model.

The assumption made by Equation 3.1 is that the hidden state sequence Y_t should be an informative lower dimensional projection of more complicated observations X_t. With linear dynamics and noises, these states should summarize driving factors of observations more compactly. Thus, LDS models work with low dimensional states rather than high dimensional observables ($k \ll p$).

Two most popular methods in temporal LDS are the Kalman filter [96] and Hidden Markov Model [152], which usually assume Gaussian inputs. The popularity of linear Gaussian models comes from two useful analytical properties: first, the sum of two independent Gaussian distributed quantities is also Gaussian distributed; second, the output of a linear system whose input is Gaussian distributed is again Gaussian distributed. Accordingly, if $Y_1 \sim \mathcal{N}(\mu_1, Q_1)$, then all future states Y_t and observations X_t will also be Gaussian distributed, where t indicates a time tick. The explicit formula for the conditional expectations of the states and observables can be written as:

$$\begin{aligned} P(Y_{t+1}|Y_t) &= \mathcal{N}(AY_t, Q)|_{Y_{t+1}}, \\ P(X_t|Y_t) &= \mathcal{N}(CY_t, R)|_{X_t}. \end{aligned}$$

The joint probability of a sequence of states and outputs can be written as,

$$P(\{X_1, \ldots, X_T\}, \{Y_1, \ldots, Y_T\}) = P(Y_1) \prod_{t=1}^{T-1} P(Y_{t+1}|P_t) \prod_{t=1}^{T-1} P(X_t|Y_t),$$

where T is the total period of training. The negative log probability is just the sum of matrix quadratic forms,

$$
\begin{aligned}
&- 2\log P(\{X_1,\ldots,X_T\},\{Y_1,\ldots,Y_T\}) \\
&= \sum_{t=1}^{T}[(X_t - CY_t)^{'} R^{-1}(X_t - CY_t) + \log|R|] + \sum_{t=1}^{T-1}[(Y_{t+1} - AY_t)^{'} Q^{-1}(Y_{t+1} - AY_t) + \log|Q|] \\
&+ (Y_1 - \mu_1)Q_1^{-1}(Y_1 - \mu_1) + \log|Q_1| + T(p+k)\log 2\pi. \qquad (3.3)
\end{aligned}
$$

Interpretation: The objective function takes the log-likelihood form rather than the productions for two reasons: 1) it helps to alleviate a floating point error due to the proximity of tiny numbers of production; 2) it offers a more convenient operation to factorize the model.

Given fixed model parameters $\{A, C, Q, R, \mu_1, Q_1\}$, a very basic task is to estimate the total likelihood of an observation sequence,

$$
p(\{X_1 \ldots X_T) = \int_{\forall\{Y_1 \ldots Y_T\}} P(\{X_1,\ldots,X_T\},\{Y_1,\ldots,Y_T\}) d(\{Y_1,\ldots,Y_T\}, \qquad (3.4)
$$

the marginalization of which requires an efficient way of integrating the joint probability (results of Equation 3.3) over all possible configurations. With this integral, it is simple to compute the conditional distribution for a proposed hidden state sequence given the observations by dividing the joint probability by the total likelihood of the observations,

$$
P(\{Y_1,\ldots,Y_T\}|\{X_1,\ldots,X_T\}) = \frac{P(\{X_1,\ldots,X_T\},\{Y_1,\ldots,Y_T\})}{P(\{X_1,\ldots,X_T\})}. \qquad (3.5)
$$

Interpretation: The conditional probability is closely related to Equation 3.4 and often is the intermediate value of a recursive method such as the Viterbi algorithm. Most discriminative structured learning models optimize the same objective but with different regularization criteria.

The next problem is about learning: given the observed sequence (or perhaps several sequences) of outputs $\{X_1,\ldots,X_T\}$, the question is how to find the parameters $\{A, C, Q, R, \mu_1, Q_1\}$ that maximize the likelihood of the observations? One approach is to use the expectation-maximization (EM) algorithm originally

proposed by Shumway et al. [163], and extended by Ghahramani et al. [75]. The objective of the algorithm is to maximize the likelihood of the observation in the presence of hidden variables. Let $\mathbf{Y} = \{Y_1, \ldots, Y_T\}$, $\mathbf{X} = \{X_1, \ldots, X_T\}$ and denote parameters as θ. Hence, maximizing the likelihood as a function of θ is equivalent to maximizing the following log-likelihood:

$$\mathcal{L}(\theta) = \log P(\mathbf{Y}|\theta) = \log \int_{\mathbf{X}} p(\mathbf{X}, \mathbf{Y}|\theta) d\mathbf{X}.$$

I can obtain a lower bound on $\mathcal{L}$ use any distribution Q over hidden variables,

$$\begin{aligned}
\log \int_{\mathbf{X}} P(\mathbf{Y}, \mathbf{X}|\theta) d\mathbf{X} &= \log \int_{\mathbf{X}} Q(\mathbf{X}) \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{Q(\mathbf{X})} d\mathbf{X} \\
&\geq \int_{\mathbf{X}} Q(\mathbf{X}) \log \frac{P(\mathbf{Y}, \mathbf{X}|\theta)}{Q(\mathbf{X})} d\mathbf{X} \qquad (3.6) \\
&= \int_{\mathbf{X}} Q(\mathbf{X}) \log P(\mathbf{Y}, \mathbf{X}|\theta) d\mathbf{X} - \int_{\mathbf{X}} Q(\mathbf{X}) \log Q(\mathbf{X}) d\mathbf{X} \qquad (3.7) \\
&= \mathcal{F}(Q, \theta), \qquad (3.8)
\end{aligned}$$

where Equation 3.6, is known as Jensen's inequality.

> **Interpretation:** Equation 3.6 provides a lower bound of the log-likelihood function, which is not factorizable. Maximizing this lower bound encourages a maximization of the loglikelihood objective.

The EM algorithm alternates between maximizing $\mathcal{F}$ with respect to the distribution Q and the parameter θ, respectively, keeping the other fixed. Starting from some initial parameters θ_0,

$$\begin{aligned}
\textbf{E - step}: \quad Q_{k+1} &\leftarrow \operatorname{argmax}_Q \mathcal{F}(Q, \theta_k), \qquad (3.9) \\
\textbf{M - step}: \quad \theta_{k+1} &\leftarrow \operatorname{argmax}_\theta \mathcal{F}(Q_{k+1}, \theta). \qquad (3.10)
\end{aligned}$$

The maximum in the E-step results is when Q is exactly the conditional distribution of $\mathbf{X}$ is $Q_{k+1}(\mathbf{X}) = P(\mathbf{X}|\mathbf{Y}, \theta_k)$, at which point the bound becomes an equality: $\mathcal{F}(Q_{k+1}, \theta_k) = \mathcal{L}(\theta_k)$.

Interpretation: Equation 3.9 and Equation 3.10 take alternative steps to maximize a non-convex function. By fixing one objective at a time, the EM algorithm reaches a local optimal of the objective (Equation 3.6) quickly.

The maximum of the M-step is obtained by maximizing the first term in Equation 3.7, because the entropy of Q does not depend on θ,

$$\textbf{M - step}: \quad \theta_{k+1} \leftarrow \text{argmax}_\theta \int_{\mathbf{X}} P(\mathbf{X}|\mathbf{Y}, \theta_k) \log P(\mathbf{Y}, \mathbf{X}|\theta) d, \tag{3.11}$$

since at the beginning of each M-step $\mathcal{F}(Q_{k+1}, \theta_k) = \mathcal{L}(\theta_k)$ and E-step does not change θ, the EM algorithm is guaranteed not to decrease the likelihood at each iteration.

3.1.1.1 Kalman Filter

For continuous time series data, temporal ordering is critical. The state evaluation dynamics, which provide the only aspect of temporal ordering, thus should not be ignored. Such models have traditionally been the focus of the control community and received a lot of attention from the power industry as well.

A popular model, described by Equation 3.14 and Equation 3.13, is called Kalman filter [96].

$$Y_{t+1} = AY_t + w, \quad , w \sim \mathcal{N}(0, Q), \tag{3.12}$$

$$X_t = CY_t + v, \quad , v \sim \mathcal{N}(0, R), \tag{3.13}$$

where A is no longer the state transition matrix as before. The k-dimension vector w and the p-dimension vector v are temporally white and spatially Gaussian distributed noises independent of each other.

Interpretation: Kalman filter is a typical Linear Dynamic System. The states are modeled by a Markov chain, and its observations are assumed to be induced by the hidden states.

Kalman [96] proposed an efficient recursive solution to the inference problem. The learning problem (estimating model parameters) was later studied by Ghahramani et al. [75] and Digalakis et al. [57].

In the process of imaging in the state-space, the points are embedded by a sphere (described by Q). The embedding sphere is stretched into a ellipsoid in the observation space by C. This ellipsoid is convolved

with the observation noise covariance (described by R). The center of the state-space ball moves over time, and its location is determined by eigenvalues and eigenvectors of the matrix A. The center is moved to a new point according to this flow field (induced by eigenvalues and eigenvectors of A); then it relocates to pick a new state. From the new state, it moves to a new point and iterates. If A is an identity matrix (not the zero matrix), the "flow" does not move but evolves according to a random walk of the noise set by Q.

3.1.1.2 Hidden Markov Model

The Hidden Markov Model (HMM) can be obtained by some simple modifications to the previous continuous state model. The HMM specifies the state at any time with a discrete value. Many processes, especially those with distinct modes of operation, are better described by internal states that are non-continuous.

The HMM state evolution is still a first-order Markovian, and the observation process is a linear process with an additive Gaussian noise. The difference between HMM and the Kalman filter is the use of the winner-take-it-all operation $\mathrm{WTA}(\cdot)$, which is defined as a unity vector X in the position of the largest coordinate of HMM but as zeros in all other positions. The HMM model is:

$$Y_{t+1} = \mathrm{WTA}(AY_t + w), \quad , w \sim \mathcal{N}(0, Q), \tag{3.14}$$

$$X_t = CY_t + v, \quad , v \sim \mathcal{N}(0, R), \tag{3.15}$$

where A is a matrix, but it is no longer called the state transition matrix. Like the Kalman filter, the k-dimension vector w and p-dimension vector v are independent white noises.

> **Interpretation:** The Hidden Markov Model uses the same LDS formulation of Kalman Filter, but its states take discrete values rather than continuous values.

The initial state Y_1 is generated as the following,

$$Y_1 = \mathrm{WTA}(\mathcal{N}(\mu_1, Q_1)). \tag{3.16}$$

The difference between the HMM and the Kalman filter models is the WTA operation. Indeed, I can construct an equivalence transformation from Equation 3.14 to Equation 3.12 by mapping $\mathrm{WTA}(A)$ to a state transition matrix T, where $T_{ij} = P(Y_{t+1} = e_j | Y_t = e_i)$. T can be computed easily given A and Q:

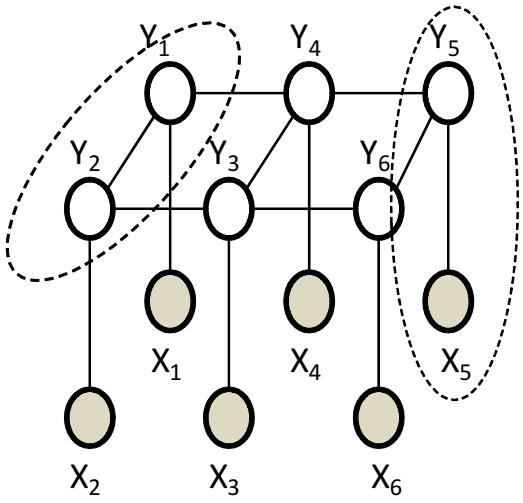


Figure 3.2: Graphical model of CRFs and M3N. X and Y correspond to local observations and their corresponding labels. The circles $[X_5, Y_5]$ and $[Y_1, Y_2]$ represent a unary feature and a pairwise Markovian feature.

T_{ij} is the probability assigned by the Gaussian whose mean is the i-th column of A (with covariance Q) to the region of k-space in which the j-th coordinate is larger than all the others. For any transition matrix T (whose rows sum to unity), there exist matrices A and Q such that the dynamics are equivalent. Similarly, the initial probability mass function for Y_1 can be computed from μ_1 and Q_1 .

For any noise covariance Q, the means in columns of A can be chosen to be any equivalent transition probabilities T_{ij}. I thus restrict Q to be the identity matrix and use only the means in columns of A to set probabilities. Equivalently, I can restrict $Q_1 = I$ and use only the mean μ_1 to set the probabilities for the initial state Y_1.

For HMM, the likelihood estimation and inference are performed with the forward (alpha) recursions; learning is conducted with the forward-backward (alpha-beta) recursions. The EM algorithm for learning is the exact well known Baum-Welch re-estimation procedure [18].

3.1.2 Relational Models

On the other hand, structural priors play crucial roles in many vision and natural language processing tasks, e.g. optical character recognition [151], object detection [52], and scene understanding [78]. Although the basic idea is to recognize an entity, the algorithm should consider beyond local observations and take the context into consideration. Please refers to Figure 3.2. Markov Random Fields (MRFs) have been

considered a natural model for incorporating such priors, however, it is trained in a generative way, which involves expensive computation.

Recent development in discriminative training techniques show prominent advantages over generative training approaches. For example, in Conditional Random Fields (CRFs) [107], relaxing the independence assumption by being conditionally trained gives significant performance improvement to discriminative models. Another state-of-the-art method, Maximum Margin Markov Networks (M3N) [172] incorporates large margin mechanisms into MRFs, making them very appealing. Theoretically, CRFs and M3N differ only in their loss functions. Both methods can be formulated in the same framework: structured linear discriminant function.

Let (X, Y^*) denote a pair of the local observation and its label. The goal of discriminative structure prediction can be thought of learning a W-parametrized linear discriminant function,

$$F(W, X, Y) =< W, \phi(X, Y) >, \tag{3.17}$$

where $\phi(\cdot)$ maps the pattern (X, Y) from the input space $\mathcal{D} = [\mathcal{X} \times \mathcal{Y}]$ to a feature vector $\phi(X, Y) \in \mathcal{R}^Q$; W is the weight vector in $\mathcal{R}^Q$. The definition of the feature representation ϕ depends on the application. With the discriminative function 3.17, the prediction rule is determined by,

$$Y^* = f(W, X) = \text{argmax}_{Y \in G(X)} F(W, X, Y), \tag{3.18}$$

where $G(X)$ enumerates various label configuration candidates for input X; the value of $F(W, X, Y)$ can be understood as a score evaluating the compatibility between the pair: X and Y.

> **Interpretation:** The objective above takes unary feature and pairwise state co-occurance into consideration, concurrently. The global optimization synthesized information from both perspectives to reduce ambiguities due to noisy observations.

The output not only labels individual entity but also explores meaningful internal structures within Y. Both Conditional Random Fields and Maximum Margin Markov Network are instances of such discriminative structured prediction framework.

3.1.2.1 Conditional Random Fields

CRFs first defines a conditional distribution over labels with function $F(W, X, Y)$,

$$p(Y|X, W) = \frac{1}{Z(W, X)} \exp\{F(W, X, Y)\}, \tag{3.19}$$

where $Z(W, X) = \sum_{y \in G(X)} \exp\{F(W, X, Y)\}$ is called the partition function. Given training set $\{X_i, Y_i^*\}|_{i=1}^N$, the parameters W can be learned by minimizing the following regularized log-loss,

$$W* = \text{argmin}_W \sum_{i=1}^{N} l_{crf}(i) + \frac{\lambda}{2}||W||^2, \tag{3.20}$$

where $l_{crf}(i) = -\log p(Y_i|X_i, W)$ and λ is a constant determining the trade-off between empirical risk and model complexity.

> **Interpretation:** Conditional Random Field implements the joint optimization framework (Equation 3.17) using a maximum likelihood criteria.

3.1.2.2 Maximum Margin Markov Network

The Maximum Margin Markov Network is a model of Support Vector Machines with structured output. Learning parameter W amounts to solving the following constraint quadratic optimization problem,

$$\begin{aligned} &\text{argmin}_W \sum_{i=1}^{N} \xi_i + \frac{\lambda}{2}||W||^2 \\ &s.t. \langle W, \Phi(X_i, Y)\rangle \geq e_i(Y, Y_i^*) - \xi_i, \forall Y \in G(X_i), \end{aligned} \tag{3.21}$$

where $\Phi(X_i, Y) = \phi(X_i, Y_i^*) - \phi(X_i, Y)$ and $e_i(Y, Y_i^*)$ is the hamming distance between the configuration Y and the true label Y_i^*. The hinge loss of M3N can be written as:

$$l_{m3n}(i) = \max_{Y \in G(X_i)} [e_i(Y, Y_i^*) - \langle W, \Phi(X_i, Y)\rangle]. \tag{3.22}$$

Hence, the constrained optimization in Equation 3.21 can be written as:

$$W* \quad = \quad \text{argmin}_W \sum_{i=1}^{N} l_{m3n}(i) + \frac{\lambda}{2}||W||^2. \tag{3.23}$$

Comparing Equation 3.20 and 3.23, it is easy to see that CRFs and M3N differ only in their loss functions. Both models have a regularization term, which is understood as Bayesian parameter estimation with Gaussian prior for CRFs and as large margins for M3N.

> **Interpretation:** Maximum Margin Markov Network implements the joint optimization framework (Equation 3.17) using a maximum margin criteria.

3.2 Evaluation Metrics

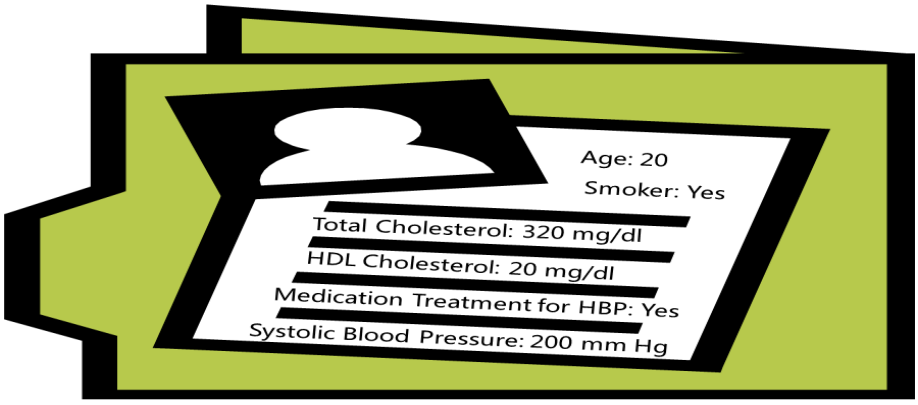
Predictive models are developed to assist clinicians to decide on treatment options and follow-up management [170]. Accurate prediction is critical so that treatment can be given to those individuals who are most likely to develop the disease [80]. In personalized medicine, the utility of predictive models is dependent on three parameters: sensitivity, specificity and reliability of predicted values.

Specificity and sensitivity are about drawing a decision boundary between positive and negative cases. Good specificity means superior performance of identifying sick individuals while specificity indicates the percentage of healthy people who are correctly identified as not having the condition. A reciprocal relationship exists between sensitivity and specificity. Thus, successful model should be highly specific without sacrificing sensitivity. While sensitivity and specificity are widely known measures for *discrimination*, the area under the ROC curve is used more frequently, as a one-number summary of *discrimination* [83]. Historically, supervised classifiers (e.g., Support Vector Machines and Decision Trees) were designed to provide a mapping between features and the outcome (represented by 0/1 class membership in many cases). These classifiers aimed to optimize the ranking of their outputs in a cohort study so that positive samples ranked higher than negative samples.

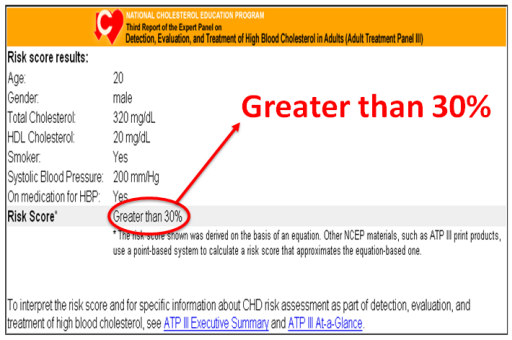
This objective is known as *discrimination*, which defines how well the predictive model can distinguish between two or more classes [12, 129, 168, 179]. A classifier with good *discrimination* ability, e.g. a large Area Under the ROC Curve (AUC) [83, 123], suffices to handle many learning tasks that require only decision boundary. However, it is not necessarily a good probabilistic model as it is not "calibrated". The

term "*calibration*" indicates how close the system estimates is to the "true" probability of a disease [37, 111, 176]. For example, Support Vector Machine defines a decision boundary to maximize the *discrimination* performance but its outputs are not calibrated, thus cannot be used as estimations of the "true probability".

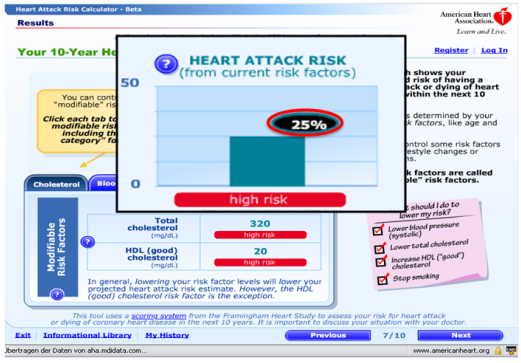
Due to the difficulty of defining what exactly constitutes the true probability in predictive models, *calibration* measures have received less attention than *discrimination*. Nevertheless, it is imperative that predicted probabilities reflect the true probability of disease as closely as possible [117], because only then can these predictions be used for individual risk assessment, which is a critical component to personalized clinical decision support. More importantly, each patient i should have his or her own probability parameter π_i, and not the probability parameter of the cohort π. Inconsistency between a model's outputs and true probability could significantly reduce the model's reliability for decision makers.



(a) A made-up medical profile.



(b) Output from the NHLBI's risk assessment web-tool.



(c) Risk estimated by the American Heart Association's online tool.

Figure 3.3: Lack of *calibration* can cause inconsistent risk predictions. The same patient got different risk scores from different online tools due to lack of *calibration*.

Figure 3.3 demonstrated a case in which the same made-up patient record got different heart attack risk

scores from different online risk estimation systems. The inconsistency provides yet another motivation for calibrating probabilistic outputs.

Ideally, a predictive model for supporting clinical decision making requires outputs to be indistinguishable from "true probability" of an individual patient being diagnosed. That is, predictions useful in supporting clinical decisions should be "calibrated". We all know now that diagnosis or treatment good for a population in general might not work for individual patients in this diagnostic group. In terms of personalized clinical decisions, it is important to count what matters the most for individual patients to estimate their own risks. However, this is a nontrivial task.

A few difficulties complicate the investigation of "ideal" probabilistic models: first, there is no clear definition of what constitutes "true probability"; second, we need a good measurement of how well a probabilistic model "calibrates" [111]. Unfortunately, there is no one number summary of a model's *calibration* ability such as AUC for *discrimination*.

Recent research in machine learning has shown the benefits of calibrating predictive models, which becomes especially important when probability estimates are used for clinical decision making [140, 147, 205, 206]. In summary, the quality of probabilistic predictive models depends on two major indices: *discrimination* and *calibration*. Importantly, the success of personalized clinical decision support systems depends on a comprehensive consideration of both metrics.

3.2.1 Discrimination

Discrimination (also known as resolution or refinement) is the ability of the probabilistic model to correctly separate co-variate vectors into two sets corresponding to observed outcomes, which take dichotomous outcome values of 0 or 1.

One popular *discrimination* measurement is a receiver operating characteristic (ROC) curve, which is a graphical plot of sensitivity (true positive rate) vs. 1-specificity (false positive rate) [108]. The ROC curve was first developed under the signal detection theory for a binary classifier system where its *discrimination* threshold is varied. The curve compares two operating characteristic (True Positive Rate vs. False Positive Rate).

Every point on a ROC curve corresponds to a unique pair of True Positive Rate (TPR) and False Positive Rate (FPR), as indicated by Figure 3.4. The Area Under the ROC Curve (AUC) [83, 181] provides a one number summary to evaluate a probabilistic model's ability to discriminate between positive and negative

co-variate vectors at various discriminative thresholds. Concretely, the AUC can be expressed by integrating TPR over FPR:

$$\begin{aligned} AUC &= \int_0^n (TPR)\, d(FPR) \\ &= \frac{1}{nm} \sum_{X \in \{+\}} \sum_{O \in \{-\}} (P(X) > P(O)), \end{aligned}$$

where $P(X)$ and $P(O)$ correspond to the posterior probability of a positive sample X and a negative sample O, respectively. The values $\{+\}$ and $\{-\}$ indicate the positive and negative observations, respectively. The quantities m and n correspond to the cardinality of the positive and negative classes. Figure 3.4 illustrates the relationships between ROC, AUC and its calculation.

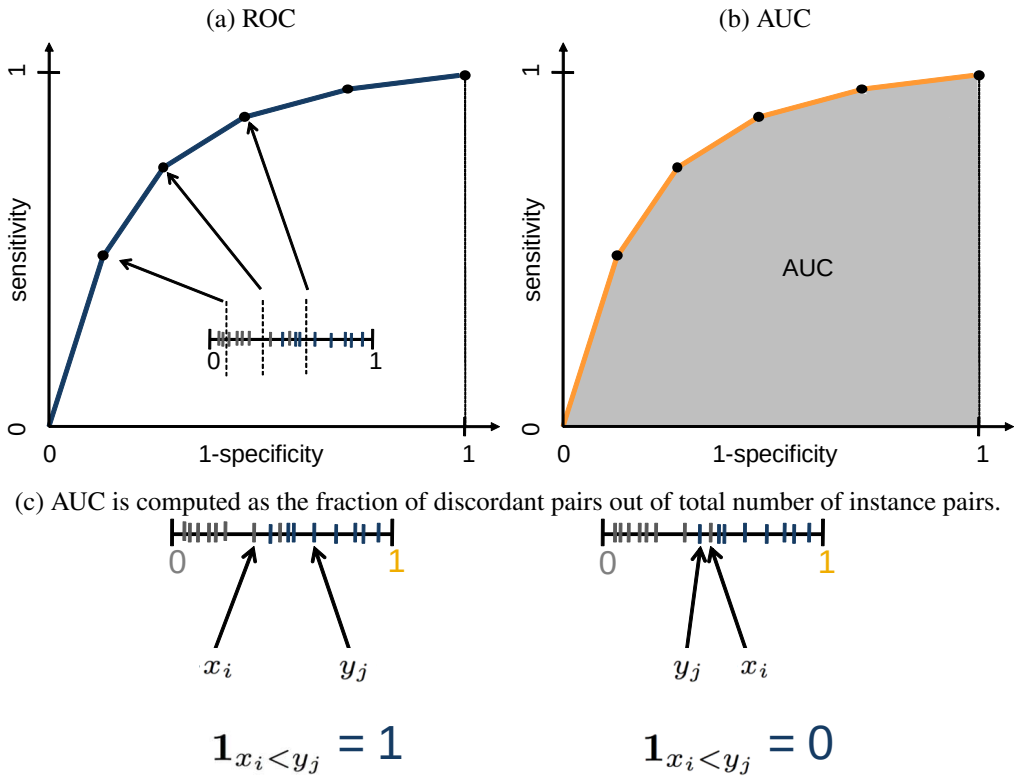


Figure 3.4: ROC, AUC and the calculation.

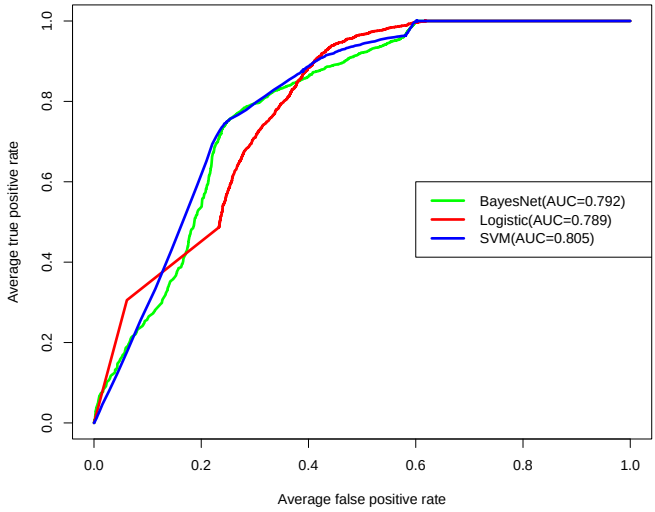


Figure 3.5: Demo plots of ROC curve for different models: Bayes Net, Logistic Regression and Support Vector Machine. Models are trained using synthetic data to illustrate the curves for illustration purposes.

A random classifier corresponds to an AUC of 0.5; a model that perfectly separates two classes of covariate vectors corresponds to an AUC of 1. For illustration purpose, I plotted ROC curves for three different models: Bayes Net, Logistic Regression and Support Vector Machine in Figure 3.5. AUCs are shown in the legend next to their names.

> **Interpretation:** The area under the ROC curve (AUC) evaluates model discrimination by calculating the probability that among all possible pairs of individuals with two different outcomes, the predicted value for the one with positive outcome is higher than for the one with negative outcome. [41]

However, this index is a global measure. Recent research in biomedical informatics has shown the insufficiency of this metric [12, 176]. With numerical simulations, Pepe et al. [146] demonstrated the relation between association (measured in odds ratios) and classification, depicted by AUC, and concluded that the statistical significance in association with itself does not characterize the discriminatory capacity.

3.2.2 Calibration

Supervised learning often concentrates on a learning problem to infer a function that separates given instances to discrete categories they belong to. However, it is often desirable to output the probability of an

instance belonging to a particular class. Ideally, these probabilities estimate converge to the true underlying probability distribution, namely, the proportion of events in a group of cases that have an estimate of p is exactly p. If a model predicts that it is going to rain on a particular day with 10% probability, then users should expect that there is genuinely a 10% chance that it is going to rain.

For a binary outcome case, that is, whenever a classifier outputs a probability estimates of instance s, the fraction of times this instance in question is positive is roundly s. This problem is often know as "*calibration*", another important probabilistic model evaluation criteria in addition to "*discrimination*". Well-*calibrated* probabilities are useful whenever users want to make not just classification, but actions. Intuitively, *calibrated* probability estimates can be interpreted as genuine frequency-based estimates.

More formally, *calibration* (also called reliability) is how well probabilistic outputs numerically correspond to the observed outcomes. Let us, for example, denote a probabilistic classifier that assigns a probability p to every sample i. That is:

$$y = \text{class label}, \; p = \text{estimated probability}.$$

Mathematically, this intuition can be can be defined as follows:

$$\rho_N^{\epsilon}(p) = \frac{\sum_{i=1}^{N} y_t \mathbb{I}[p_i \in (p-\epsilon, p+\epsilon)]}{\sum_{i=1}^{N} \mathbb{I}[p_i \in (p-\epsilon, p+\epsilon)]}, \tag{3.24}$$

($\mathbb{I}[\cdot]$ is the indicator function.) In other words, $\rho_N^{\epsilon}(p)$ is the empirical frequency of the outcome on just the rounds when the predictions were "roughly" p. This frequency is expected to be close to p. Thus, a forecaster is calibrated if every p and every ϵ,

$$\lim \sup_{N \to \infty} |\rho_N^{\epsilon}(p) - p| \leq \epsilon.$$

holds.

> **Interpretation:** *Calibration* (Equation 3.24) counts the frequency of outcomes when its values is "roughly" p.

In practice, when there are not many samples with the same estimated probabilities, samples with similar estimated probabilities $\rho_i^{\epsilon}(p)$ are grouped for evaluation by partitioning the sample set into groups (or bins). To estimate the unknown true probabilities, I divided the prediction space is into a number of bins. These can be determined by fixed thresholds (e.g., cases with predicted value between 0 and 0.1 fall in the first bin and between 0.1 and 0.2 in the second bin.), or by percentiles (e.g., deciles). For each bin, the mean

predicted probability is plotted against the observed fraction of positive cases. If the model is well calibrated, the points fall near the diagonal line. This kind of plot is known as a reliability diagram [33], as shown in Figure 3.6.

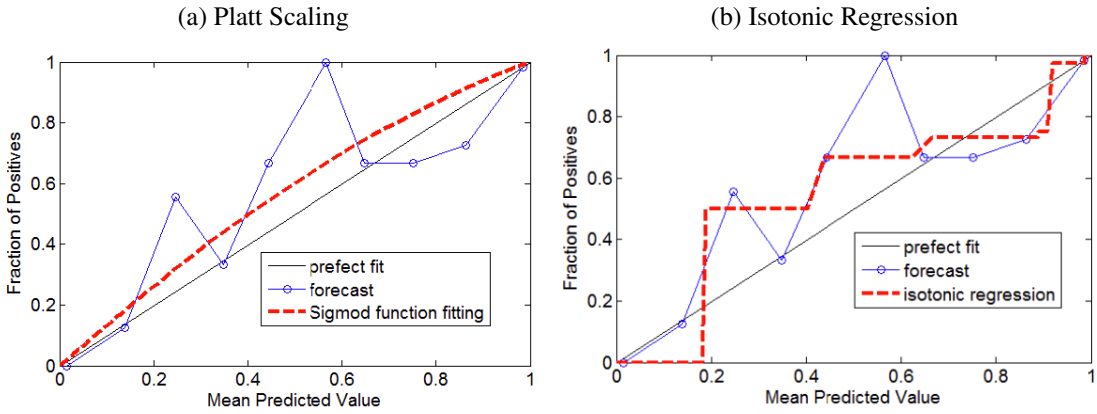
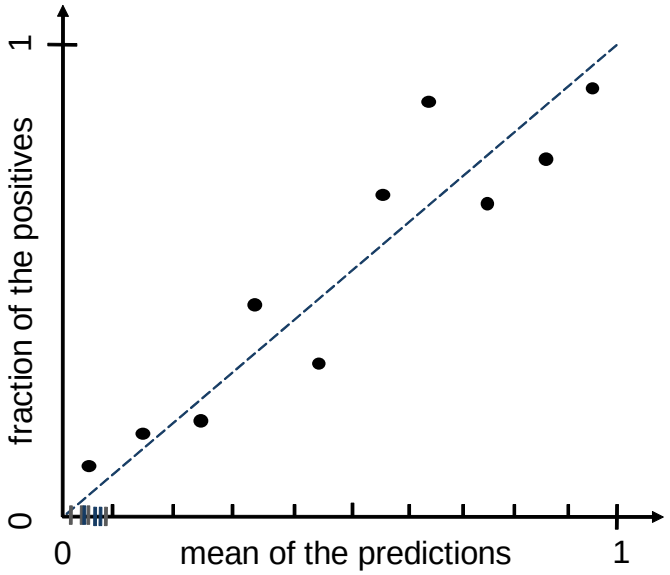


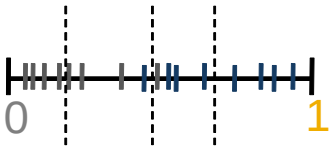
Figure 3.6: Reliability diagrams for two *calibration* approaches. Subfigure (a) corresponds to Platt Scaling and the subfigure (b) corresponds to Isotonic Regression. The blue circles are connected to help visualization; the red dotted lines are the results for the *calibration* algorithms. A perfectly calibrated classifier must generate predictions that lie on the 45 degree diagonal line.

Statistically, *calibration* is also defined synonymously to goodness of fit. Such accessing goodness of fit can be conducted in checking model assumptions. A widely used goodness-of-fit statistic is the Hosmer-Lemeshow test (HL-test) [90, 103]. Although HL-test has important limitations, few alternatives have been proposed. In addition, most of these alternatives are model-specific *calibration* measurements, which makes them unattractive for evaluating probabilistic outputs across different models. For this reason, I decided to use the HL-test as my *calibration* evaluation measurement in this thesis.

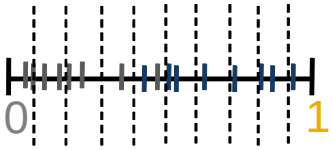
The HL-test statistic can be written as: $H = \sum_{g=1}^{10} \frac{(O_g - E_g)^2}{N_g \pi_g (1-\pi_g)}$, where O_g, E_g, N_g and π_g correspond to observed positive events, expected positive events, number of total observations, and predicted risk for the g^{th} risk deciles, respectively. H is called the Hosmer-Lemeshow H test statistic if deciles are defined as equal-length subgroups of fitted risk values; otherwise, H is called the Hosmer-Lemeshow C test statistic if deciles are defined as equal-size subgroups of fitted risk values. The distribution of the statistics H is approximated by a chi-square with eight degrees of freedom. Figure 3.7 illustrates the relationship of reliability diagram and two types of HL-test.



(a) Reliability diagram provides a visual evidence of goodness-of-fit. The averaged point to line distance roughly indicates how well a model is calibrated.



(b) HL-C test.



(c) HL-H test.

Figure 3.7: Reliability diagrams and two types of HL-test.

Chapter 4

Co-Estimating Hidden States in Predicting Bio-terrorism Related Outbreaks

[1]Bioterrorism related outbreak prediction is receiving more and more attentions since the 9/11 attack [34, 97, 134, 184, 190]. Since then, the country channeled enormous financial resources in bioterrorism research: in 2001, for example, the annual budget for the National Institute of Allergy and Infectious Disease (NIAID) in Bethesda, Maryland— the division of the National Institutes of Health that carries biodefence and infectious-disease research — was $42 million. By 2002, it had escalated to $187 million, a 345% increase [97]. Despite many efforts devoted to fuse the public health information and social networks [99], our abilities to successfully detect, monitor and foresee bioterrorism-related disease outbreaks are not yet sufficient.

The ability to foresee bioterrorism attacks and their impact on the public is necessary to ensure the efficacy of responding from the planning and preparation perspectives [46, 114, 154, 190]. To tackle this problem, the first step is to determine the manifestation process of large scale disease spread. Predictive models can be developed on the basis of the manifestations of the disease, to assist in the decision making process of responding to and controlling it.

A widely applied epidemiological forecasting method called susceptible-infected-recovered (SIR) mod-

[1]A version of this chapter has been published in ECML'10 and Lecture Notes in Computer Science [92].

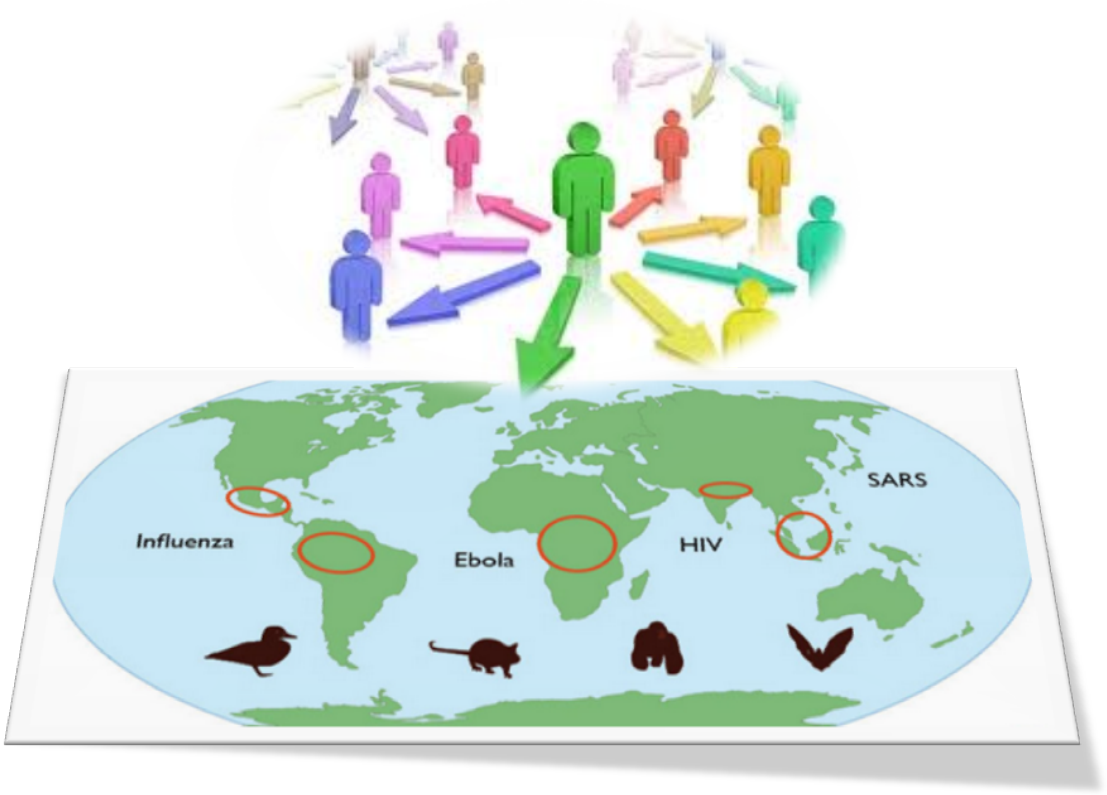


Figure 4.1: Infectious disease spreads through human networks over time.

els manifestations of disease process and suggests response strategies [43, 77, 161, 175, 208]. The model categorizes the entire population into three groups – susceptible, infected, and recovered. Individuals in each group are assumed to have the same states and SIR uses predefined transition probability to model the disease progression.

However, a "population-based" disease progression processes model like SIR assumes a homogeneous mixing of individuals. Its prediction over a network assumes that contacts between individuals are fixed, at least for the duration of an outbreak, however, in reality, contact patterns may be quite fluid, with individuals frequently making and breaking social or sexual relationships [182]. We also know that symptoms or manifestations for one member of a diagnosis population might not be the same for all members. SIR model thus lacks the ability to adequately capture disease transmission on dynamic networks in which each individual has a characteristic while identities of their contacts change in time [39].

Recent studies of aviation influenza have confirmed that large-scale diseases quickly spread over a net-

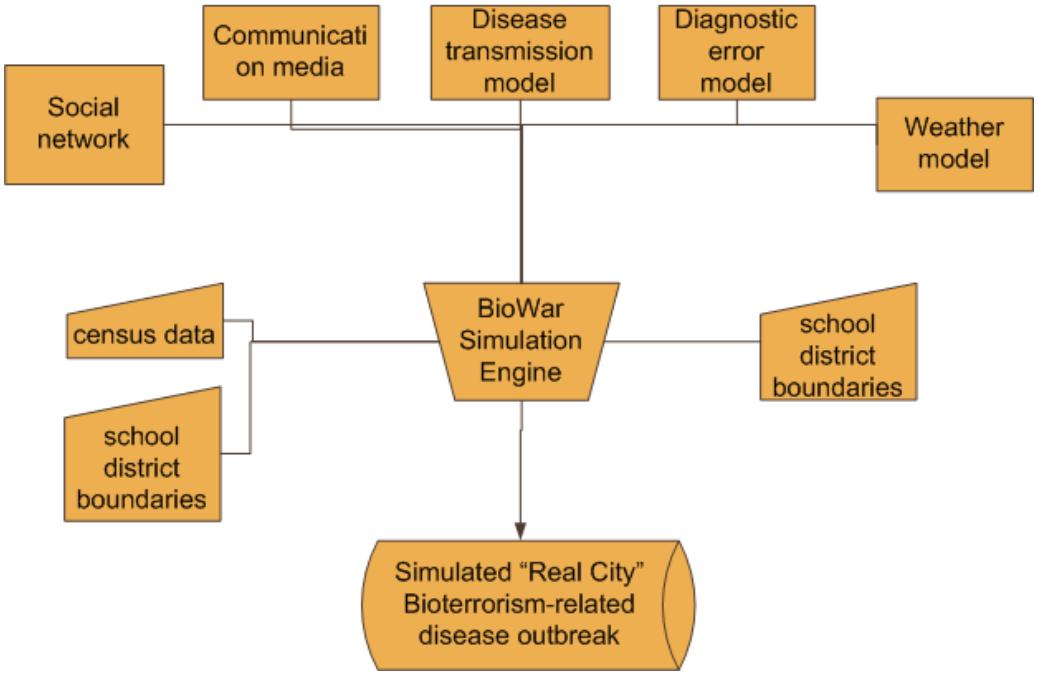


Figure 4.2: BioWar simulation engine.

work of people [48, 76, 149, 150], as illustrated in Figure 4.1 Actually, the spread of disease is a complex process involving multiple aspects of social life, i.e., social networks, communication media, demographic information, urban spatial structure, and weather conditions. Recently, a more sophisticated agent-based simulation engine was developed to take into account of above mentioned aspects concurrently with an advanced diagnostic model to determine the disease process at a much finer scale than SIR. This engine, known as BioWar, was developed by CASOS lab at CMU (`www.casos.cs.cmu.edu/projects/biowar/`) to provide a single integrated interface that simulates the impact of a bio-terrorist attack in a U. S. city. As opposed to traditional methods that model hypothetical cites, the BioWar engine used real city data like census, school track, and other public available sources to output various manifestations of the disease as the simulated agents go about their lives, as illustrated in Figure 4.2. These manifestations include daily observations like doctor visit rate, school attendance, and weblookup rate. The validity of the simulation model has been successfully confirmed by a number of previously published articles [30, 31, 32, 38].

In traditional supervised learning methods, the goals are to separate instances into discrete categories. These methods aim to optimize the 0/1 loss function so that the empirical mistakes on the training set minimized. Among these approaches, Support Vector Machine (SVMs) [180] has demonstrated impressive

success on a broad range applications such as image segmentation [89, 202], text mining [55, 109, 210] and many more. These successes have been achieved to a large extent on the ability of the kernels [17, 159], which can map low-dimensional inputs to a very high dimensional space where the separations become easier. In addition to many empirical success, SVMs have strong generalization guarantees, derived from the margin-maximizing properties of the learning algorithm [25, 56, 63].

However, many supervised learning applications contains rich contextual information in addition to individual manifestations. These contextual information are usual reflected in relationships with time and space. For example, it might be easier to label a set of correlated instances such as optical character recognition, part of speech tagging and video segmentation, which involves labeling an entire sequences of instances into discrete non-exclusive categories. An easy solution, which is often conducted in practice, is to treat this sequence of labeling tasks as a group of independent tasks, handling each of them separately. This approach, however, fails to exploit significant correlation information, and is often insufficient to meet the performance requirements. Typical approaches include Hidden Markov Network [152] that learns temporal correlated states and Conditional Random Fields [107] that learn relational dependent states. These approaches use probabilistic inference algorithms, i.e. Balm-Welch's backward and forward, to jointly assign manifestations with their most likely states. That is, they are capable of modeling correlations between different labels, often outperforms ad-hoc methods that classify instances separately. Unfortunately, these probabilistic models are not generalizable to more complex situations where both temporal and relational dependency are involved. Moreover, graphical models cannot provide generalization bounds that maximum-margin classifiers could offer.

The above discussion is closely related to the bioterrorism-related disease outbreak problem, which involves observing manifestations of different sources over a continuous period. The objective is to estimate hidden states that drive these manifestations to assist decision makers in emergency response and disaster preparedness.

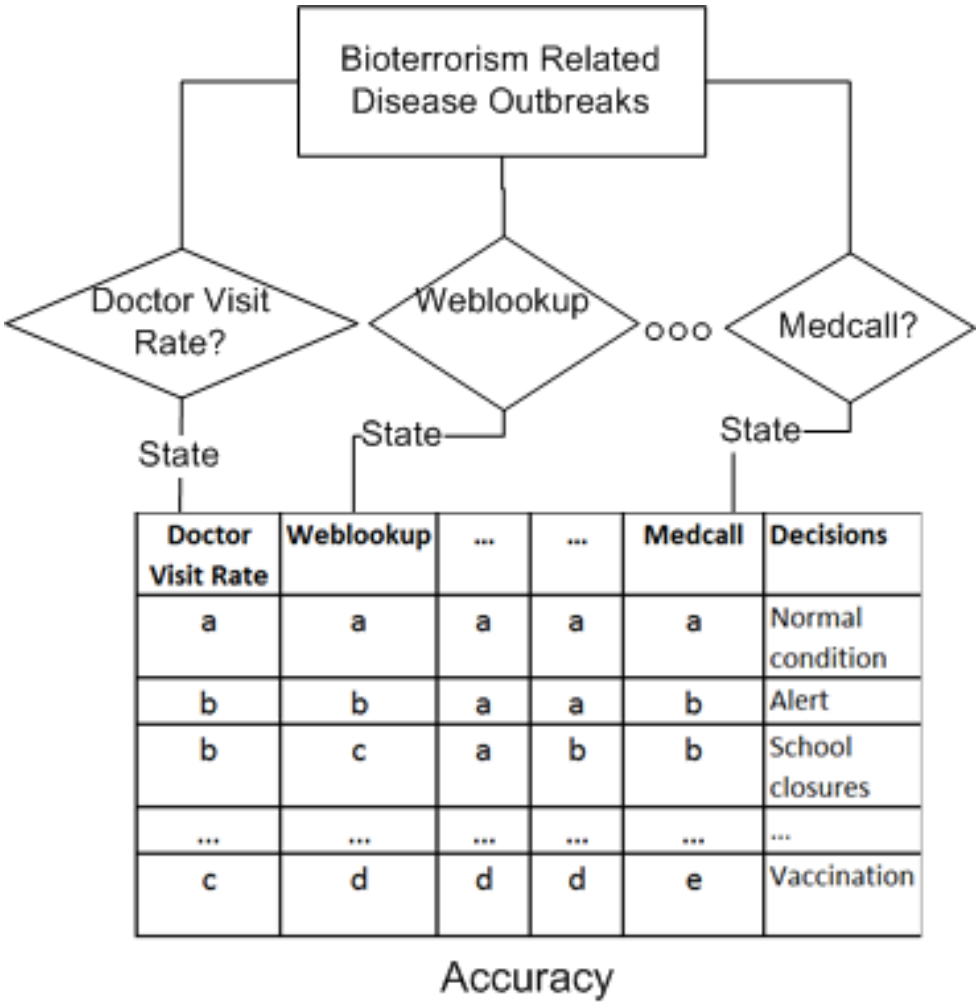


Figure 4.3: An example of using co-estimated variables to support emergency responses of disease outbreaks.

Figure 4.3 illustrates an example of using co-estimated states to support emergency responses. Interventions are suggested by not a single state but a set of estimated states together. Obviously, the effectiveness of interventions relies on how accuracy the state estimation is. Unfortunately, neither modeling manifestations from a single perspective over time (HMM model [152]) nor modeling all manifestations at one time tick (CRFs [107]) can provide highly accurate state predictions in bioterrorism-related disease outbreaks. Combining manipulations over temporal and spatial aspects could offer a more comprehensive understanding of the disease processes by reducing the ambiguity in individual instances if they are considered separately. However, a direct solution of this joint optimization problems could be computationally infeasible (i.e., ex-

ponentially more expensive than that of the individual optimization). To this end, I focused on developing tractable temporal and relational prediction frameworks to model manifestations of simulated avian influenza using the BioWar engine. The new framework called Temporal Maximum Margin Markov Network to consider factorizable temporal dependency and relational correlation concurrently for better performance.

4.1 Motivation

The public fear of potential Bio-terrorism attack draws our attention to an old problem: predicting the spread of infectious diseases on a large scale [19, 24, 46]. Accurate prediction of manifestations of the disease process helps the policy makers to intervene in a more timely and appropriate manner in various situation, e.g., resource allocation, vaccination, and school closures. For example, if a reliable prediction model indicates that the hospital inpatient rate will increase sharply, health officials can devote more attention and resources to get better prepared [46].

Intuitively, when people people interact closely, their chances of becoming infected by a contagious disease increases. However, the spread of diseases, especially the airborne ones, is linked with multiple environmental factors, and it is difficult to predict future disease outbreaks due to the evolving nature of threats and vulnerabilities [3]. On a city scale, individuals interact with one another with a large degree of randomness. Diseases spread through the social and activity networks over time, which further complicate the modeling process.

To model the complex nature of disease outbreak more faithfully, I developed a more comprehensive approach. As opposed to the traditional single-target modeling (e.g., number of people at work, number of doctor visits, web look up rates and inpatient rates), my approach provides a concurrent prediction of multiple hidden factors that are linked to the same disease outbreak. The benefit of introducing concurrent co-estimation is twofold: improving the overall prediction accuracy by leveraging mutual correlations among various hidden factors and predicting the ranking of the hierarchy of contributing factors.

The first benefit is obvious as it provides early detection, which helps decision makers. The latter is useful to the human experts in generating "profiles" of different disease outbreak stages. The main contributions of this study are:

- I propose a novel symbolic representation (SAX+ coding) to model different but correlated continuous outcome variables.

The approach aims at finding the best trade-off between quantitative methods that are precise and formally rigorous and qualitative methods that are intuitive and easy to understand. My approach maps various sources of continuous observations to discrete-valued states (symbolic representation) at comparable scales, and thus makes semantic correlations among heterogeneous observations more meaningful and provides an intuitive interface to watch concept drifts that relate to adverse events.

- I improve the overall prediction accuracy by inferring a relational network of the states, concurrently.

My approach differs from previous studies that predict individual states. I advocate predicting relational-related states simultaneously to best utilize mutual correlations between different perspectives of collected observations. For example, a sharp increase of flu-related keywords on Weblookup indicates a potential flu outbreak, which increases the probability of doctor visiting rates. Predictions of the latter state (doctor_visit_rate), if predicted together with the former state (Weblookup rate), could be improved over those on its own time series.

Because no bio-terrorism related attacks have been observed in U.S., I will use simulated data from the BioWar engine to validate my model. Please refer to Chapter 2 for the details about the data and BioWar simulation engine. The rest of this chapter is organized as follows: Section 4.2 describes the data; Section 4.3 discusses the related works; Section 4.4 conducts preliminary data analysis and describes my novel state representation approach; Section 4.5 discusses my machine learning model including the learning and prediction algorithms; Section 4.7 presents the result of using synthetic data and two simulated BioWar datasets. I also evaluate the generalization ability of the ML-Model in this section. Finally, I conclude the this chapter.

4.2 Data

I use BioWar-I simulation data in this chapter. The data contain multiple outcome variables collected over a continuous period of time. The data incorporate both relational and temporal information. Please refer to Chapter 2 for the origin of this data. The following table summarizes outcome variables and their statistics.

Table 4.1: BioWar-I summary: min, mean, median, and max for each variable.

tick	dayOfWeek	month	day	dead	is.er
Min. : 0.0	Fri:312	Aug : 186	Min. : 1.00	Min. : 0.000	Min. : 0
1st Qu.: 547.2	Mon:312	Dec : 186	1st Qu.: 8.00	1st Qu.: 0.000	1st Qu.: 0
Median :1094.5	Sat:312	Jan : 186	Median :16.00	Median : 0.000	Median : 7
Mean :1094.5	Sun:318	Jul : 186	Mean :15.72	Mean : 4.338	Mean : 696
3rd Qu.:1641.8	Thu:312	Mar : 186	3rd Qu.:23.00	3rd Qu.: 0.000	3rd Qu.: 13
Max. :2189.0	Tue:312	May : 186	Max. :31.00	Max. :97.000	Max. :19401
	Wed:312	(Other):1074			
kidsAtHome	adultsAtHome	at.work	weblookup	medcalls	num.exchanges
Min. :20959	Min. : 37447	Min. : 0	Min. : 0.0	Min. : 0	Min. : 0.0
1st Qu.:31566	1st Qu.: 86910	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0	1st Qu.: 0.0
Median :78695	Median :217960	Median : 0	Median : 8.0	Median : 0	Median : 0.0
Mean :59889	Mean :151609	Mean : 41487	Mean : 695.9	Mean : 18867	Mean : 101.9
3rd Qu.:78701	3rd Qu.:217985	3rd Qu.: 0	3rd Qu.: 15.0	3rd Qu.: 0	3rd Qu.: 0.0
Max. :79497	Max. :226684	Max. :187787	Max. :19043.0	Max. :141408	Max. :11438.0
in.hospital	is.home	is.work	is.school	is.pharmacy	is.doctor
Min. : 0	Min. : 58563	Min. :0	Min. : 0	Min. : 0.0	Min. : 0.0
1st Qu.: 0	1st Qu.:118408	1st Qu.:0	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0.0
Median : 7	Median :296655	Median :0	Median : 0	Median : 0.0	Median : 0.0
Mean : 696	Mean :211498	Mean :0	Mean : 9277	Mean : 557.6	Mean : 125.9
3rd Qu.: 13	3rd Qu.:296683	3rd Qu.:0	3rd Qu.: 0	3rd Qu.: 13.0	3rd Qu.: 3.0
Max. :19401	Max. :306181	Max. :0	Max. :58370	Max. :25178.0	Max. :10315.0
is.stadium	is.theater	is.store	is.restaurant	is.university	is.military
Min. : 0	Min. : 0	Min. :0	Min. : 0	Min. :0	Min. :0
1st Qu.: 0	1st Qu.: 0	1st Qu.:0	1st Qu.: 0	1st Qu.:0	1st Qu.:0
Median : 0	Median : 0	Median :0	Median : 0	Median :0	Median :0
Mean :1437	Mean : 3997	Mean :0	Mean : 31342	Mean :0	Mean :0
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.:0
Max. :7408	Max. :20286	Max. :0	Max. :157115	Max. :0	Max. :0

Most outcome variables are self-explanatory. To see how they are distributed, I used histograms of these variables to illustrate their distribution in Figure 4.4.

The following figure shows 20 outcome variables. It is easy to see that most clinical related outcome variables like death, adults_at_home and is-pharmacy have an outbreak during the observation period.

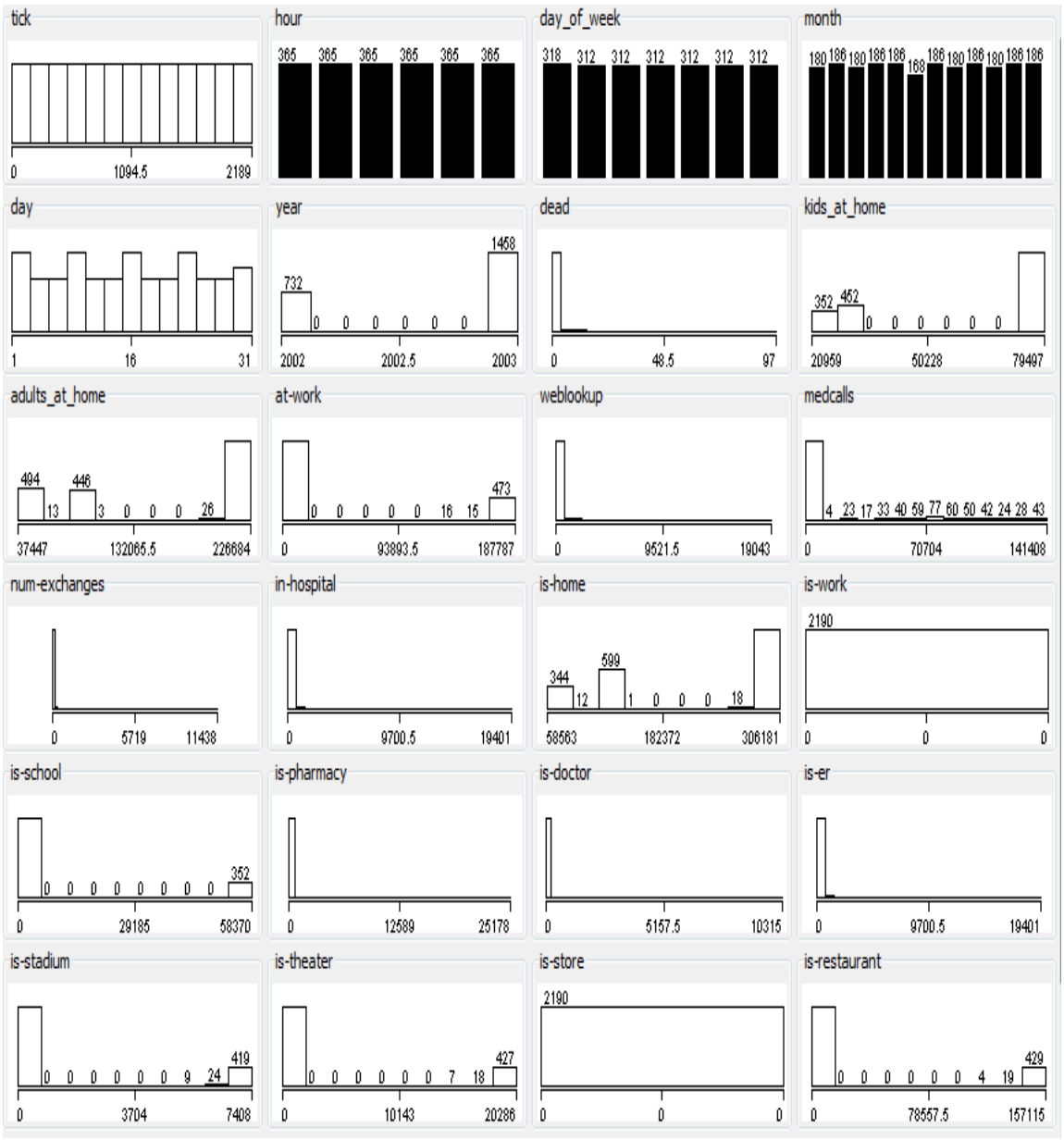


Figure 4.4: Histograms of outcome variables in the BioWar-I data.

To get a better understanding of the interaction between these outcome variables, I plotted co-variable occurrence in Figure 4.5 using matrix plot. The diagonal cells of this matrix lists variable names, and the other cells demonstrate the co-occurrence. Several variables demonstrate strong correlations, as indicated in the Figure 4.4 above.

To explore these correlations, I chose a few pairs from outcome variables to plot their grouped probability

density functions (PDFs), as indicated in Figure 4.6.

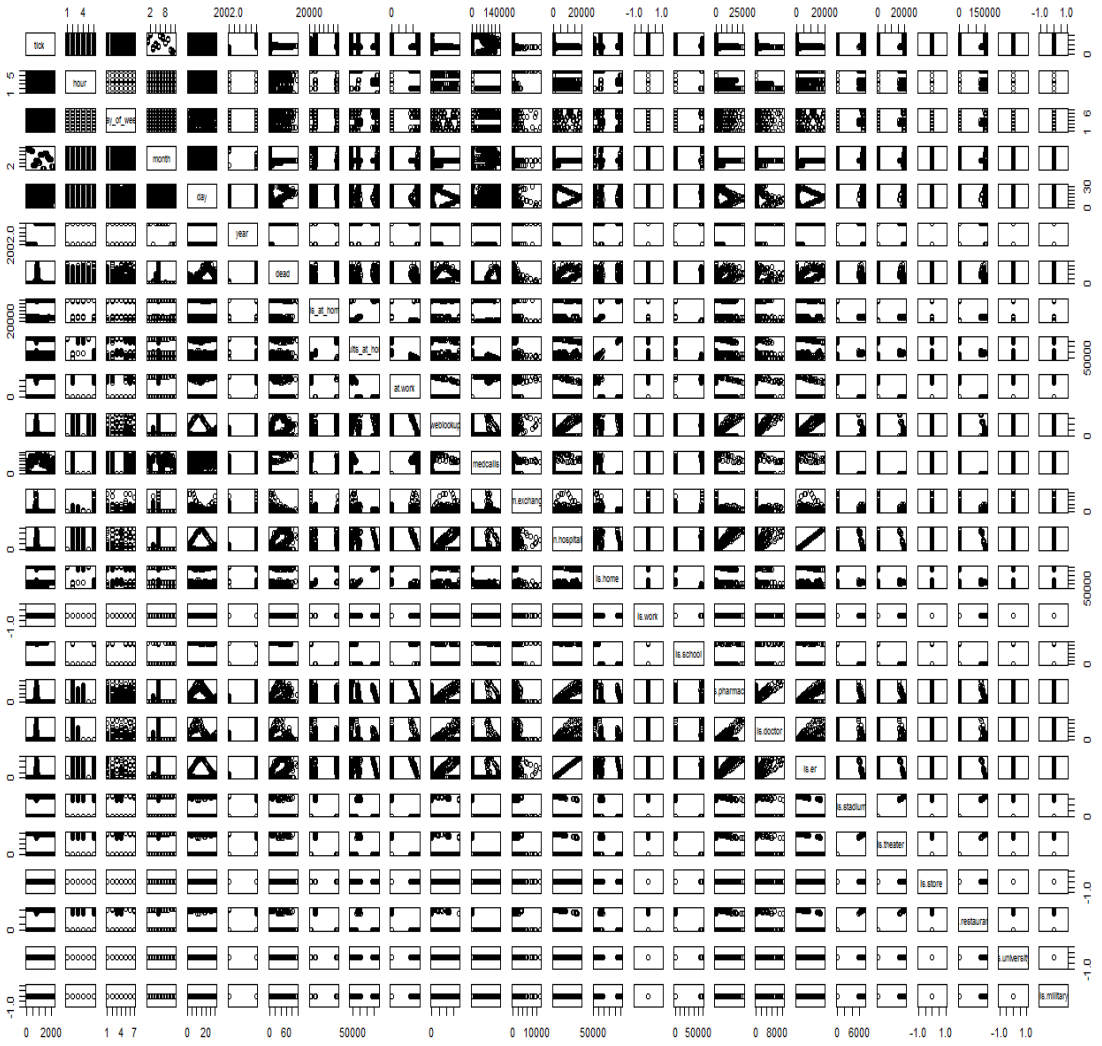


Figure 4.5: BioWar-I matrix plot.

Figure 4.6 shows four pairs of grouped PDFs. There are obvious correlations between these variables. However, some correlations occur across various levels, and others are prominent only in a few levels of the outcome variable “dead”.

Alternatively, I plotted the histogram of pairwise co-variables aggregated in time for a better understanding of outcome variable correlations over time. The following four figures demonstrate the temporal correlations among pairwise co-variables. This dataset shows strong feature correlations over time, and thus

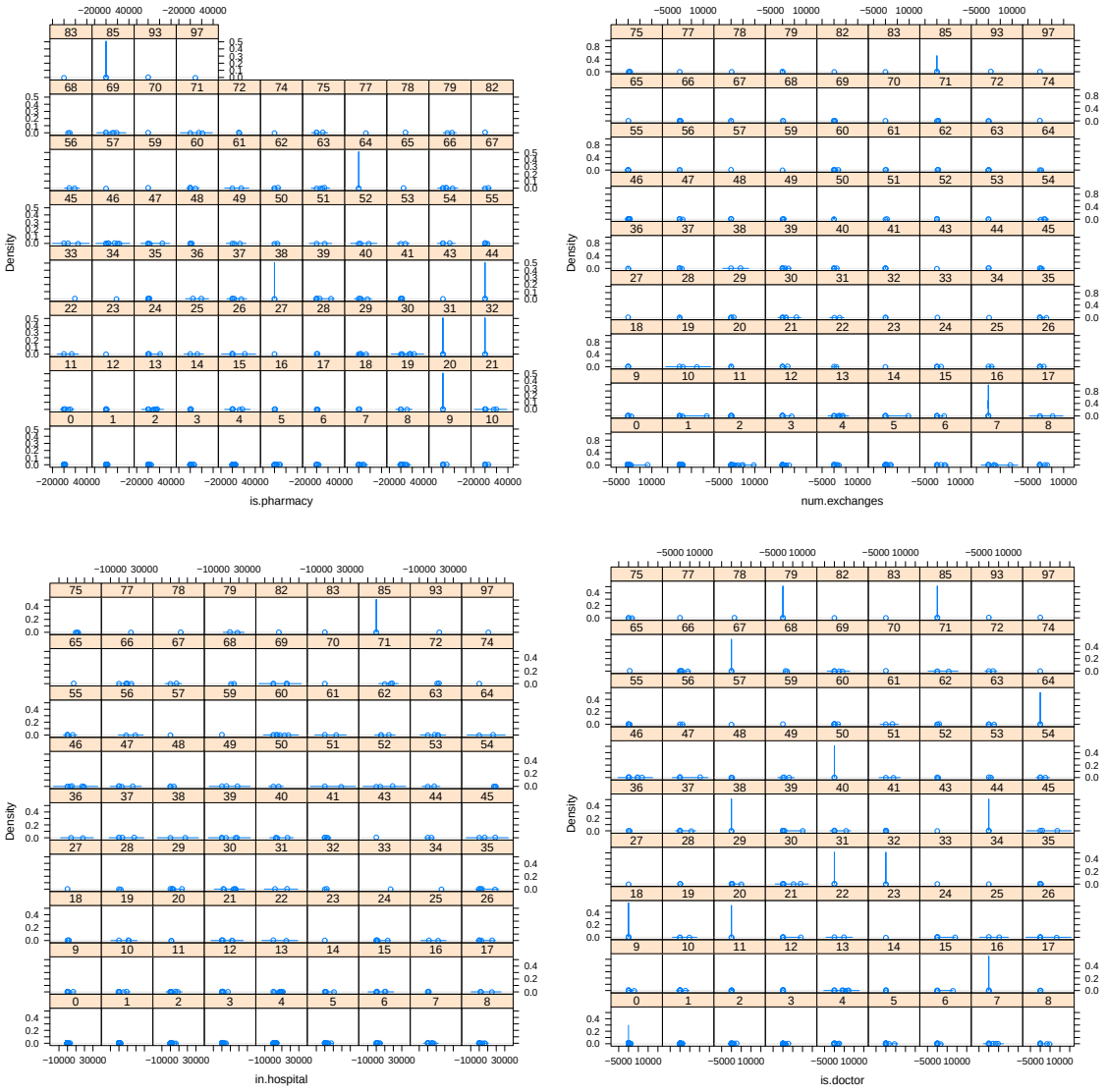


Figure 4.6: Grouped density plot for various output variables of BioWar-I. For these sub-figures, I plot probability density function of "is.pharmacy", "num.exchange", "in.hospital" and "is.doctor" at 98 levels the outcome variable "dead."

becomes a good fit to test the model developed in this chapter, which is capable of co-estimating multiple variables over time.

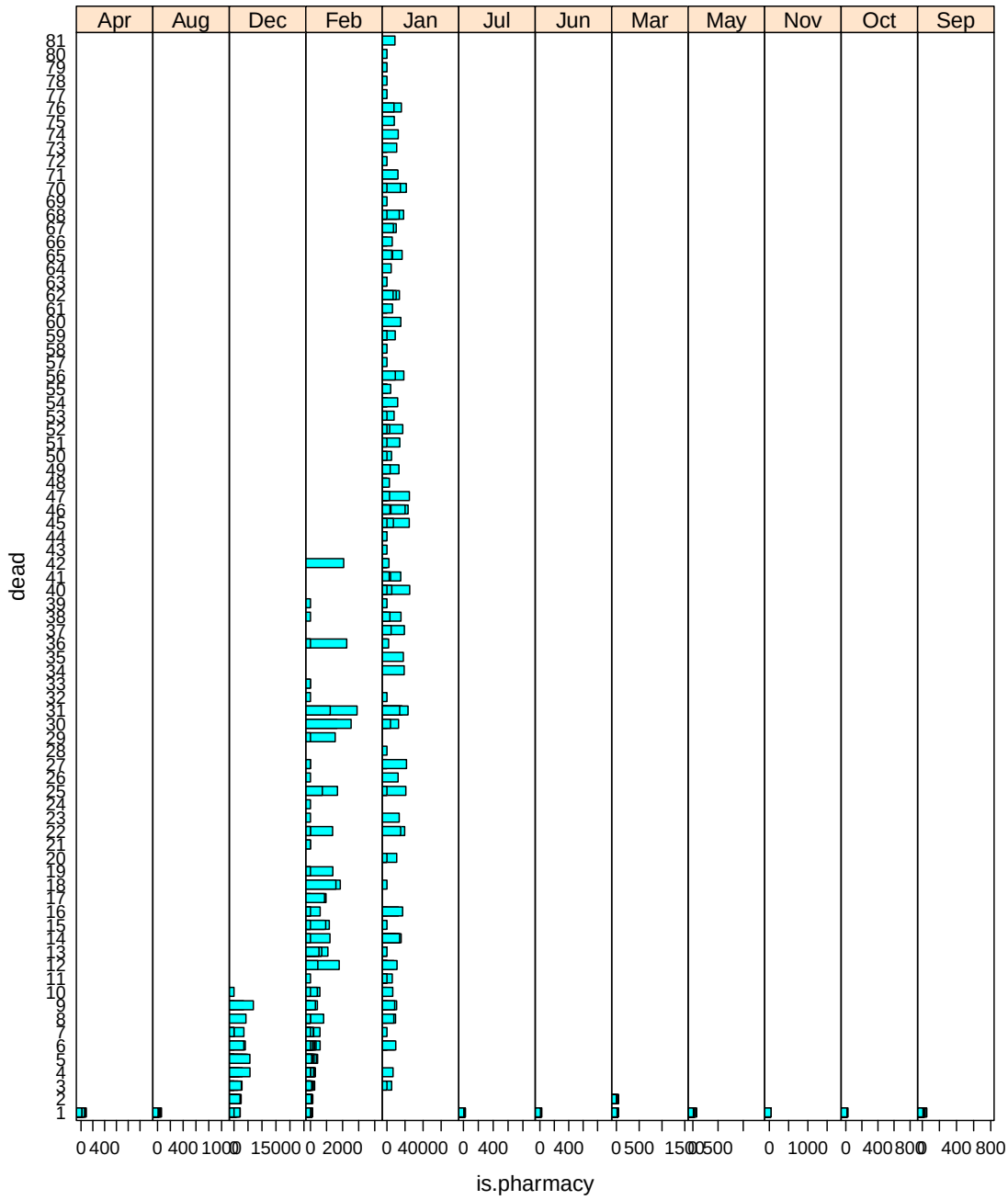


Figure 4.7: Grouped histograms for two outcome variables “dead” V.S. “in.pharmacy”. I aggregate monthly observations to plot twelve figures of histograms.

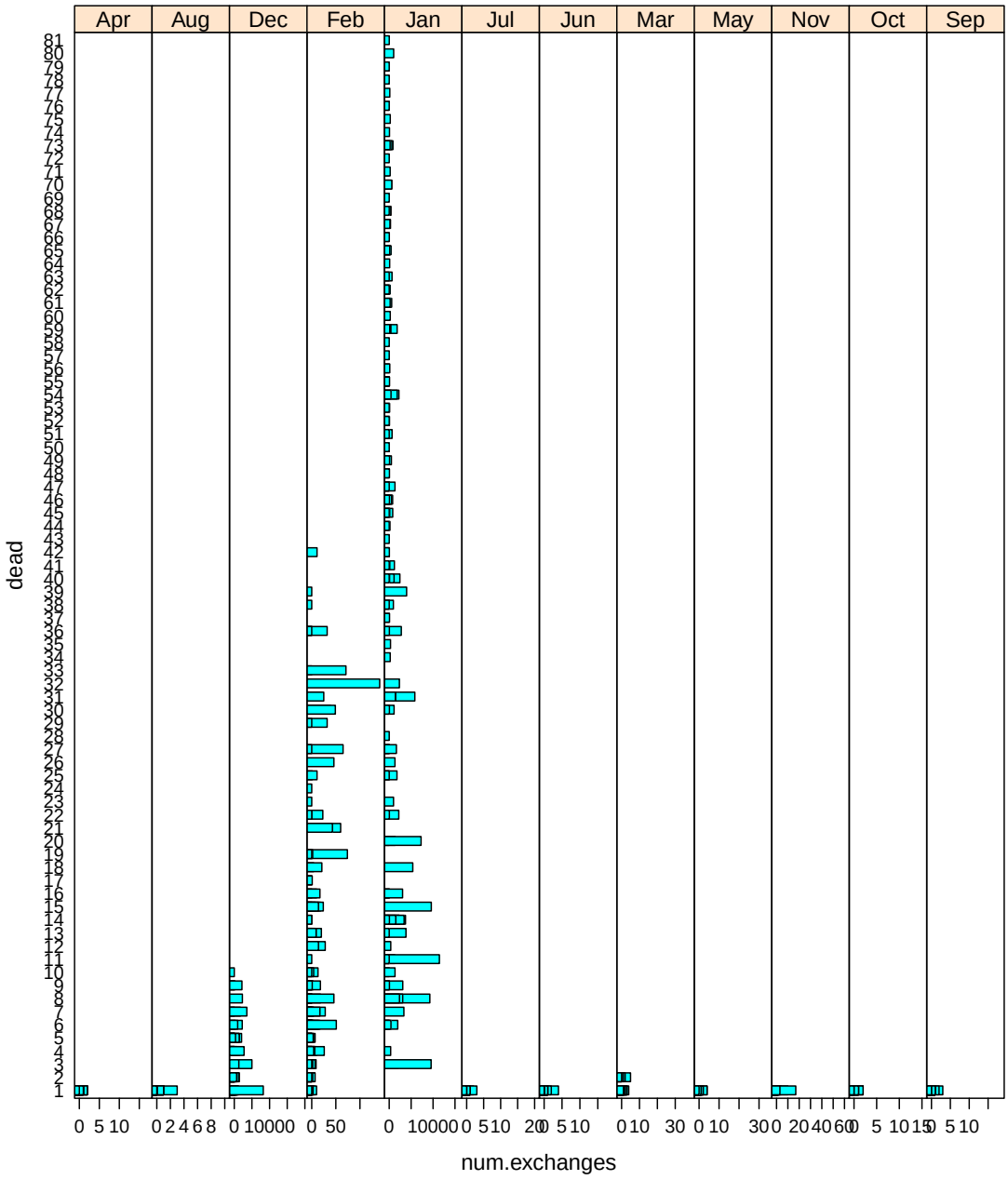


Figure 4.8: Grouped histograms for two outcome variables “dead” V.S. “num.exchanges”. I aggregate monthly observations to plot twelve figures of histograms.

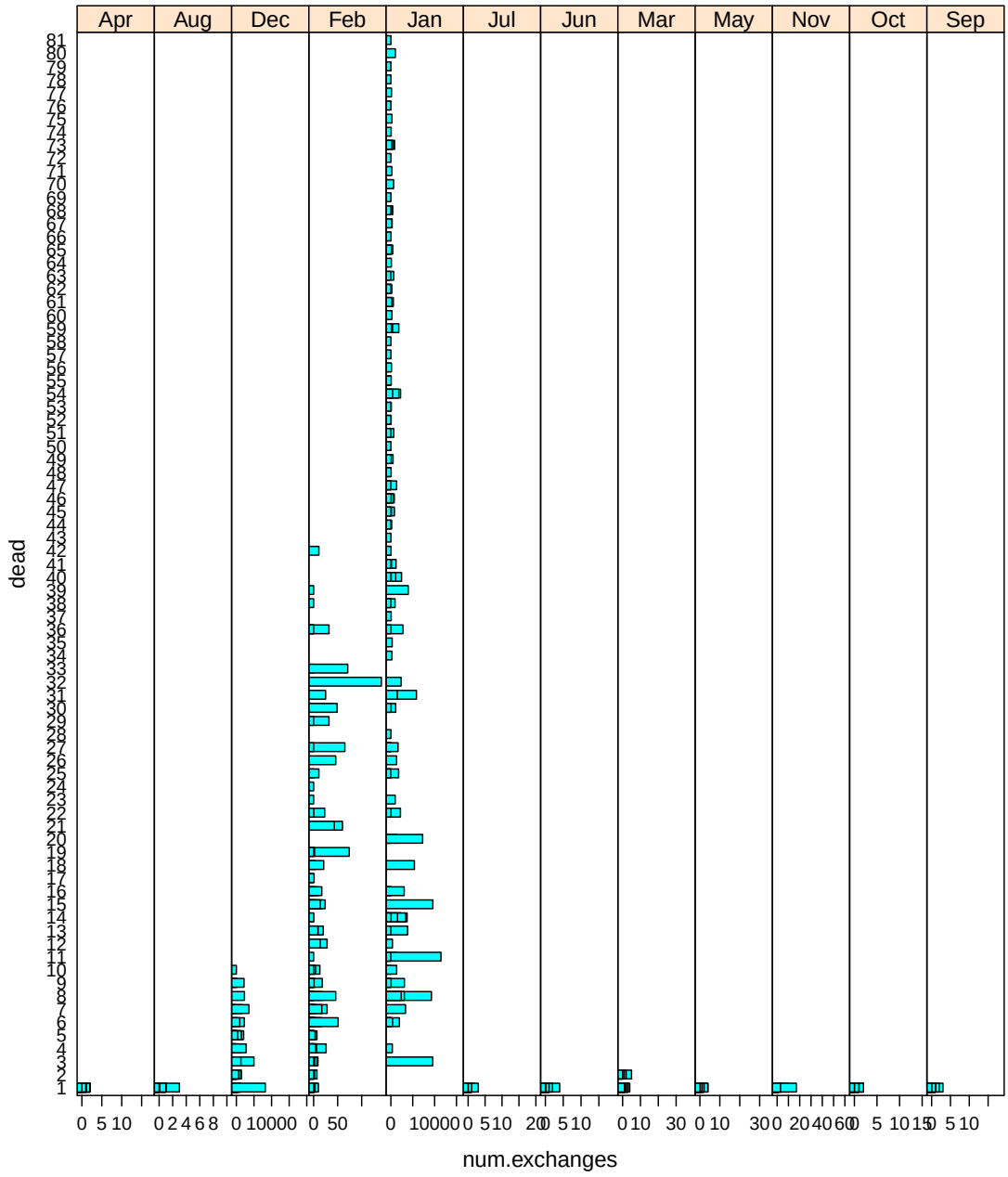


Figure 4.9: Grouped histograms for two outcome variables “dead” V.S. “in.hospital”. I aggregate monthly observations to plot twelve figures of histograms.

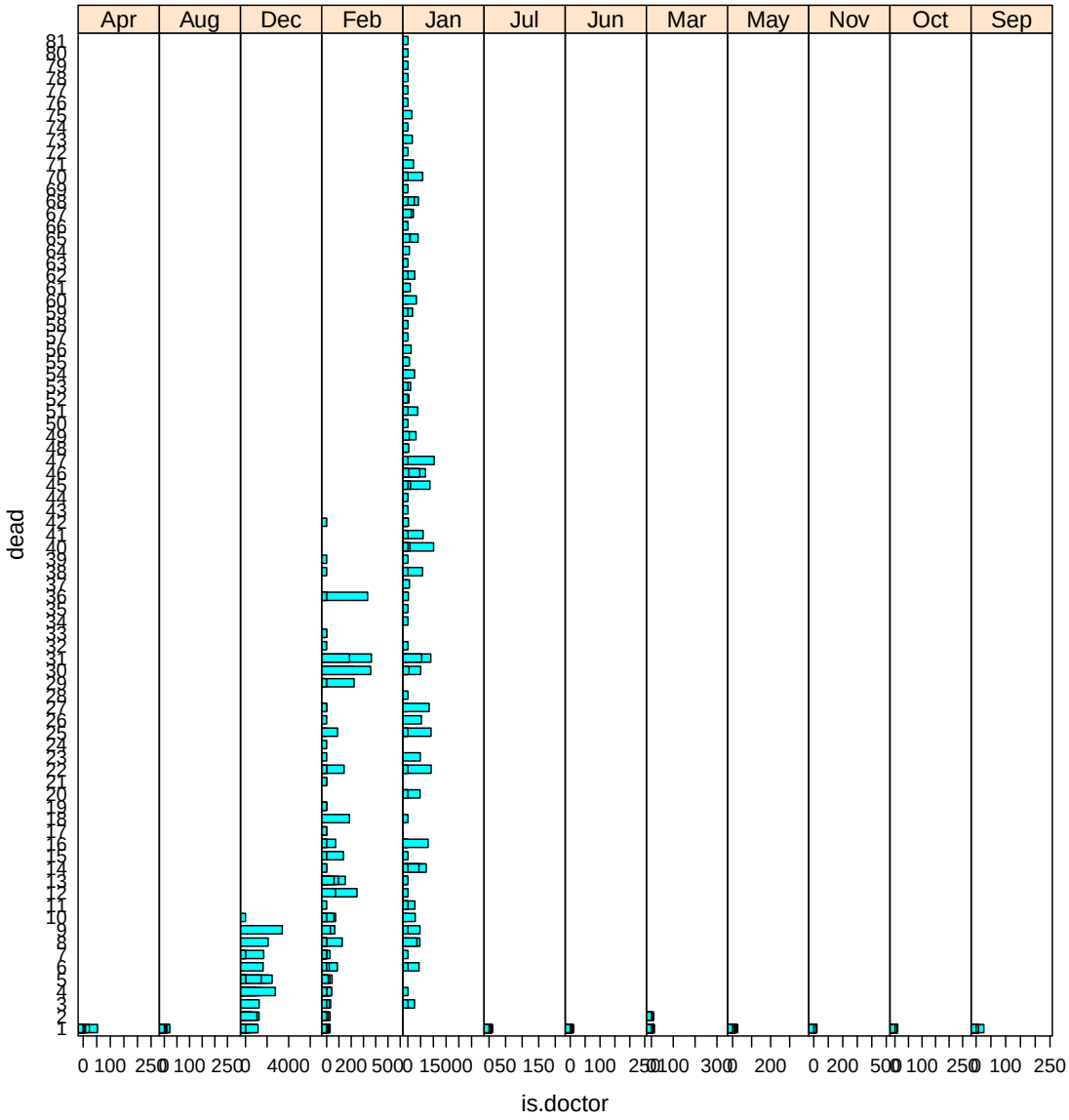


Figure 4.10: Grouped histograms for two outcome variables “dead” V.S. “in.hospital”. I aggregate monthly observations to plot twelve figures of histograms.

4.3 Related Works

Traditional predicting models focus on either the spatial dependence or the temporal correlation. Lafferty et al. [107] developed a statistical framework, Conditional Random Fields (CRFs), which accounts for

spatial dependence, in addition to the explanatory variables (observations). Later, Taskar [172] extended the Support Vector Machine (SVM) to the Maximum Margin Markov Network (M3N), which has the same modeling capacity of the CRFs but can be computed more efficiently. Similar models considering spatial dependence include the structured SVM [177] and the Maximum Margin Training [156]. All these models aim to combine spatial dependence and the information from observations for a single end task, multivariate classification. They have been successfully applied to problems like optical character recognition [151], object detection [52] and scene understanding [78]. However, these models overlook the state correlations over time, and hence, are insufficient to handle data with strong temporal pattern.

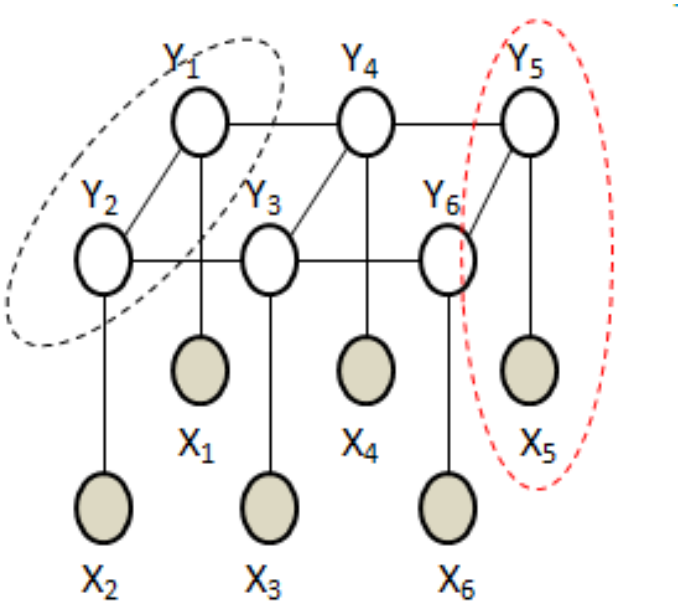


Figure 4.11: Graphical model of CRFs and M3N. X_i and Y_i correspond to the local observations and their labels. The two dashed ovals encompass $[X_5, Y_5]$ and $[Y_1, Y_2]$, which correspond to a unary feature and a pairwise Markovian spatial feature, respectively.

On the other hand, temporal correlated models such as Kalman filter [96], HMM [189] have been developed over decades and have been carefully studied by the optimization and control community. Successful applications include time series forecasting [65], speech recognition [152] and behavior classification [178]. These models are well known for their capability of capturing hidden temporal correlations; modeling the unknown state process from observations made in noisy environments. However, they ignore the structural correlations in the environment, which oftentimes hurt their performance.

Clearly, both temporal correlated models and spatial dependent models have limitations. An innovative

work [139] advocated a variational inference method for switching Linear Dynamical system (SLDS) that learns different dynamical processes at various time ticks; [113] extended this work to combine HMM and LDS with tractable computation. However, these methods treat temporal and spatial (structural) information one at a time; they fail to provide a comprehensive interface to model the temporal and spatial correlated real-world scenarios.

To close the gap, I propose a novel model that considers spatial correlations aggregated over time for tractable inference. The model has advantages over models concentrating on either aspect, as the temporal and structural information are oftentimes complementary. I intend to provide a principled approach that accounts for spatial dependence and temporal correlations, simultaneously.

4.4 Data Representation

Representing and managing uncertainty is central to understanding and supporting biomedical decision making. However, implementing effective decision models for practical applications presents a dilemma: on the one hand, informal and qualitative representations of uncertainty may be natural for people to understand but they often lack formal rigour; on the other hand formal approaches based on probability theory are precise but can be awkward and non-intuitive to use [68]. To balance these approaches, I suggest a principled data representation method to convert multiple sources of continuous observations to discrete states that are statistical comparable yet easy for people to interpret.

In decision support tasks, people are more interested in states of raw observations, e.g., “normal,” “alarm level,” and “outbreak” instead of the actual values of these observations. While there are many techniques for converting a continuous-valued time series to a discrete-valued state sequence, decision makers need one approach that can map multiple continuous observations to comparative scales; thus their relational dependence can be taken into consideration.

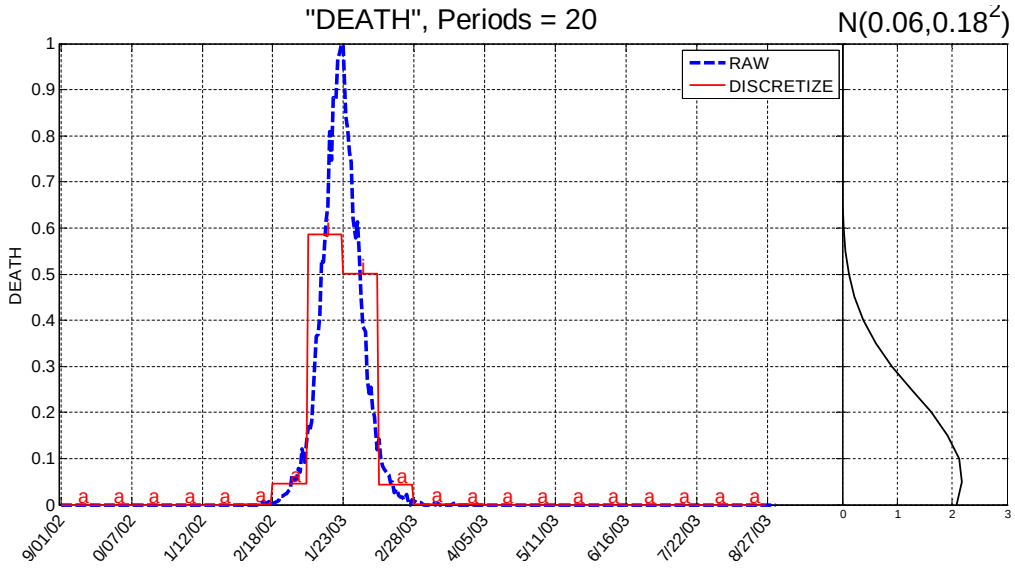


Figure 4.12: The SAX+ maps a continuous time series to 20 discretized symbols. The "death" time series in the figure is aggregated on a daily basis; then normalized into the (0,1) interval. A normal distribution was calculated to its right figure. The blue curve corresponds to a continuous time series; the red lines indicate its mean levels of every 18 days (thus, 20 periods for a year). These mean levels are discretized by calculating how many standard deviations away from the mean of the calculated normal distribution; and then corresponding letters are assigned.

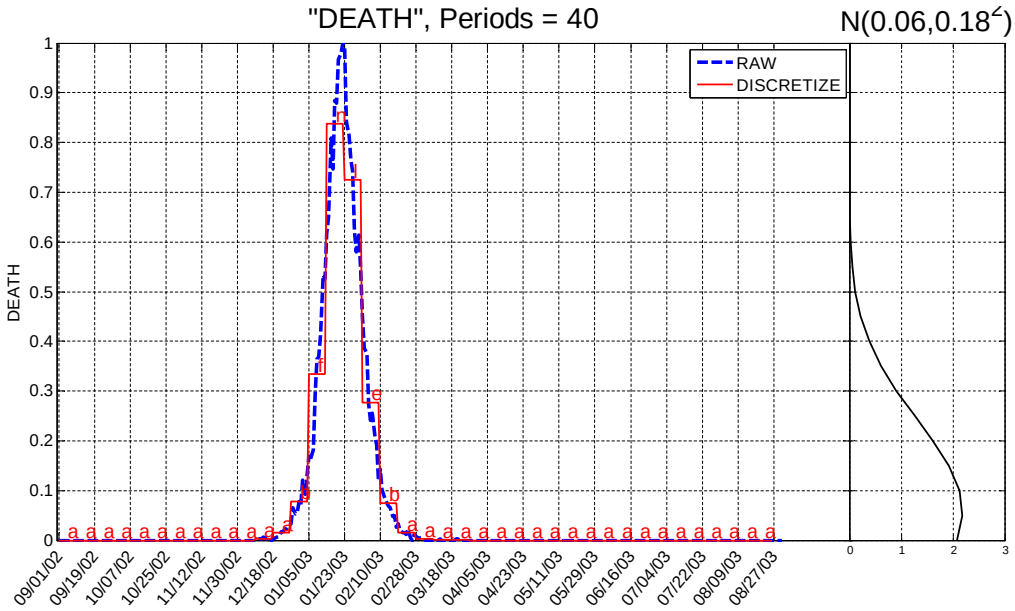


Figure 4.13: The SAX+ maps a continuous time series to 40 discretized symbols. The window size can be adjusted to achieve a finer granularity. In this new representation, I set the total periods to be 40, which reduces the window size to nine days. I thus obtained a longer sequence of symbols at a finer scale.

Inspired by [116], I develop an improved coding technique (SAX+) to map continuous time series data into discrete-valued states. The following figures (4.12, 4.13) illustrate two examples of applying SAX+ technique to the raw observation of "death" in the BioWar I.

My SAX+ approach takes a real-value time series $\mathbf{Y} = \{Y_1, \dots, Y_t, \dots\}$ and divides it into equal-sized intervals. The mean level of each interval is obtained. Each mean level is assigned a symbolic state (letter or alphabetic order) by calculating its distance from the mean. Thus, SAX+ generates a reduced dimensional piecewise-constant approximation of the raw continuous-valued observation. Figure 4.13 shows the "death" time series converted into a sequence of states: "a...abbedba....a".

My major improvement to [116] is the followings: 1) I fit a normalized time series to a Gaussian distribution; 2) I assign discrete valued states use an equal-deviation criteria as opposed to the equal-probable criteria. Due to these improvements, the converted states using SAX+ have transferable "implications". For example, letter "c" means observations are three standard deviations away from the mean and are more likely to be an unusual situation; letter "b", at two standard deviations away from the mean, may represent an alarm level; and letter "a" corresponds to a normal situation. Algorithm 1 describes the details of SAX+ state representation technique.

Algorithm 1 SAX+ state representation algorithm.

Input: Raw observation of continuous variables $X_1, \dots, X_t \dots X_T$, where $X_t = \{x_t^1, \dots x_t^m\}$ are the co-variate vector at time t.
Output: State representation $Y_1, \dots, Y_{t'} \dots Y_{T'}$, where $Y_{t'} = \{y_{t'}^1, \dots y_{t'}^m\}$ are the state representation of the corresponding co-variate patterns of $X_1, \dots, X_t \dots X_T$.
Parameters: Length of the state period: z.

1: Fit each co-variable x_*^i using a Gaussian distribution.
2: Calculate the fitted statistic and for each.
3: Calculate the average of , where $|z|$ is the length of the period.
4: Assign each $y_{t'}^{i'} = (\frac{|\bar{x}_{i \in z}^i - \mu_{x_*^i}|}{\sigma_{x_*^i}})$.

4.5 Methodology

In this section, I introduce a new machine learning framework to co-estimate multiple unknown states by exploiting their mutual interactions. The idea is to consider temporal and relational correlations globally. I start with notation and some background, followed by model details and algorithmic analysis.

4.5.1 Notation

The notations for this section are defined in the following table.

Table 4.2: Notation for Temporal Maximum Margin Markov Network.

Variables	Summary
$X_{k,i,t}$	The k dimensional feature at site i in time tick t.
$Y_{i,t}$	The discrete valued state at site i in time tick t.
$\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})'$	The estimated states at time tick t.
$\mathbf{Y}_t^* = (Y_{1,t}^*, \ldots, Y_{n,t}^*)'$	The ground-truth states at time tick t.
$\theta_k^1, \theta_{ij,t}^2$ and $\theta_{it,t-1}^3$	The unary, spatial and temporal regression coefficients.
$\varphi(\cdot)$	The feature function.
$\ell_t(\mathbf{Y}_t)$	The loss at time t.
$h_\theta(\mathbf{X}_t)$	The state estimation function.

4.5.2 Backgrounds

I summarize the framework that encompasses relational dependent models on a regular or irregular lattice. I define $s_1, \ldots, s_n$ as the sites on a relational lattice. For notational convenience, let $j \in \mathcal{N}_i$, where $\mathcal{N}_i$=$\{j$: s_j is a neighbor of $s_i\}$ defines the neighbors of the site s_i. Let $Y_1, \ldots, Y_n$ denote hidden states on the lattice, where $Y_i = Y(s_i) \in (1, \ldots, C)$, and C is the number of classes.

The joint distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ can be formulated as:

$$p_\theta(\mathbf{X}, \mathbf{Y}) \propto \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k\varphi(X_{k,i}, Y_i) + \sum_{j\in\mathcal{N}_i}\theta_{ij}\varphi(Y_i, Y_j)\right\}, \tag{4.1}$$

$$p_\theta(\mathbf{Y}|\mathbf{X}) = \frac{e^{\left\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k\varphi(X_{k,i},Y_i)+\sum_{j\in\mathcal{N}_i}\theta_{ij}\varphi(Y_i,Y_j)\right\}}}{Z_\theta(\mathbf{X})}, \tag{4.2}$$

where

$$Z_\theta(\mathbf{X}) = \sum_{Y}\exp\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k\varphi(X_{k,i}, Y_i) + \sum_{j\in\mathcal{N}_i}\theta_{ij}\varphi(Y_i, Y_j)\} \tag{4.3}$$

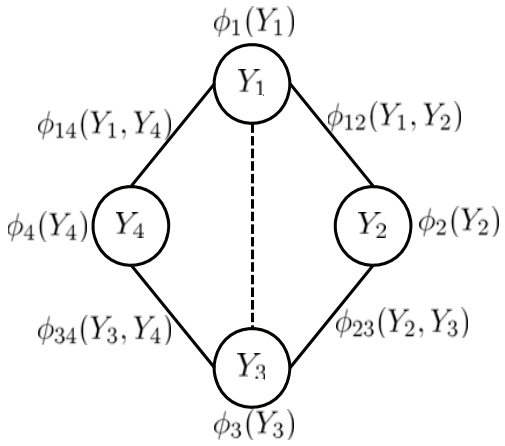
is called the partition function; $X_{k,i} = X_k(s_i)$ denotes the k-th explanatory variable at site s_i; θ_k denotes the k-th regression coefficients correspond to the feature function $\varphi(X_{k,i}, Y_i)$, with $k = 1, \ldots, p$; θ_{ij} denotes the relational dependent regression coefficients for the i-th and j-th sites so that $\theta_{ij} = \theta_{ji}$ and $\theta_{ij} \geq 0$ if $j \in \mathcal{N}_i$.

> **Interpretation:** The joint optimization framework integrates unary potential of features and pairwise potential between states for a global optimization.

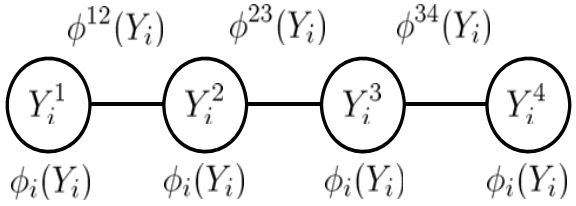
This model relates a discrete valued response variable to a hidden state by two regression components; and it is capable of estimating the probability of hidden states at a given site and predicting a certain outcome at an unsampled site. However, this formulation ignores the fact that observations are oftentimes made repeatedly over time, and past states on the same relational lattice may contribute to the states in future time ticks. That is, for a given site s_i at a given time tick t, the state is $Y(s_i, t) = Y_{i,t} \perp (Y_{i,t-1} \cup \{Y_{j,t}\}_{j \in \mathcal{N}_i})$, where $i = 1, \ldots, n$ and $t = 1, 2, \ldots$. To close the gap, I extend the model to include temporal correlations.

4.5.3 Spatial-Temporal Structured Model

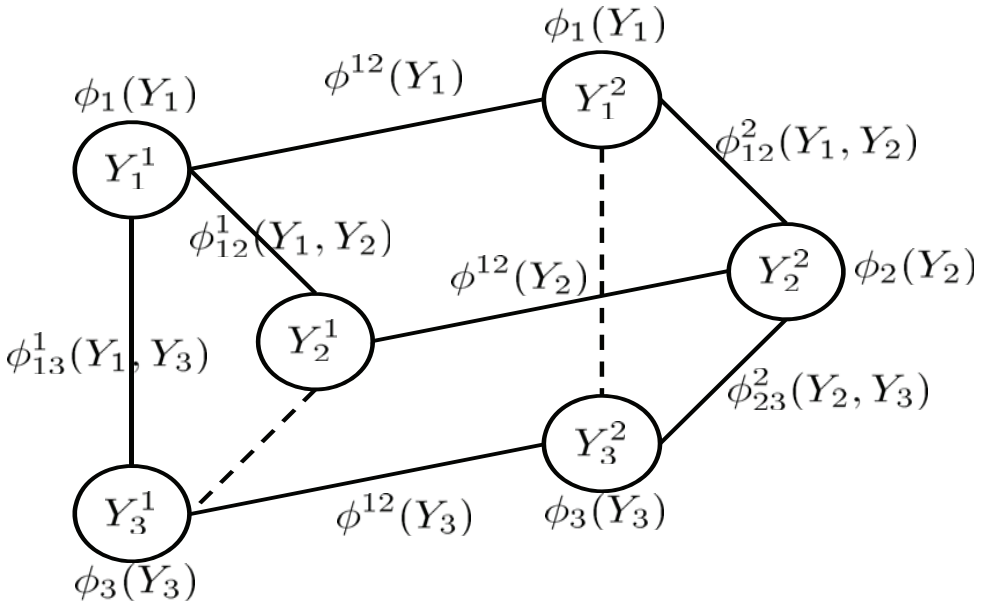
I generalize the previous framework to include an additional temporal component. With the additional regression term, the new framework is capable of modeling information carried by observations, spatial dependence at fixed time tick, and temporal correlations of the hidden states.



(a) Structural Model



(b) Temporal Model



(c) Temporal-Structural Model

Figure 4.14: Graphical models for three structured learning frameworks. (a) Typical relational dependent model - first order Markov network: $\phi_i(Y_i) = \exp\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k\varphi(X_{k,i}, Y_i)\}$ correspond to node potentials, $\phi_{i,i+1}(Y_i, Y_{i+1}) = \exp\left\{\sum_{j\in\mathcal{N}_i}\theta_{ij}\varphi(Y_i, Y_j)\right\}$ correspond to relational edge potentials. (b) Typical temporal correlated model - first order Markov chain: $\phi_i(Y_i) = \exp\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k\varphi(X_{k,i}, Y_i)\}$ correspond to node potentials, $\phi^{t-1,t}(Y_i) = \exp\left\{\sum_{i=1}^{n}\theta_{it,t-1}^{3}\varphi^3(Y_{i,t}, Y_{i,t-1})\right\}$ correspond to temporal edge potentials. (c) I propose a new framework that generalizes both relational dependent models and temporal correlated models. For illustration purposes, I only show correlated states of two consequent time ticks, but the framework indeed depicts a gigantic network over time. Thus, traditional approaches such as CRFs and M3N fail to solve it with tractable computation.

Consider a discrete valued spatial-temporal process $\{Y_{i,t} : i = 1, \ldots, n, t = 1, 2, \ldots\}$, where $Y_{i,t} = Y(s_i, t) \in (1, \ldots, C)$ corresponds to the i-th site s_i at the time tick t; $i = 1, \ldots, n$ and $t = 1, 2, \ldots$. For a given time tick t, let $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})'$ denote the discrete valued hidden states on a graph structure $\{(s_i), (s_i \times s_j)\}\,|_{i,j=1}^{n}$. I model $\{\mathbf{Y}_t : t = 1, 2, \ldots\}$ by a n-dimensional Markov chain with the following transition probability:

$$p_\theta(\mathbf{Y}_t|\mathbf{Y}_{t-1}) = q(\mathbf{Y}_t|\mathbf{Y}_{t-1})/G_{\mathbf{X}}. \tag{4.4}$$

Here $G_{\mathbf{X}}$ is a normalization constant and,

$$q(\mathbf{Y}_t|\mathbf{Y}_{t-1}) = \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k^1\varphi^1(X_{k,i,t}, Y_{i,t}) + \right.$$
$$\left.\sum_{j\in\mathcal{N}_i}\theta_{ij,t}^2\varphi^2(Y_{i,t}, Y_{j,t}) + \sum_{i=1}^{n}\theta_{it,t-1}^3\varphi^3(Y_{i,t}, Y_{i,t-1})\right\}, \tag{4.5}$$

where $X_{k,i,t} = X_k(s_i, t)$ denotes the k-th explanatory variable at site s_i and time tick t; θ_k is the linear regression coefficient corresponding to explanatory feature $\varphi^1(X_{k,i,t}, Y_{i,t}), k = 1, \ldots, p$; $\theta_{ij,t}^2$, which represents the relational regression coefficients. The difference between the Equation 4.5 and the Equation 4.2 is the additional parameters $\theta_{it,t-1}^3$ that represent the temporal coefficients. When $\theta_{it,t-1} = 0$, there is no correlation over time and the Markov network of $\{\mathbf{Y}_t\}$ reduces to a sequence of independent random vectors, and each represents a set of relational dependent observations at a given time tick. Clearly, the new framework incorporates the one previous described in Section 4.5 as $p_\theta(\mathbf{Y}_t|\mathbf{Y}_{t-1})$ reduces to $p_\theta(\mathbf{Y}_t)$.

> **Interpretation:** The new framework represents two different pairwise potentials, relational dependency and temporal correspondence, which joinly reduce the ambiguity in noisy and limited observations.

On the other hand, when $\theta_{it,t-1} \neq 0$, the framework considers state correlations over time; the magnitude of $\theta_{it,t-1}$ is related to the mean difference between two consecutive time ticks of the same site. To simplify the representations, I abbreviate the model parameters by $\theta = \left(\{\theta^1\}, \{\theta^2\}, \{\theta^3\}\right)'$; model features by $\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}_{t-1}) = \left(\{\varphi^1(X_{k,i,t}, Y_{i,t})\}, \{\varphi^2(Y_{i,t}, Y_{j,t})\}, \{\varphi^3(Y_{i,t}, Y_{i,t-1})\}\right)'$; and observations from T time points by $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$, where $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})', t = 1, \ldots, T$.

The equation 4.5 represents a general framework by considering relational-temporal correlations, which generalizes both temporal correlated models and relational dependent models, as indicated by Figure 4.14. However, there are more states to be considered together in the new framework due to the relational and temporal coupling. Thus, traditional solutions such as constructing a gigantic CRFs network would be computationally intractable.

4.5.4 Temporal Maximum Margin Markov Network

There are two typical tasks in a machine learning problem like Equation 4.5: learning and predicting. For learning, I want to estimate parameters θ so that

$$h_\theta(\mathbf{X}_t) = \text{argmax}_{\mathbf{Y}} \theta' \psi(\mathbf{X}_t, \mathbf{Y}, \mathbf{Y}^*_{t-1}) \approx \mathbf{Y}^*_t, \forall t, \tag{4.6}$$

where $\hat{\mathbf{Y}}_t$ is the ground-truth states. For predicting, I would like to infer the most likely states

$$\mathbf{Y}^*_{t+1} = \text{argmax}_{\mathbf{Y}} \theta' \psi(\mathbf{X}_{t+1}, \mathbf{Y}, \mathbf{Y}^*_t), \tag{4.7}$$

given the parameter θ and the novel observation $\mathbf{X}_{t+1}$ and past states $\mathbf{Y}^*_t$. I will now describe a convex instantiation of the spatial-temporal correlated framework to handle both tasks.

First, I need to measure the error of the approximation $h(\cdot)$ using a loss function ℓ. Here I use a Hamming distance error measurement $\ell_t(\mathbf{Y}_t)$ to indicate the number of variables predicted incorrectly, which essentially measures the loss on the label sequences,

$$\ell_t(\mathbf{Y}_t) = \sum_i \Delta(Y_{i,t}, \hat{Y}_{i,t}) \text{ and } \Delta(Y_{i,t}, \hat{Y}_{i,t}) = \begin{cases} 1 & Y_{i,t} \neq \hat{Y}_{i,t} \\ 0 & Y_{i,t} = \hat{Y}_{i,t} \end{cases}.$$

I adapt the hinge upper bound $\bar{\ell}(h_\theta(\mathbf{X}_t))$ on the loss function for structured classification inspired by max-margin criterion:

$$\begin{aligned}\bar{\ell}_t(h_\theta(\mathbf{X}_t)) = \max_{\mathbf{Y}_t} \left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}^*_{t-1})) + \ell_t(\mathbf{Y}_t)\right] \\ - \theta'\psi(\mathbf{X}_t, \mathbf{Y}^*_t, \mathbf{Y}^*_{t-1})) \\ \geq \ell_t(h_\theta(\mathbf{X}_t)),\end{aligned} \tag{4.8}$$

where $\bar{\ell}_t(h_\theta(\mathbf{X}_t)) = \bar{\ell}(h_\theta(\mathbf{X}_t), \mathbf{Y}^*_t)$ and $\ell_t(h_\theta(\mathbf{X}_t)) = \ell(h_\theta(\mathbf{X}_t), \mathbf{Y}^*_t)$.

> **Interpretation:** I optimize the lower bound of a maximization problem. The maximization of its lower bound encourages the maximization of the NP-hard objective induced by Equation 4.8.

With this upper bound, the min-max formulation for the structured classification problem is analogous to SVM,

$$\min_{\theta, \mathbf{Y}_t} \frac{\lambda}{2}||\theta||^2 + \frac{1}{T}\sum_{t=1}^{T}\xi_t \tag{4.9}$$

$$s.t. \left\langle \theta, \Phi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}^*_{t-1}, \mathbf{Y}^*_t)\right\rangle \geq \bar{\ell}(\mathbf{Y}_t, \mathbf{Y}^*_t) - \xi_t, \forall t, \forall \mathbf{Y}_t, \tag{4.10}$$

where $\Phi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}^*_{t-1}, \mathbf{Y}^*_t) = \psi(\mathbf{X}_t, \mathbf{Y}^*_t, \mathbf{Y}^*_{t-1}) - \psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}^*_{t-1})$.

> **Interpretation:** The final formulation of TM3N is similar to the structured Support Vector Machine, but TM3N extends it to consider state correlations from different perspectives jointly.

This formulation incorporates the "maximum margin" criteria. The following formulation indicates the margin of the state configuration $\mathbf{Y}^*_t$ over another state configuration $\mathbf{Y}_t$.

$$\mathrm{M} = \frac{1}{||\theta||}\left\langle \theta, \Phi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}^*_{t-1}, \mathbf{Y}^*_t)\right\rangle \tag{4.11}$$

Assuming ξ_i are all zeros (because λ is very small), the constraints enforce,

$$\theta'\left(\psi(\mathbf{X}_t, \mathbf{Y}^*_t, \mathbf{Y}^*_{t-1}) - \psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}^*_{t-1})\right) \geq \bar{\ell}(\mathbf{Y}_t, \mathbf{Y}^*_t), \tag{4.12}$$

so minimizing $||\theta||^2$ essentially maximizes the smallest of such margins, scaled by the loss $\ell_i(\mathbf{Y}_t, \mathbf{Y}_t^*)$. The above formulation is a standard QP and can be solved using optimization packages, but it is exponential in the size and computation is generally prohibitive. Another way to express this problem is the following representation,

$$\begin{aligned} &\min_{\theta, \mathbf{Y}_t} \frac{\lambda}{2}||\theta||^2 + \frac{1}{T}\sum_{t=1}^{T}\xi_t \\ &s.t. \theta'\psi(\mathbf{X}_t, \mathbf{Y}_t^*, \mathbf{Y}_{t-1}^*) + \xi_t \geq \\ &\quad \max_{\mathbf{Y}_t} \left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}_{t-1}^*)) + \ell_t(\mathbf{Y}_t)\right], \forall t, \end{aligned} \tag{4.13}$$

which is a convex quadratic program in θ, since

$$\max_{\mathbf{Y}_t} \left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}_{t-1}^*)) + \ell_t(\mathbf{Y}_t)\right], \tag{4.14}$$

is convex in θ.

> **Interpretation:** The global optimization induced by Equation 4.13 is a convex function and can be optimized if the embedded maximization problem is solved in polynomial time. However, the intermediate value of $\max_{\mathbf{Y}_t} \left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}_{t-1}^*)) + \ell_t(\mathbf{Y}_t)\right]$ is obtained through a non-convex optimization, which is introduced later.

It might be easier to interpret Equation 4.13 in its alternative representation Equation 4.15 by eliminating the constraints,

$$\min_{\theta, \mathbf{Y}_t} \frac{\lambda}{2}||\theta||^2 + \frac{1}{T}\sum_{i=1}^{T}\Big\{\max_{\mathbf{Y}_t} \left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}_{t-1}^*)) + \ell_t(\mathbf{Y}_t)\right] \tag{4.15}$$

$$-\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t^*, \mathbf{Y}_{t-1}^*)\Big\}. \tag{4.16}$$

Careful readers might notice that $\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t^*, \mathbf{Y}_{t-1}^*)$ is invariant to $\mathbf{Y}_t$ and I can run the algorithm in two separate steps: first, fix θ and optimize $\max_{\mathbf{Y}_t} \left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}_{t-1}^*)) + \ell_t(\mathbf{Y}_t)\right]$; second, fix $\mathbf{Y}_t$ obtained in the first step to calculate θ that minimize Equation 4.15. The procedure is similar to the Expectation-

Maximization algorithm and I are guaranteed not to increase the objective function at each step.

4.5.4.1 Learning

Recalling the objective in Equation 4.15 is a convex function, an intuitive way to estimate its parameters θ is to use a gradient descent approach. In this case, the gradients only depends on the most violated state configuration,

$$\mathbf{Y}_t^* = \text{argmax}_{\mathbf{Y_t}} \left(\theta' \psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}_{t-1}^*)) + \ell_t(\mathbf{Y}_t) \right), \tag{4.17}$$

which can be computed as:

$$g(\theta) = \lambda\theta + \frac{1}{T} \sum_{i=1}^{T} \left(\psi(\mathbf{X}_t, \mathbf{Y}_t^*, \mathbf{Y}_{t-1}^*) - \psi(\mathbf{X}_t, \mathbf{Y}_t^*, \mathbf{Y}_{t-1}^*) \right). \tag{4.18}$$

> **Interpretation:** The equation above is a subgradient of the objective in Equation 4.15. Since this objective function is convex, its optimality can be obtained by tracing its subgradient direction stepwise.

The following algorithm thus summarizes the procedure of gradient optimization,

Algorithm 2 Sub-gradient Optimization

Input: training data $\mathcal{D} = \{(\mathbf{X}_t, \mathbf{Y}_t)\}|_{t=1}^{T}$, regularization parameter λ, step size σ, tolerance ϵ, number of iterations T
Output: parameter vector θ

```
Initialize θ ← 0, t ← 1
repeat
    for t = 1 to T do
        Set violation function H(Y_t) = θ'ψ(X_t, Y_t, Y*_{t-1})) + ℓ_t(Y_t) − θ'ψ(X_t, Y*_t, Y*_{t-1})
        Find most violated label for (X_t, Y_t) : Y*_t = arg max_{Y_t} H(Y_t)
    end for
    Compute g(θ), update θ^(t) ← θ^(t−1) − σg(θ).
    Update t ← t + 1
until t ≥ T or MSE(||θ^(t)|| − ||θ^(t−1)||) ≤ ε
```

A critical step of Algorithm 2 is to compute the most violated constraint at each time step efficiently. The exact inference of this step is usually intractable as irregular lattices often involve loops that cannot be handled by deterministic algorithms in polynomial time. To this end, I leverage on a well established approximation algorithm, loopy belief propagation (LBP) [132] to solve this problem. To use LBP, I define the following potentials:

- **Unary potentials** represent the impact of local observation in $\mathbf{X}_t$ to the states $\mathbf{Y}_t$, this potential function at each site s_i takes the form,

$$\exp\left(\sum_{k=1}^{p}\theta_k^1\varphi^1(X_{k,i,t},Y_{i,t})+\ell_t(Y_{i,t})\right),\forall i, \tag{4.19}$$

- **Environmental potentials** represent the influence between states and over time, these potential functions take the form,

$$\exp\left(\theta_{ij,t}^2\varphi^2(Y_{i,t},Y_{j,t})\right),\forall i,j\sim i,\ \text{Structural Potential}, \tag{4.20}$$

$$\exp\left(\theta_{it,t-1}^3\varphi^3(Y_{i,t},Y_{i,t-1})\right),\forall i,\quad \text{Temporal Potential}. \tag{4.21}$$

4.5.4.2 Predicting

Now I will introduce my linear integer programming interface for predicting. The goal is to predict a hidden state as the most likely configuration:

$$\hat{\mathbf{Y}}_{T+1}=\text{argmax}_{\mathbf{Y_{T+1}}}\left(\theta'\psi(\mathbf{X}_{T+1},\mathbf{Y}_{T+1},\mathbf{Y}_T^*)\right). \tag{4.22}$$

Denote $Z^t=(\{z_i^t\}|_{i=1}^n,\{z_{ij}^t\}|_{i=1}^{n,j\sim i},\{z_i^{t,t-1}\}|_{i=1}^n)$ as indicator variables at time t so that: $z_i(m)=1$ indicates i-th site takes state m, $z_{ij}(m,n)$ indicates i and j-th sites take states m and n, and $z_i^{t,t-1}(m,n)=1$ indicates i-th site take states m and n at time t and $t-1$, respectively. If I factorize Equation 4.22, the following linear integer programming interface defines an exact mapping,

$$\max_{Z^t} \quad \sum_{i,m} z_i^t(m) \left[\theta^1_{(\cdot)}\varphi^1(X_{(\cdot),i,t}, m)\right] + \sum_{i,j,m,n} z_{ij}^t(m,n)\left[\theta^2_{ij,t}\varphi^2(m,n)\right] + \sum_i \left[\theta^3_{it,t-1}\varphi^3(m, Y^*_{i,t-1})\right], \tag{4.23}$$

$$s.t. \quad z_i^t(m) \in \{0,1\},\ z_{ij}^t(m,n) \in \{0,1\}, \tag{4.24}$$

$$\sum_m z_i^t(m) = 1, \tag{4.25}$$

$$\sum_n z_{ij}^t(m,n) = z_i^t(m), \tag{4.26}$$

$$\sum_m z_{ij}^t(m,n) = z_j^t(n). \tag{4.27}$$

Interpretation: I used binary indicator functions to encode the variables to state correlations over time. Ideally, this would give the optimal solution of objective function (Equation 4.22). However, this linear integer programming is NP-hard to solve, so I have to use some approximation.

The constraint Equation 4.25 enforces only one state allocated for each site s_i; the constraint equation 4.26 enforces the structural consistency. Note that I assign $z_i^{t,t-1}(m, Y^*_{i,t-1}) = 1, \forall i$ so that $Y_{i,t}$ is influenced by its previous state $Y^*_{i,t-1}$of the same site s_i. The above linear integer programming is an intractable combinatorial problem, but I can obtain an approximated solution by relaxing the binary constraint in Equation 4.24 to be $z_i^t(m) \geq 0,\ z_{ij}^t(m,n) \geq 0$. A threshold χ, usually equals to 0.5, is used to discretize the final outputs Z^t for predicting the states.

4.6 Model Complexity

The computational cost of TM3N is the complexity of $\max_{\mathbf{Y}_t}(\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}^*_{t-1})) + \ell_t(\mathbf{Y}_t))$ because the search space of $\mathbf{Y}$ is exponentially large. The computation of $\mathbf{Y}$ is intractable if computed brutal force, which costs $O(K^{(N-1)})$ time and $O\left((N-1)K^C\right)$ space. Here C is the max clique size, K is the number of states and N is the number of nodes in the graph G. Dynamic algorithms, on the other hand, trade space for time. Viterbi and forward-backward algorithms can do exact inference if the underlying structure is a

chordal graph.

1. For a chain structure, forward-backward algorithms can compute $\max_{\mathbf{Y}_t}(\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}^*_{t-1}))+\ell_t(\mathbf{Y}_t))$ in $O\left((N-1)K^2\right)$.
2. On a tree structure, exact inference is possible through belief propagation. It computes all N marginals in 2 passes over graph and the computational cost is still $O\left((N-1)K^2\right)$. In general, the exact inference algorithm for singly-connected networks and the beliefs converge to the correct marginals in a number of iterations equal to the diameter of the graph [132].
3. Exact inference in any graphical model is made by converting to a tree and running BP. The resulting Junction Tree has nodes with $O(K^{w+1})$ states, so inference time is bounded by $O\left((N-1)K^{2w+2}\right)$ [w=clique size of triangulated graph], which states that:
   (a) Complexity of BP is exponential in the size of the nodes in the maximum clique.
   (b) In most real world applications, node size in equivalent triangulated tree is huge.
4. Another option is to apply Linear Programming to the original graph in $O\left((N-1)K^2\right)$ time. [2].

I decided to use Linear Programming strategy for optimizing the $\max_{\mathbf{Y}_t}(\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}^*_{t-1})) + \ell_t(\mathbf{Y}_t))$ for generalization purpose.

4.7 Experiments

4.7.1 Synthetic Results

I use the following temporal-spatial correlated Linear Dynamic System (LDS) to generate the simulation. This system specifies the hidden state Y_t^i, which depends temporally on the previous state Y_{t-1}^i and correlates spatially with the states of the neighboring sites Y_t^j, $j \in \mathcal{N}_i$.

[2]This is empirically successful but results are approximations to the exact solution.

$$Y_t^i = \alpha Y_{t-1}^i + (1-\alpha) \sum_{j \in \mathcal{N}_i} \beta^j Y_t^j + e_1, \tag{4.28}$$

$$X_t^i = A Y_t^i + e_2, \tag{4.29}$$

$$e_1 \sim N(0, \sigma_{e_1}^2), \tag{4.30}$$

$$e_2 \sim N(0, \sigma_{e_2}^2), \tag{4.31}$$

where $\mathcal{N}_i$ corresponds to the neighboring sites of i but excludes i; A is a projection vector that maps hidden states to the observations; X_t^i corresponds to the observations at site i, time tick t; e_1 and e_2 are the environmental Gaussian noises;α represents the temporal/relational trade-off parameter. If α is set to zero, the system considers no time dependence. Otherwise, if α is set to one, the system ignores relational correlations.

> **Interpretation:** The linear dynamic systems induced by Equation 4.28 and Equation 4.29 are temporal and relational dependant. The outputs are generated by states with slight perturbation.

I initialized $Y_t^i \sim \mathrm{Uniform}(0,1)$, $\beta^j \sim \mathrm{Uniform}(0,1)$; set total sites number N equals to four; specified the error term $e_1 \sim N(0, 0.05)$ and $e_l \sim N(0,1)$; and let the projection matrix be $A = [10; 20]$. To simulate the hidden states, I used an approach similar to Gibbs sampling that iteratively samples Y_t^i until the system converges. These simulated states were rounded to real valued states and the simulated observations were calculated using Equation 4.29.

In the experiment, I varied the temporal/spatial trade-off parameter α from 0.1 to 0.8 at an interval of 0.1 to evaluate the performances of four different models: HMM, M3N, CRF and TM3N. For every α value, I ran the experiment 50 times to calculate the averaged accuracy. The results are demonstrated in Figure 4.15, where the blue curve corresponds to the accuracy of TM3N model at various α values. Obviously, TM3N shows superior performance compared to HMM, CRF and M3N.

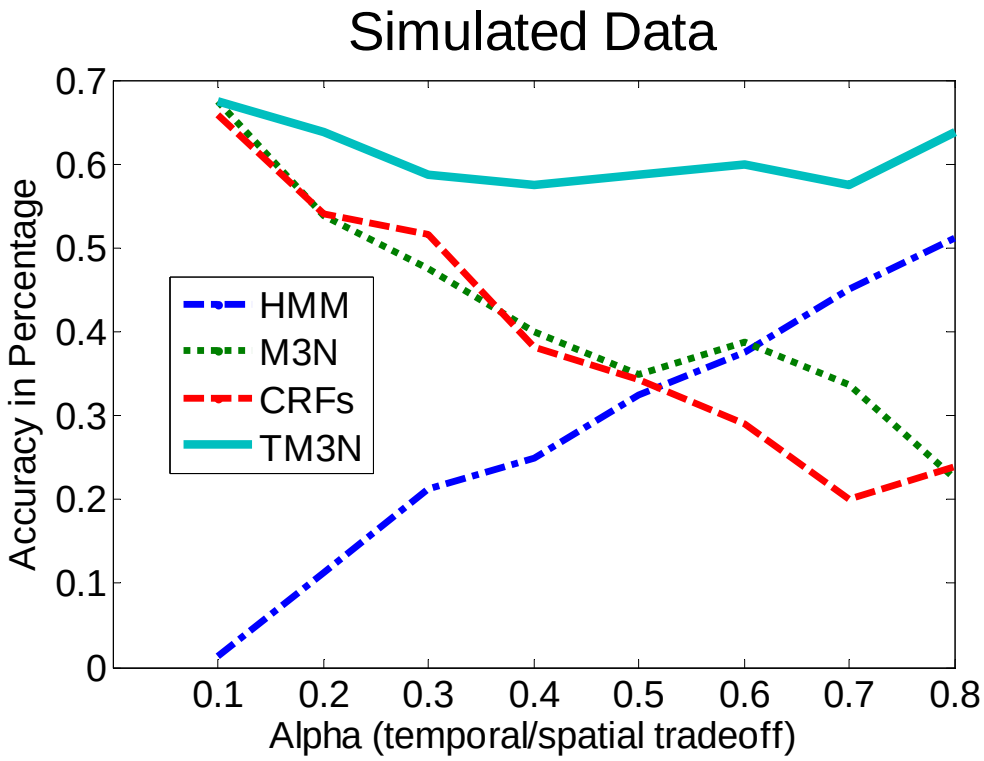


Figure 4.15: Structured learning model comparison using synthetic temporal-relational correlated data. The X axis corresponds to the Alpha (temporal/relational trade-off parameter) value and Y axis represents the accuracy in percentage. HMM's performance increases as the temporal influence becomes larger while CRFs/M3N's accuracy decreases at the same time. TM3N outperforms all other models and demonstrates its efficacy.

Table 4.3: Averaged accuracy of four different methods using synthetic LDS data with various α value. The number in each cell indicates the averaged accuracy.

	Value of α							
Models	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
HMM	0.01	0.11	0.21	0.25	0.33	0.38	0.45	0.51
CRFs	0.66	0.54	0.52	0.38	0.34	0.29	0.2	0.23
M3N	0.68	0.54	0.47	0.4	0.35	0.39	0.34	0.23
TM3N	0.68	0.64	0.59	0.58	0.59	0.6	0.58	0.63

Interpretation: The results of applying TM3N to synthetic linear dynamic system outputs indicates the model's superior performance over the other methods in comparison, including CRF, M3N and HMM. This indicates the benefit of joint optimization and verifies my earlier assertion that information from complementary perspectives helps to resolve ambiguities.

4.7.2 BioWar Hidden States Co-Estimation

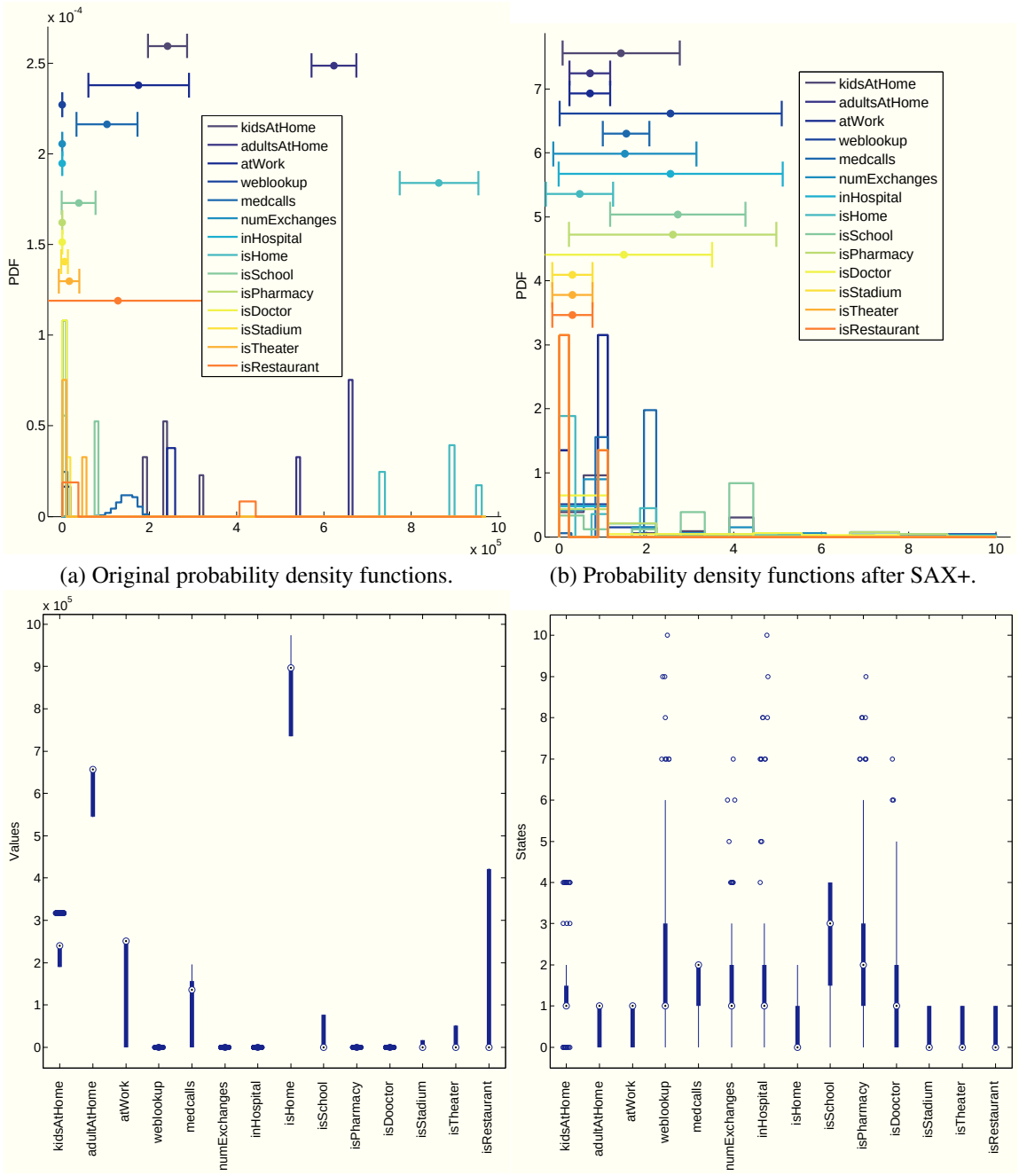
4.7.2.1 Data Processing

We use a period $P = 52$ to divide an entire year of BioWar-I, II into 52 week-long periods and apply the SAX+ State coding technique to obtain discrete valued states, which is illustrated in Table 4.4

Table 4.4: Discrete-valued states obtained using the SAX+ approach.

death	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	1	1	1	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kids_at_home	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	3	3	0	0	0	0	4	4	4	4	3	4	4	4	4	4	4	4	4
adults_at_home	2	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1
at_work	1	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	0	0	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	1	2	2	2	2	2	2	2	2
weblookup	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6	1	1	7	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
medcalls	0	1	2	2	2	1	3	3	3	3	2	3	2	4	4	3	1	1	2	3	4	5	5	5	5	5	5	4	2	3	3	3	1	2	2	3	3	2	3	2	3	3	2	0	2	0	2	1	0	0	2	2
num_exchanges	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6	1	1	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
in_hospital	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	1	1	9	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
is_home	1	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	0	3	3	3	3	1	3	3	3	3	3	3	3	3
is_school	3	4	4	4	4	4	3	4	4	4	4	4	3	4	4	4	0	1	4	4	3	4	4	4	3	4	4	4	4	4	4	4	4	1	1	4	4	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0
is_pharmacy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5	1	1	8	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
is_doctor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	7	1	1	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
is_stadium	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
is_theater	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
is_restaurant	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0

Generally, a larger letter indicates more severity of the corresponding observations, e.g., Weblookup peaks in the week of Jan. 5, 2003 with a state "e", which indicates the observation is standard deviations from the normal situation (mean). I can therefore speculate that the situation could be very severe. However, there are a few exceptions when Gaussian model does not handle the pattern well, for example, kids_at_home and is_school shows significant pattern changes from Jun. 1, 2008, which is not because of disease outbreak but due to the fact of summer vacation. Although concept drifts of these converted states can still be observed during disease outbreak, a better scaling would be achieved with Gaussian Mixture Models (GMM).



(a) Original probability density functions.

(b) Probability density functions after SAX+.

(c) Boxplot for original co-variables.

(d) Boxplot for SAX+ encoded co-variables.

Figure 4.16: Probability density functions and box plots before and after SAX+ encoding. The states generated after SAX+ encoding are more comparable than their raw values, which makes co-prediction of multiple factors meaningful.

Figure 4.16 illustrates the probability density function of the observation variables before and after apply-

ing our proposed SAX+ algorithm. The figures on the right column indicate that the scale of these variables are closer.

4.7.2.2 Results

Given the symbolic representation introduced in Section 4.7.2.1, I convert raw observations as the synchronized states over time. In this representation, I denote Y^1 as death rates, Y^2 as infection rates, Y^3 as absenteeism from work and school. Let $\mathbf{Y} = \{Y^1, Y^2, Y^3, Y^4, Y^5, \ldots, Y^{15}\}$ be the states in the period z.

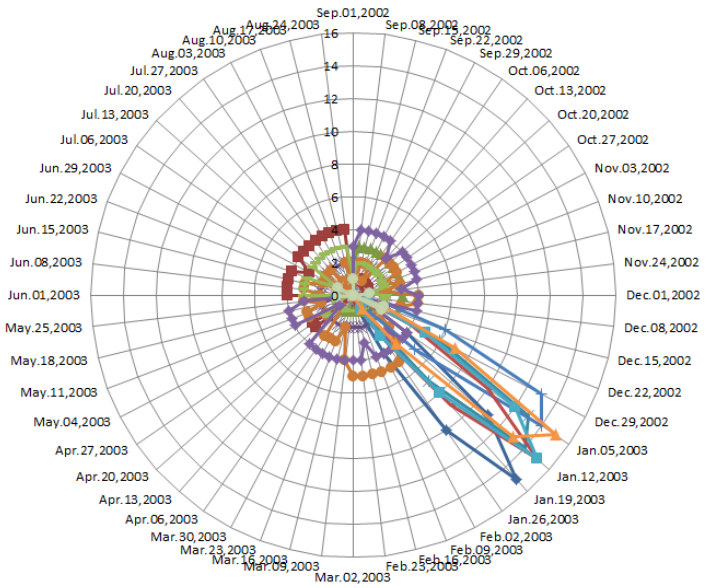
We also denote the features as observations in the previous period $\mathbf{X} = \{X^1, X^2, X^3, X^4, X^5, \ldots, X^{15}\}$, where each X^i contains a continuous period of observations from the previous time tick. The prediction of M3N are compared with three other models: HMM, CRF and M3N.

Table 4.5: Accuracy of state estimation, TM3N vs. CRF, HMM and M3N.

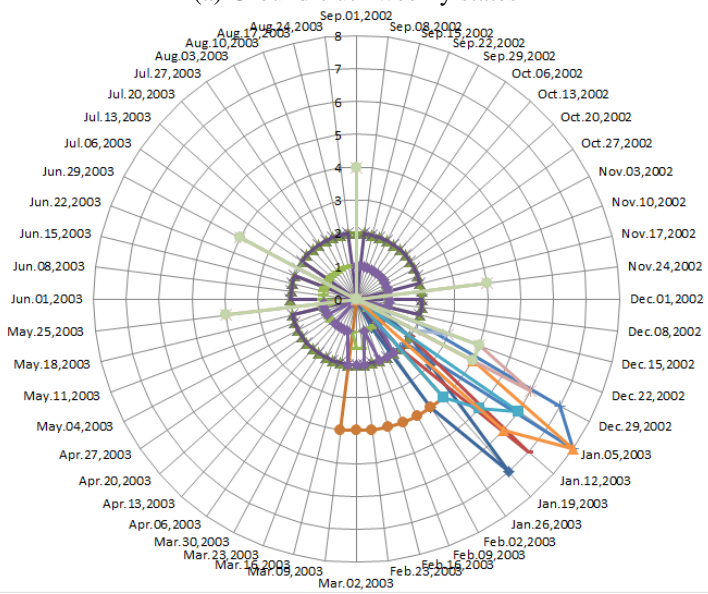
	HMM	CRF	M3N	TM3N
Accuracy	0.65	0.57	0.58	0.69

Our evaluation standard is the overall state estimation accuracy $Accuracy = \frac{F_1}{F_1+F_2}$, where F_1 is the number of correctly classified states and F_2 is the number of misclassified states. Table 4.5 summarizes the performance of various models. Obviously, TM3N outperforms the rest models in comparison by combining temporal and relational information into consideration, simultaneously.

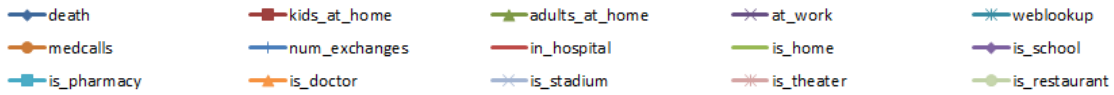
Interpretation: TM3N demonstrates its efficacy on the BioWar-I data, which empirically confirms the advantage of this global optimization framework over existing approaches.



(a) Ground-truth weekly states



(b) TM3N predicted weekly states



(c) Legend for both sub-figures a and b.

Figure 4.17: Radar diagram of original and M3N predicted states. It can be observed that the ordering of peaks is consistent between the ground-truth states and the predicted states: num_exchange>is_doctor$\geq$ is_pharmacy>in_hospital>medcalls>death.

Interpretation: I can aggregate predictions over time to obtain a reduced form of the BioWar-I model, which is generated by the Biowar simulation engine. The reduced form provides quick interpretation of observations and assists decision makers for efficient responses.

Figure 4.17 illustrates both a) ground-truth states and b) predicted states over the period of a year. It can be seen that an outbreak of avian influenza occurred during the period from Dec. 2002 to Jan. 2003, as a few random variables deviates more than 3 standard deviations away from their normal patterns. I verified the relative peaking time of predicted state variables matches various sources of raw observations, e.g., num_exchange>is_doctor$\geq$ is_pharmacy>in_hospital>medcalls>death.

4.8 Discussion

Historically, manifestations of the disease process and response strategies are modeled by the susceptible-infected-recovered (SIR) model[39]. This model categorizes the entire population into three groups of susceptible, infected, and recovered. Individuals in each group are assumed to have the same states, and SIR uses predefined transition probability to model the disease progression. However, the "population-based" disease progression processes model assuming a homogeneous mixing of individuals has limited capability of modeling disease spreading at a finer granularity.

This chapter presents a methodology for predicting large-scale disease outbreaks using observations generated by a next-generation disease spreading simulation engine: BioWar. The simulation engine simulates the agent-level impact of a bio-terrorist attack in a U. S. city using social networks. As opposed to traditional methods that model hypothetical cites, the BioWar engine uses real city data like census, school track, and other public available sources to output various manifestations of the disease as the simulated agents go about their lives.

To make the prediction tractable and meaningful, I developed a discrete-valued state representation method, SAX+, which balances quantitative methods that are formally rigorous and qualitative methods that are intuitive and easy to understand. By mapping various sources of continuous observations to discrete-valued states (symbolic representation) at comparable scales, SAX+ provides an intuitive interface to watch concept drifts that might relate to adverse events. On top of the state representation, I developed a framework to estimate relational-correlated hidden states. As opposed to traditional approaches that focus on a single

outcome, my framework models semantic correlations among heterogeneous observations in addition to an individual state. My framework demonstrated superior performance over existing approaches by considering temporal and relational correlations globally.

If all predictions over time are aggregated, I obtain a reduced form of the BioWar model. Thus, the same technique can be applied to other high fidelity models to assess how they are the same or different; i.e., do their reduced form versions look the same or different in terms of their outcome metric. The state representation approach has limitations in modeling a time series that demonstrates multiple patterns. But this can be improved by a more sophisticated Gaussian Mixture Model (GMM) at the cost of increasing computational complexity. On the other hand, my structured prediction framework offers *discrimination* against states but cannot be used to generate new samples due to its *discriminative* modeling assumption. Although not comprehensive, the framework serves many decision support purposes. Finally, results on both synthetic data and BioWar simulation data indicate the framework's applicability and efficacy in modeling large-scale disease outbreaks and encourage more efforts in the same direction.

4.9 Conclusion

In this chapter, I revealed that considering information from different sources concurrently could improve the prediction accuracy. I designed a state coding technique, SAX+, to discretize continuous observations into human interpretable values. Together with TM3N, a new framework developed in this chapter, I optimized temporal coherence together with relational dependence with tractable computation. As opposed to traditional approaches that predict states of outcome variables independently, my framework described semantic correlations of heterogeneous variables and observations of individual variables in a global manner. The joint optimization reduced the ambiguity in estimating multiple outcome variables independently. In addition, this approach seamlessly integrated state-of-the-art maximum margin based kernel methods developed for classifying independent instances with the rich description ability of graphical model that can exploit the contextual structure of complex data.

In summary, TM3N framework offers flexibility in modeling complex systems; and it can be easily generalized to other dynamic systems involving temporal, spatial, and relational dependencies. The framework can be applied to cases involving multiple dependent manifestations, which are time correlated. For example, ICU patients management, multiple speaker speech tagging and object tracking in video sequences are

good candidates for applying TM3N framework. More specifically, inputs to TM3N can be any features (continuous or categorical) from a set of instances while the outputs are some or all of their corresponding states. Note these states must be discrete values as TM3N is a discriminative model that cares about the accuracy of predictions.

However, it has its own limitations. Due to the adoption of maximum margin criteria, the framework inherited discriminative power of SVMS but loosing the generative ability of graphical models. That is, it cannot simulate novel data with learned parameters. This is a tradeoff between classification performance and modeling capacity, I preferred former as it is most critical to decision support in emergency response. To verify the efficacy, I conducted experiments using both synthetic and BioWar simulation data. The synthetic experiments using linear dynamic systems simulated outputs demonstrated TM3N constantly outperformed HMM, CRF and M3N at various temporal and relational impact ratios. The BioWar experiment indicated that TM3N again led the competition with 69% of accuracy, outperforming CRF (57%), M3N (58%), and HMM (65%).

Chapter 5

A Unified View of Discrimination and Calibration

Psychologically, people can make more informed decisions when they are aware of the progression of the events and the consequences of their actions [51, 204]. To assist the decision making process, a common strategy is to use a predictive model to determine the class membership of novel observations based on empirical evidences of the same event [10, 11, 79].

Given a reasonably large amount of observations and their labels, the prediction task is often generalized to supervised learning problems, which provide a mapping between features and outcomes (usually represented by the 0/1 class membership). *Discrimination* and *calibration* are two major families of measurements in the evaluation of model performance [12, 42, 54, 124]. In clinical predictions, *discrimination* measures the ability of a model to separate patients with different outcomes; in the case of a binary outcome, good *discrimination* indicates adequate distinction in the distributions of predicted values, based on the model, between the two classes, defined by the binary outcome [41]. On the other hand, *calibration* measures the similarity between the predicted values and the observed outcomes. To increase the quality for classification model in clinical research, it would be more proper to calculate discrimination and calibration concurrently [42].

Oftentimes, good *calibration* tends to correspond to good *discrimination*; however, a successful classifier in terms of *discrimination* ability or Area Under ROC Curve (AUC) is not necessarily calibrated. For

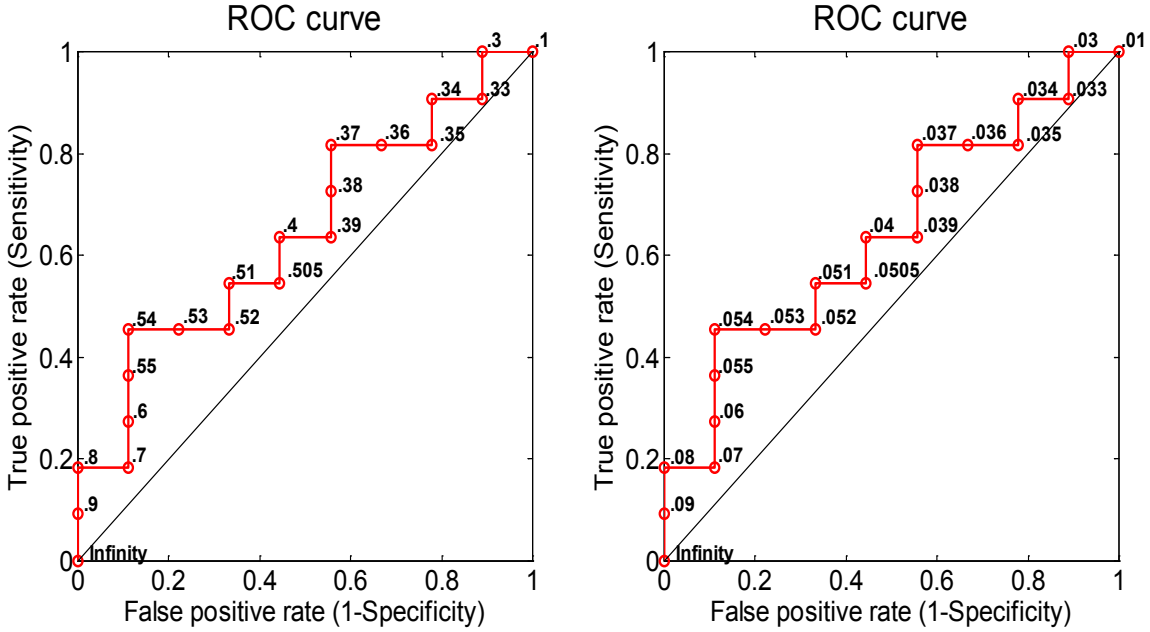
instance, a model that predicts all positive outcomes to occur with probability 0.51 and all negative outcomes to occur with probability 0.49 has perfect *discrimination* but bad *calibration*. Figure 5.1 gives another example of such conflict.

#	1	2	3	4	5	6	7	8	9	10
C	p	p	n	p	p	p	n	n	p	n
P	.9	.8	.7	.6	.55	.54	.53	.52	.51	.505
#	11	12	13	14	15	16	17	18	19	20
C	p	n	p	p	n	n	p	n	p	n
P	.4	.39	.38	.37	.36	.35	.34	.33	.3	.1

(1) Probabilistic Classifier A

1	2	3	4	5	6	7	8	9	10
p	p	n	p	p	p	n	n	p	n
.09	.08	.07	.06	.055	.054	.053	.052	.051	.0505
11	12	13	14	15	16	17	18	19	20
p	n	p	p	n	n	p	n	p	n
.04	.039	.038	.037	.036	.035	.034	.033	.03	.01

(2) Probabilistic Classifier B



(a) ROC curve for two probabilistic models (1) and (2). Notice the two figures have the same AUC except the thresholds are different

Figure 5.1: Probabilistic classifier outputs and their ROC curve. In sub-figure (1) and (2), the number indicates the sample, C corresponds to class membership and P represents the probabilistic output. In sub-figure (c), each red circle corresponds to a threshold value. Note the probabilistic classifier B has the same AUC as the probabilistic classifier A, yet it is uncalibrated.

Readers can easily observe that the probabilistic classifier A and the probabilistic classifier B have the same AUC. But probabilistic outputs of classifier B is ten times smaller that of classifier A, and thus obviously uncalibrated. Of course, a classifier that assigns sample $\#1$ to the positive class with a likelihood estimate of 0.09 is useless to any decision maker. Figure 5.1 also reveals that models with indistinguishable AUCs can differ significantly in terms of *calibration*, which is also indicated that traditional supervised learning algorithms developed under cohort study theory are not suitable for personalized clinical decision

even with a good AUC. For instance, a classifier can draw a perfect decision boundary but fails to estimate the true risk about individual patients.

As personalized medicine is inevitable [145], for the best performance of diagnostic and treatment for individuals, physicians usually conduct individualized lab tests to determine diseases or conditions of their patients instead of relying on their knowledge of diagnostic groups at large. Analogously for the risk estimation problem, careproviders need personalized risk scores rather than the risk of a group to make more informed decisions regarding individual patients. That is, if careproviders can determine the risk of their patients in developing a disease or condition individually, they can treat them with respect to their own situation to improve the quality of their services.

Unfortunately, traditional supervised learning algorithms have focused on models' *discriminative* ability and ignored their *calibration* performance. For instance, Decision Trees (DT) and Support Vector Machine (SVM) provide a decision boundary for a cohort study, but their outputs do not closely represent the "true" probability of adverse events, as indicated in Figure 5.2.

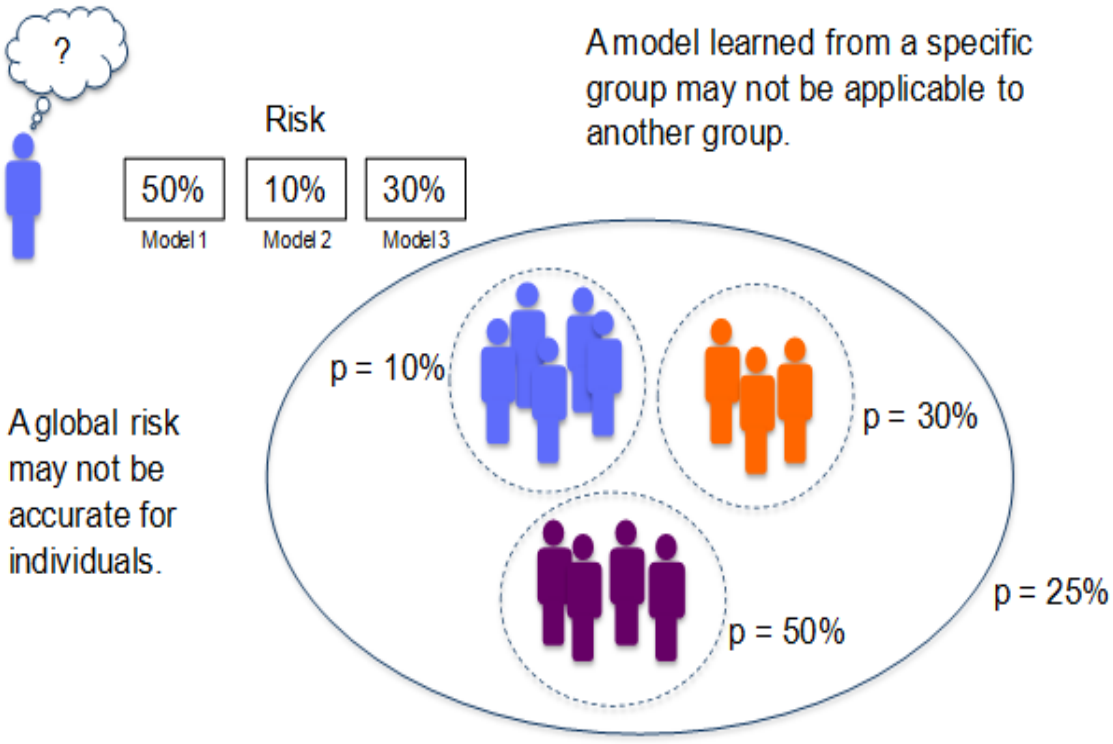


Figure 5.2: Traditional learning approaches are not appropriate for personalized medicine as global risks make little sense to individuals.

In order to evaluate different courses of actions, it is useful to obtain accurate likelihood estimates of the alternatives [45]. This *calibration* aspect of models is crucial to biomedical decision making tasks, because only "*calibrated*" outputs can be used for individual risk assessment and further combined with other pieces of information in a consistent manner for personalized clinical decision making.

Clinically, risk assessment tools such as Cox proportional hazard model, logistic regression model, and other machine-learning based predictive models are widely used in patient diagnosis, prognosis, and clinical studies. Accurate *calibration* of these models is important if the outputs are going to be applied to cohorts other than those the model was initially built upon [45, 160].

One example is the Gail model, a prediction model of a woman's risk of developing invasive breast cancer. However, the Gail model is reported to underestimate the risk among women over 50 years old and high-risk populations such as patients with family history or atypical hyperplasia. The calibrated model found more patient that would benefit from chemoprevention than the original model [5, 160].

Another example is the Framingham Heart Study model, a gender-specific coronary heart disease (CHD) prediction functions to assess the risk of developing incident CHD in a white middle-class population. While the original model overestimated the risk of 5-year CHD events among Japanese American men, Hispanic men and Native American women, the re-calibrated risk score based on the new cohort's own average incidence rate performed well [50].

In summary, *calibration* is fundamental to achieving consistency of measurement [27]. Thus, methods have to provide a reliable estimate of the "true probability" of the class membership for application to many clinical decision making problems.

To face this challenge, this chapter is dedicated to investigating the possibility of providing an integrated framework that concurrently bears both *discrimination* and *calibration* criteria. However, this is a non-trivial task, as Habbema and his colleagues denoted: "We would like to devise a statistic which reflects *discriminatory* ability and reliability (*calibration*) at the same time, the two ingredients being mixed in justifiable proportions. But we have not been able to do so. Nor are we able to prove that it cannot be done [81]."

Through careful investigation of probabilistic method's reliability and validity, I revealed insufficiency of state-of-the-art *calibration* approaches (e.g., Platt Scaling (PS) [147] and Isotonic Regression (IR) [9]) and why they could have poor performance in various situations. As opposed to these methods using a post-processing step to "*recalibrate*" probabilistic outputs [37, 176, 205, 206], I developed a unified framework

that synthesized both model optimization objectives as a global optimization problem. I derived a simple formulation that accounts for the joint optimization problem. My framework incorporates, extends and improves state-of-the-art *calibration* approaches (e.g., Platt Scalingand Isotonic Regression). I showed both approaches are degenerated instances of my framework under parametric and nonparametric assumptions, respectively.

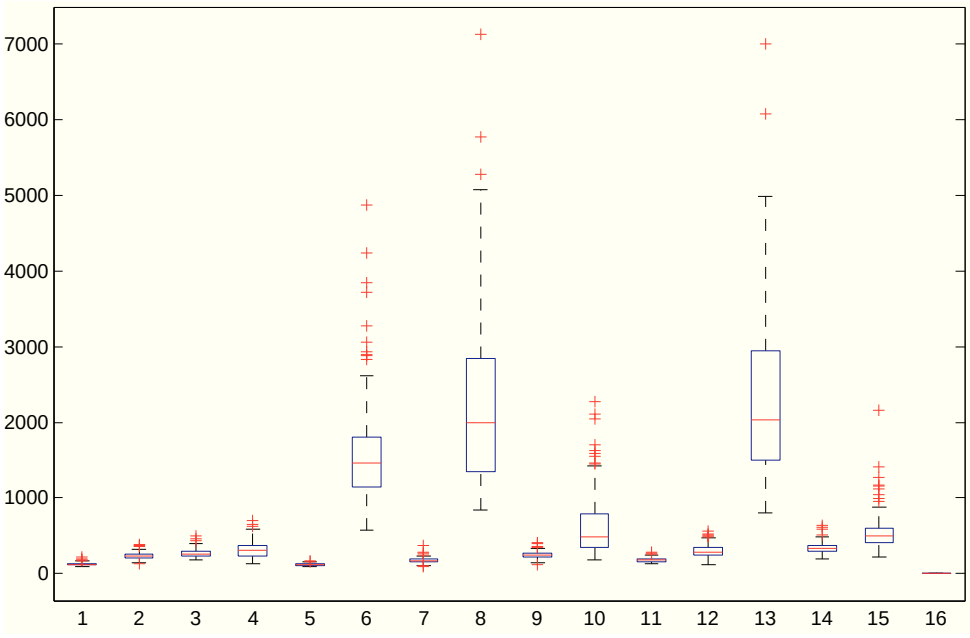
I also implemented a generalized method under this integrated framework, which extends traditional Support Vector Machines to consider a *calibration*-related loss function in addition to a discriminative hinge loss function. The new model is called "Doubly Penalized Support Vector Machine", as it is regularized by two rather than one loss functions. Finally, I evaluated my approach using clinical related data and demonstrated improved performance.

5.1 Data

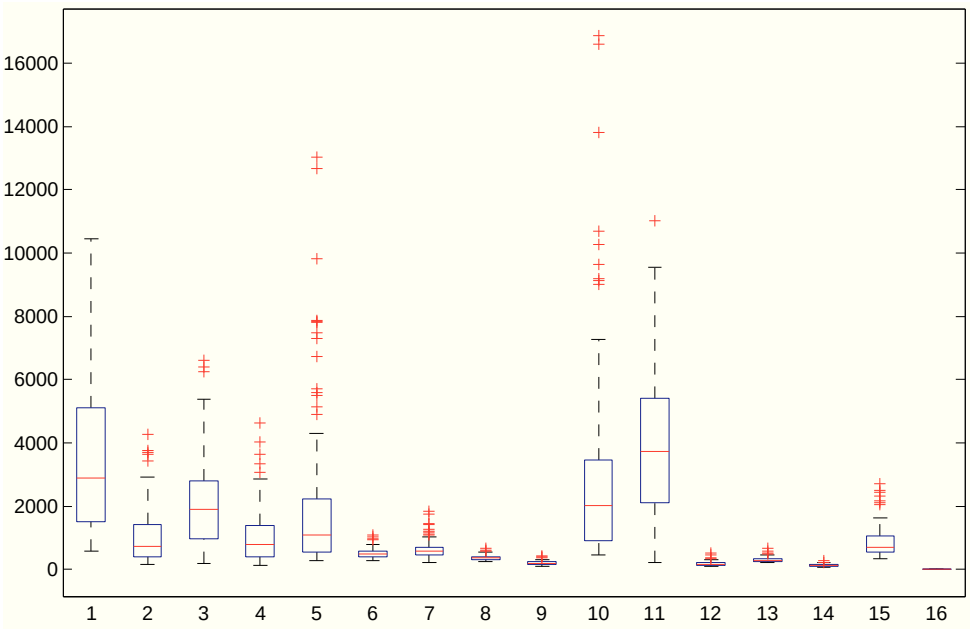
In additional to simulated data, I used a clinical dataset, Breast Cancer Gene Expression, to evaluate the efficacy of my approach in this chapter. Specifically, this dataset contains GSE2034 and GSE2990, which are both gene expression data related to breast cancer. The dataset was obtained from the NCBI Gene Expression Omnibus (GEO). Two individual data downloaded were previously studied by Wang et al. (GSE2034) [187] and Sotiriou et al. (GSE2990) [166], respectively.

To make my data comparable to the previous studies, I follow the criteria in [140] to select patients, who did not receive any treatment and had negative lymph node status. Among these pre-selected candidates, only patients with extreme outcomes, either poor outcomes (recurrence or metastasis within five years) or very good outcomes (neither recurrence nor metastasis within eight years) were selected. The number of samples after filtering are: 209 for GSE2034 (114 good/95 poor), 90 for GSE2990 (60 good/30 poor).

Both data have a feature size of 247,965, corresponds to the gene expression results obtained from microarray experiments. They were preprocessed to keep only the top 15 features ranked using a t-test (see [140] for details). I showed boxplots and matrix plots for both data in Figure 5.3 and Figure 5.4, respectively.

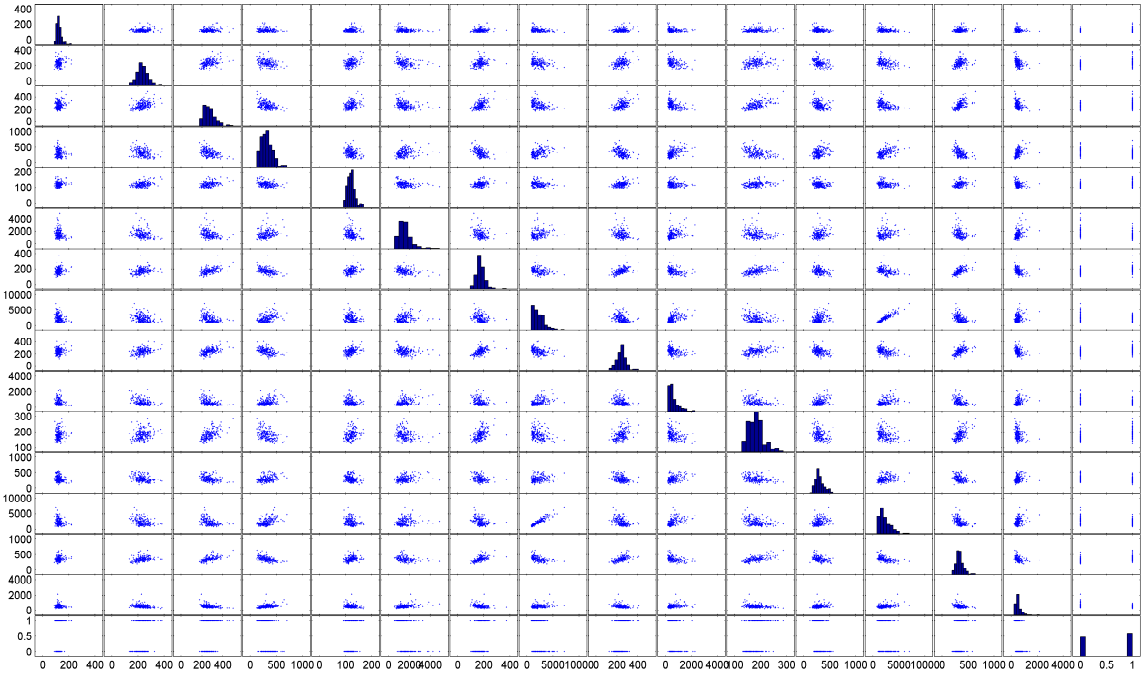


(a) GSE_2034

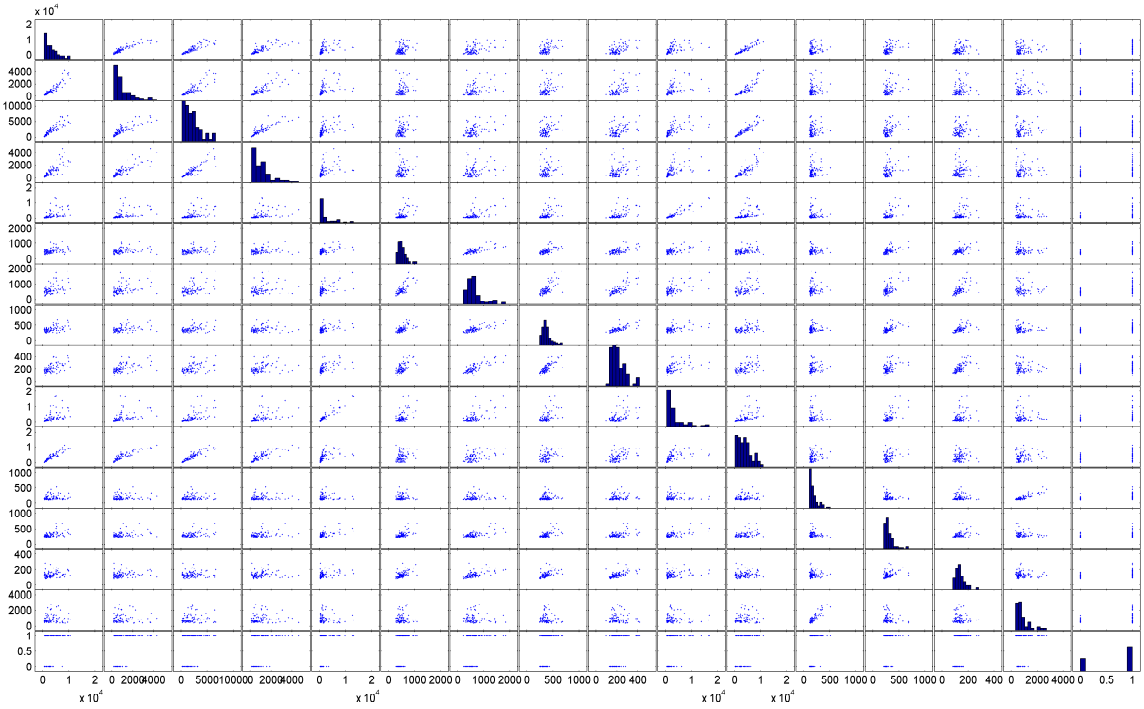


(b) GSE_2990

Figure 5.3: Boxplots of Breast Cancer Gene Expression Data. Each column corresponds to one feature vector, and the last column indicates the outcome variable.



(a) GSE_2034



(b) GSE_2990

Figure 5.4: Matrix plots of Breast Cancer Gene Expression Data. Each subfigure corresponds to a matrix plot of one data set.

5.2 Related Works

Classifier *calibration* is the process of converting classifier scores into reliable probability estimates [62]. Various procedures for *calibrating* classifiers have been proposed in different contexts: in forecasting the I/O response time of large storage arrays [45], in pattern classification [60, 136] and in game theory [67].

In general, existing *calibration* methods are mostly post-processing steps that applies to a probabilistic model by "rectifying" their outputs. Among *calibration* approaches have been proposed, there are two main streams: the parametric models and non-parametric models. Platt suggested transforming SVM predictions to posterior probabilities by passing them through a sigmoid function $P(y_i =' +'|\hat{y}_i) = \frac{1}{1+\exp(A\hat{y}_i+B)}$, where y_i corresponds to the class label, $\hat{y}_i$ is the prediction probability and A, B are the model parameters [147]. The idea is to use a sigmoid function to map model predictions to posterior probabilities that minimize the errors to the ground-truth. The parameters of the sigmoid function can be efficiently estimated by the gradient descent algorithm. However, it is problematic when the outputs cannot be fitted with a sigmoid function; in this case, approximations could result in poor *calibration* results.

Recently, a *calibration* technique based on Isotonic Regression has gained attention within machine learning as a flexible and effective way to calibrate classifiers [207]. This non-parametric model aims to find a weighted least square fit with the following form: $\min \sum_i w_i(x_i - y_i)^k$ subject to $x_i \geq x_j, \forall i > j$. Note that k is the order of the norm, w_i corresponds to the weight, y_i is the binomial class label and x_i is the calibrated estimation variable. The formula is subject to a set of monotonicity constraints giving a simple or partial order over the variables. When $k = 2$, there is an efficient pair-adjacent violators (PAV) algorithm to solve this problem [26]. However, the result of the *calibration* is not continuous and the result could be easily over-fitting.

> **Interpretation:** If classification is the goal, it is generally recommended to choose a model with good discrimination over one with good calibration. If a predictive model has poor discrimination, no adjustment or calibration can correct the model. On the other hand, if discrimination is good, but calibration is poor, the model can be re-calibrated without sacrificing the discrimination. [41]

Figure 5.5 illustrates both *calibration* approaches using a made-up example. The blue circles correspond to un-calibrated probability estimates and the red dashed curves represent the calibrate transformations. The Platt Scaling *calibration* is more smooth but does not represent the underlying pattern sufficiently while the

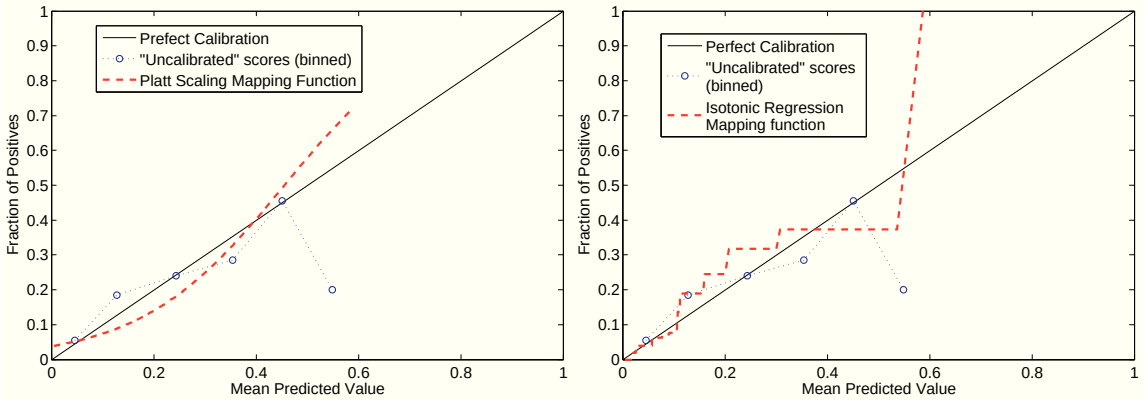


Figure 5.5: State-of-the-art *calibration* approaches.

Isotonic Regression seems quite zigzag and tends to over-fit.

Figure 5.6: Existing approaches of *calibration*. These methods take place in probabilistic outputs of "uncalibrated" classifiers. They are executed in a sequential manner, that is, they first optimize a model that maximize *discrimination* and its outputs are calibrated by these "rescue" methods.

> **Interpretation:** Existing *calibration* approaches apply monotonic mapping functions on outputs of probabilistic models that maximize AUC in a sequantial manner. However, their capabilities are limited by two things: 1) the monotonic constraints; 2) the seperated considerations of *calibration* and *discrimination*.

Although many empirical results have demonstrated benefits of *calibration*, the proposed models Platt Scaling and Isotonic Regression are both post-processing methods that work with outputs of other *discriminative* classifiers as a separate step rather than taking *calibration* into account concurrently with *discrimination* at the very beginning.

A theoretical foundation for joint optimization was still missing, and it was unclear if a global maximization of *discrimination* and *calibration* could impact the models in a positive way. This chapter aims

to address the joint optimization of *discrimination* and *calibration* in a principled way and suggest novel approaches for implementing this framework.

5.3 Preliminaries

I briefly review *discrimination* (AUC) and *calibration* before introducing the methodology details. Figure 5.7 demonstrates both evaluations metrics with a simple example.

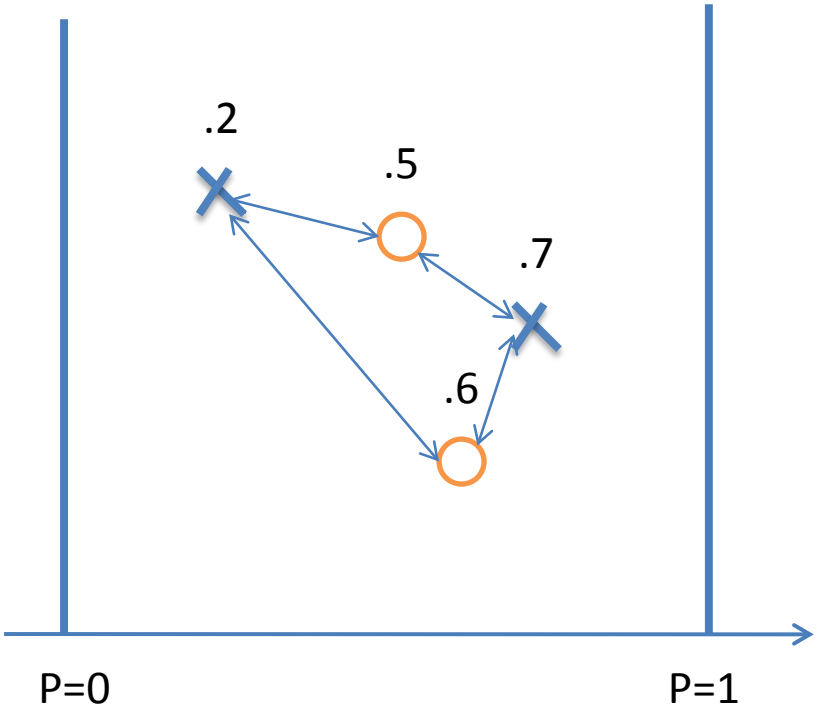


Figure 5.7: Illustration of *discrimination* and *calibration* measures on a made-up example. The "X" represents positive-labeled observations and "O" corresponds to negative-labeled observations. There are two discordant pairs out of four, thus $AUC = 0.5$ and the L-1 *calibration* error equals to $\frac{0.5+0.6+|0.2-1|+|0.7-1|}{4} = 0.55$.

Interpretation: The seemingly unrelated model evaluation metrics: *discrimination* and *calibration* are indeed correlated. Generally, good calibration encourages good discrimination and vice versa.

5.3.1 Area Under Curve (AUC)

Area Under ROC Curve (AUC) is often used as a measure of quality of a probabilistic classifier, e.g., a random classifier like a coin toss has an AUC of 0.5; a correct classifier has 1. Every point on a ROC curve corresponds to a unique pair of True Positive Rate (TPR), and False Positive Rate (FPR); please refer to Figure 5.1(c). I can thus express this area as the following integration of TPR over FPR:

$$\begin{aligned} AUC &= \int_0^1 \frac{TP}{P} d\frac{FP}{N} \\ &= \frac{1}{mn} \int_0^n TP dFP \\ &= \frac{1}{mn} \sum_{X \in \{+\}} \sum_{O \in \{-\}} \left(P(X) > P(O) \right), \end{aligned} \tag{5.1}$$

where $P(X)$ and $P(O)$ correspond to the posterior probability of positive sample X and negative sample O, respectively. m and n correspond to the cardinality of the positive and negative classes. The last line of Eq. 5.1 can be interpreted as the count of concordant pairs out of all positive and negative sample pairs. For example, if all the positive samples are ranked higher than any of the negative samples, the AUC equals to 1; conversely, if none of the positive samples is ranked higher than any of the negative samples, the AUC equals to 0.

Obviously, AUC concerns the relative ordering of the estimated probabilities rather than their actual values. I can divide the probabilities of a classifier A by 10 without affecting AUC at all, as indicated by Figure 5.1. This operation, however, makes the probabilistic predictions unreliable as $\sum_{X \in \{+\}} P(X) \ll \frac{m}{m+n}$ and $\sum_{O \in \{-\}} P(O) \ll \frac{n}{m+n}$. I now introduce the concept of *calibration*.

5.3.2 Calibration

Calibration is a standard to evaluate whether a probabilistic classifier is suitable. Recall that a probabilistic classifier assigns a probability p_i to each sample i; a well calibrated classifier is that a fraction of about p_i of the events with predicted probability p_i actually occurs. Expressed with a parametric definition,

$$y_i = \text{true probability}, \quad p_i = \text{predicted probability}$$

When there are not many samples with the same probability, samples with similar probabilities are grouped by partitioning the range of predictions into sub-segments (or bins). To estimate the unknown true probabilities for many real problems, the prediction space is divided into ten bins. Cases with a predicted value between 0 and 0.1 fall in the first bin, between 0.1 and 0.2 in the second bin, and so on. For each bin, the mean predicted value is plotted against the true fraction of positive cases. If the model is well calibrated the points will fall near the diagonal line. This plot is known as the reliability diagram, as indicated by Figure 5.5.

I accessed the *calibration* with the goodness-of-fit-test: Hosmer-Lemeshow (HL) [90, 103], which verifies whether the model of interest represents the truth labels with a high statistical significance under a χ^2 test.

5.4 Methodology

I believe that *discrimination* and *calibration* are not independent perspectives of a probabilistic model. Thus, previous models that adopt sequential processing cannot offer the best of both. My exploration indicated that better models could benefit from considering *calibration* and *discrimination* together. The following subsection introduces the integrated framework.

5.4.1 An Integrated Framework

Assume there are m positive cases and n negative cases. Let us denote $x = P(X)$ as the probability prediction of an instance $X \in \{+\}$ of the positive cases, and $o = P(O)$ as the prediction probability of another instance $O \in \{-\}$ of the negative cases. Every pair (x, o) resides in a space of $\mathcal{X} \times \mathcal{O}$. Note the cardinality of $|\mathcal{X}| = m$ and $|\mathcal{O}| = n$. Consider the following formulation:

$$\sum_{(x,o)} \left(\frac{e(x)}{n} + \frac{e(o)}{m} + f(x, o) \right) \quad = \quad mn, \tag{5.2}$$

which is equivalent to:

$$\sum_{(x,o)} \left((\frac{e(x)}{n} + \frac{e(o)}{m})/mn \right) + \sum_{(x,o)} \frac{f(x,o)}{mn} \quad = \quad 1, \tag{5.3}$$

where $e(x) = |x - 1|$ and $e(o) = |o|$ denote the prediction errors. Notice that I divide the errors by the size of the opponent class size to weight their contributions.

> **Interpretation:** I constrained the sum of *discrimination* and *calibration* metrics to be mn, the total number of sample pairs. The goal is to find the tradeoff between these two metric within the simplex.

Let $I_{Err} = \left(\sum_{(x,o)} \frac{e(x)}{n} + \frac{e(o)}{m} \right) /mn$, since the errors are calculated pairwise, I can rewrite it as

$$\begin{aligned} I_{Err} &= \left(\sum_{(x,o)} \frac{e(x)}{n} + \frac{e(o)}{m} \right) /mn \\ &= \frac{\sum_{x \in \mathcal{X}} e(x) + \sum_{o \in \mathcal{O}} e(o)}{mn}, \end{aligned} \tag{5.4}$$

which corresponds to the *normalized calibration error* under L_1 norm of the finest granularity, refers to Section 5.3.2.

I define another function called compliment AUC:

$$C_{AUC} \quad = \quad \sum_{(x,o)} \frac{I(x < o)}{mn} = 1 - I_{AUC}, \tag{5.5}$$

where $I(x < o) = \begin{cases} 1 & x < o \\ 0 & \text{otherwise} \end{cases}$ and I_{AUC} denotes the area under curve.

Now assume that the function $f(x, o)$ can be decomposed so that $f(x, o) = I(x < 0) + \epsilon(x, o)$. I can plug Equation 5.4 and Equation 5.5 into Equation 5.3 to obtain:

$$I_{Err} + I_{AUC} + \sum_{(x,o)} \frac{\epsilon(x,o)}{mn} \quad = \quad 1, \tag{5.6}$$

after re-organizing it, I obtain

$$\begin{aligned} \sum_{(x,o)} \frac{\epsilon(x,o)}{mn} &= 1 - I_{Err} - C_{AUC}, \\ &= I_{AUC} - I_{Err}, \end{aligned} \tag{5.7}$$

which is the function that depicts the trade-off between *calibration* and *discrimination*. Ideally, I want larger AUC (*discrimination* power) and smaller normalized *calibration* error, which is exactly the procedure for maximizing $\sum_{(x,o)} \frac{\epsilon(x,o)}{mn}$.

> **Interpretation:** Equation 5.7 implies that *discrimination* and *calibration* are inter-dependent and thus their optimum values can be obtained simultaneously.

Assume that a probabilistic classifier outputs near-optimal I_{AUC}, in order to increase the objective in Equation 5.7, I want to transform the probability by keeping this I_{AUC} but reduce the I_{Err} as much as possible at the same time. In order word, I am seeking a transformation function $t(\cdot)$ to minimize I_{Err} while keeping the partial orders so that if $x < o$, then $t(x) < t(o)$ and vice versa.

$$\max_{t} \sum_{(x,o)} \epsilon\left(t(x), t(o)\right). \tag{5.8}$$

It is easy to prove that any monotonic function preserves the partial ordering and thus keeps the I_{AUC}. In this case, the optimization problem of Equation 5.8 is reduced to the Isotonic Regression. I can also address it with a parametric treatment using a sigmoid function. The problem induced by Equation 5.8 becomes the Platt Scaling method. Obviously, both Platt Scaling and Isotonic Regression are the AUC-preserving instantiations under the integrated framework. The Isotonic Regression, as a non-parametric model, gives the optimal solution on the training data under the monotonicity constraints. It usually has a performance advantage over the Platt Scaling on testing samples as well because it offers a higher degree of freedom to fit the data.

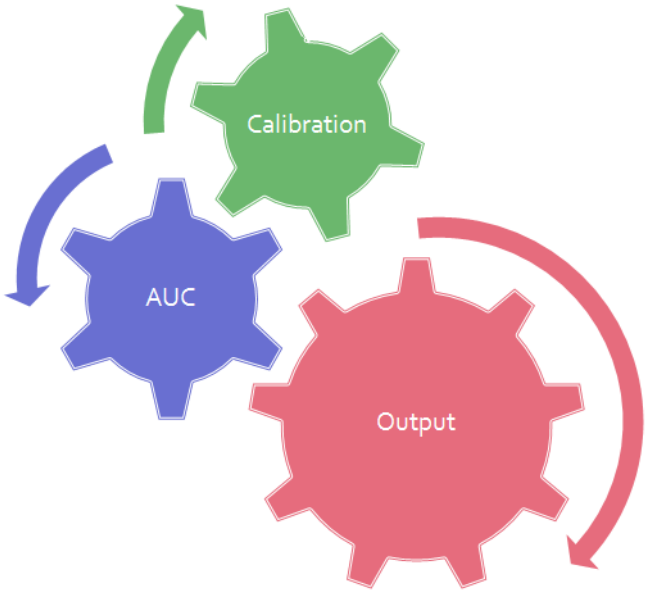


Figure 5.8: The integrated framework for joint optimization. The framework enforces maximization of both AUC and *Calibration* in a global manner.

> **Interpretation:** Good probabilities = good calibration + good discrimination. As these two aspects are complementary, they should be taken together to ensure good probabilities.

5.4.2 A Joint Optimization Implementation

To validate my framework, I implemented an instance of the framework to show the benefits of considering *calibration* and *discrimination* simultaneously. My model extended the Support Vector Machine to incorporate the *calibration*-related loss in addition to the *discrimination*-related loss. To understand the technique, I briefly review the Support Vector Machine model developed by Vanpik et al. [180] before introducing the details of my model.

5.4.2.1 Support Vector Machine

Suppose there is a training data $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\} \subset \mathcal{X} \times \mathcal{R}$, where $\mathcal{X}$ denotes the space of input patterns (e.g. $\mathcal{X} = \mathcal{R}^d$). $Y_i \in \{-1, 1\}$. The label "1" indicates a positive case and "-1" indicates a negative case. A natural way of representing a maximum margin classification optimization led to the

following format:

$$\min \ \frac{1}{2}||W||_2^2 + C\sum_i L(sign(W^T X_i + b), Y_i) \tag{5.9}$$

Here $L(sign(W^T X_i + b), Y_i)$ is the $0/1$ loss while $||W||_2^2$ is the penalty term that specifies the maximum margin between two classes of data. C trades off between model bias and variance. Unfortunately, Equation 5.9 is not convex. A relaxation of the problem leads to the using of a hinge loss function to approximate the $0/1$ loss. This result is known as the Support Vector Machine (SVM), whose objective function is $f(X) = W^T X + b$. The function thus optimizes solutions that deviate least from the ground-truth Y. It has the following form:

$$\begin{aligned} \min_{W,b,\xi,\xi^*} \frac{1}{2} W^T W + C\sum_{i=1}^{N} \xi_i \\ \text{s.t. } Y_i * (W^T X_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, \forall i, \end{aligned} \tag{5.10}$$

where ξ_i is the loss for the i-th data point $\mathbf{X}_i$; $\mathbf{W}$ and b are the weight parameters; and C is a penalty parameter to weight the loss. I reorganize the Equation 5.11 by absorbing the constraints to the target function as follows:

$$\min_{W,b,\xi,\xi^*} \frac{1}{2}||W||^2 + C\sum_{i=1}^{N} \max(1 - Y_i * (W^T X_i + b), 0). \tag{5.11}$$

> **Interpretation:** The Support Vector Machine trade off model between model smoothness and misclassification errors.

The first term $\frac{1}{2}||W||^2$ is responsible for the model complexity while the second term $\max(1 - Y_i * (W^T X_i + b), 0)$, known as the hinge loss, penalizes the model for its mis-classifications. SVM expects label "1" cases to be $f(X) > 0$ and label "0" cases to be $f(X) < 0$. The final output of this optimization is a vector of weight parameters W, which forms the decision function maximizing the margin between positive and negative samples. Figure 5.9 illustrates the separating hyperplane and maximum margin optimization.

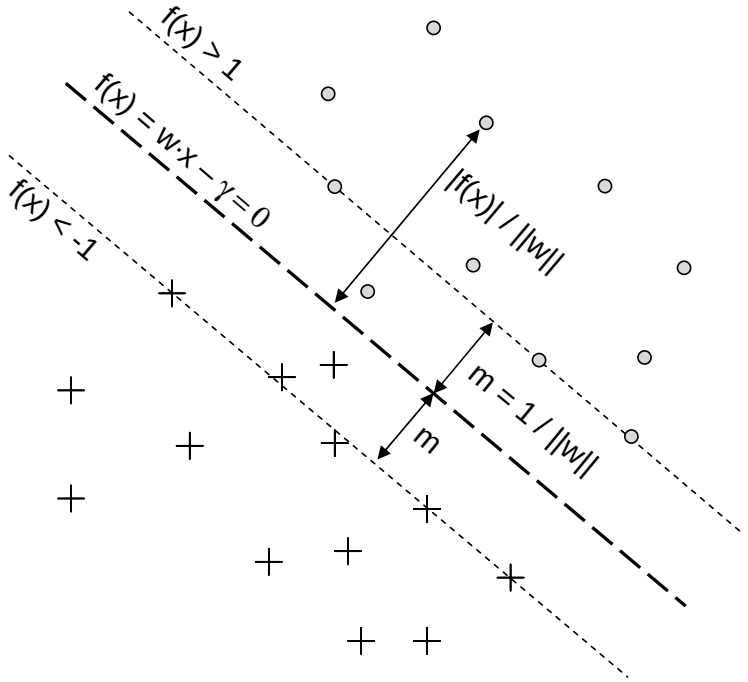


Figure 5.9: The separating hyperplane that maximizes the margin. ("o" is a positive data point, i.e., $f(\text{"}o\text{"}) > 0$, and "+" is a negative data point, i.e., $f(\text{"}+\text{"}) < 0$.

5.4.2.2 Doubly Penalized Support Vector Machine

Thanks to the hinge loss function that directly optimizes the *discrimination* ability, SVM suffices in many tasks the mission of which is to provide good classification. However, SVM does not offer reliable outputs that can be trusted as good approximations of the "true probability" of an event. The model concentrates *discrimination* ability separately from the consideration of *calibration*. That is, it does not penalize a correct case even if the predicted value is far from its label.

Platt et al. proposed transformation $P(Y_i =' +'|\hat{Y}_i) = \frac{1}{1+\exp(A\hat{Y}_i+B)}$ to rectify these "uncalibrated" outputs $\hat{Y}_i = W^T X_i + b$, where Y_i corresponds to the true class label, and A, B are the model parameters [147]. Essentially, this approach refits SVM outputs to a one-dimensional Logistic Regression model that optimizes the log-loss function. Nevertheless, such an ad-hoc approach does not always work because the monotonicity constraint of a sigmoid function limits the *calibration* ability it offers.

To address this insufficiency, I developed a hybrid model that concurrently optimizes loss functions of two different kinds: the hinge loss and the regression loss. I illustrated three typical machine learning loss functions in Figure 5.10 with their semantic meaning to help readers to understand the idea of this

combination.

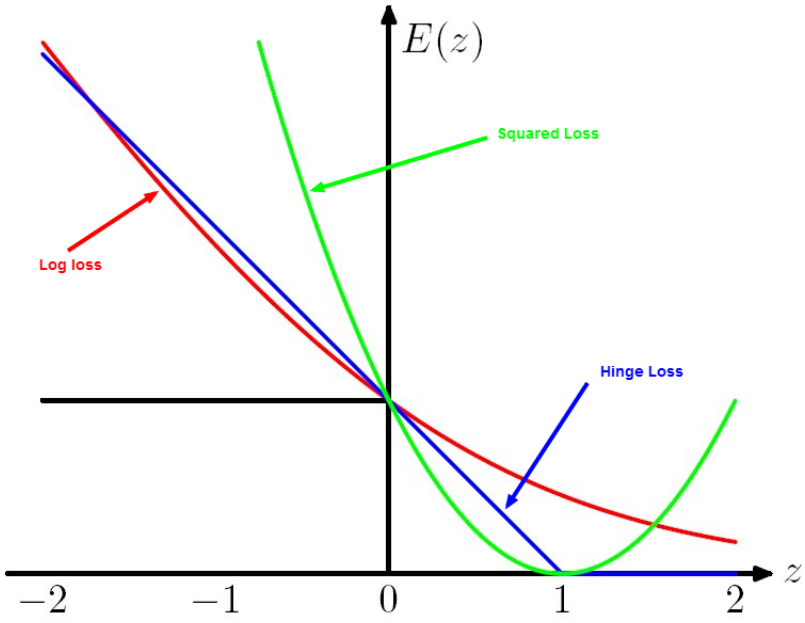


Figure 5.10: Illustration of three different loss functions. The red curve corresponds to the log-loss function (logistic regression), the green curve corresponds to the squared loss function (ridge regression) and the blue curve corresponds to the hinge loss function (SVM).

Different loss functions bear different optimization principles. For example, optimizing hinge loss means predicting the conditional median of unknown states Y; optimizing log-loss means minimizing the description length of Y; and optimizing squared loss means predicting the conditional mean of Y. The log-loss is implemented in Logistic Regression, in which the best hypothesis for a given set of data is what leads to the best compression of the data. The squared loss can be implemented in ridge regression, which is closely related to a *calibration* measurement: the Brier Score. Finally, the hinge loss function is implemented in the Support Vector Machine model to maximize the *discrimination* ability. Understanding these differences, I intentionally combine loss functions with *discrimination* and *calibration* semantics for a global optimization. The following model combines two different loss function: hinge loss and squared loss. The first is responsible for optimizing AUC and the latter is responsible for minimizing the Brier score:

$$\min_{\mathbf{w},b,\xi,\xi^*} \frac{1}{2}\mathbf{W}^T\mathbf{W} + C_1 \sum_{i=1}^{N} \xi_i + C_2 \sum_{i=1}^{N} (\mathbf{W}^T\mathbf{X}_i + b - Y_i)^2 \tag{5.12}$$

$$\text{s.t. } Y_i * (\mathbf{W}^T\mathbf{X}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, \forall i.$$

Interpretation: The doubly penalized Support Vector Machine combines two loss functions: one for *calibration* and the other for *discrimination*. The model optimizes both objectives concurrently.

The DP-SVM model extends SVM by considering *calibration*-related square loss. On one hand, computational learning theory ensures its sound generalizability to the marginal maximization property. On the other hand, thanks to the regression loss objective, the new model offers *calibration* ability that SVM cannot provide. The parameters C_1 and C_2 control the weights of *calibration* and *discrimination* impacts in building the model. To solve this quadratic programming objective (Equation 5.12), I developed the following algorithm to use subgradient descent optimization approach.

Algorithm 3 Parameter learning for DP-SVM using subgradient descent algorithm.

Input: Raw observations $\mathbf{X}_i$ where $\mathbf{X}_i = \{x_i^1, \ldots x_i^m\}$ are the co-variate vector at time i.
Output: Learned Weight Parameters $\mathbf{W}$.
Parameters: The penalty parameter for hinge loss C_1, the penalty parameters for ridge regression loss C_2 and step size η

1: Reorginize objective function to get rid of the constraints:
$\frac{1}{2}\mathbf{W}^T\mathbf{W} + C_1 \sum_{i=1}^{N} \max 0, 1 - Y_i * (\mathbf{W}^T\mathbf{X}_i + b) + C_2 \sum_{i=1}^{N} (\mathbf{W}^T\mathbf{X}_i + b - Y_i)^2$,
2: Calculate the derivative of the above objective, which can be written as
$G(\mathbf{W}) = \mathbf{W} + C_1 * \sum_i \max(1 - Y_i X_i, 0) + 2 * C_2 * \sum_i \mathbf{W} X_i$.
3: Initialize $\mathbf{W}^0 = 0$.
4: **repeat**
5: $\mathbf{W}^t = \mathbf{W}^{t-1} - \eta * G(\mathbf{W}^{t-1})$ the derivative of $\mathbf{W}$
6: **until** $\mathbf{W}$ converges

5.5 Experiments

I evaluated the approaches using: (GSE2034) [187], Sotiriou et al. (GSE2990) [166] from the Breast_Cancer clinical dataset, which were collected from NCBI Gene Expression Omnibus (GEO). Please refer to Chapter

2 for details. The experiments are carried out on a 2.00 Ghz laptop with 2GB memory. Both data-sets have been preprocessed to keep only the top 15 features, see [140] for details. The data-sets are summarized in Table 5.1.

Table 5.1: Split of the training and test data for GSE2034 and GSE2990.

	#ATTR	TRAIN SIZE	TEST SIZE	%POS
GSE2034	15	125	84	54%
GSE2990	15	54	36	67%

I qualitatively examined the *calibration* and *discrimination* of the Logistic Regression, SVM and Doubly Penalized Support Vector Machine. For both data, I trained on the 60% of the random samples. Table 5.2 and 5.3 show the performance of three models in comparison.

Table 5.2: Model performance comparison of GSE2034.

	LR	SVM	DP-SVM
AUC	0.716247	0.70595	0.737986
AUC s.d.	0.05539	0.056137	0.053685
AUC c.i. (upper)	0.607684	0.595923	0.632766
AUC c.i. (lower)	0.82481	0.815976	0.843206
F-score	0.692308	0.675	0.675
Sensitivity	0.586957	0.586957	0.586957
Specificity	0.868421	0.815789	0.815789
Type1 Error	0.413043	0.413043	0.413043
Brier Score	0.25163	0.241392	0.236004
P-value HL-H test	0	1.33E-15	1.19E-06
P-value HL-C test	0	0	1.74E-07

Table 5.3: Model performance comparison of GSE2990.

	LR	SVM	DP-SVM
AUC	0.772569	0.822917	0.861111
AUC s.d.	0.078327	0.069061	0.060775
AUC c.i. (upper)	0.619052	0.68756	0.741994
AUC c.i. (lower)	0.926087	0.958273	0.980228
F-score	0.833333	0.818182	0.808511
Sensitivity	0.833333	0.75	0.791667
Specificity	0.666667	0.833333	0.666667
Type1 Error	0.166667	0.25	0.208333
Brier Score	0.194454	0.17017	0.14737
P-value HL-H test	0	0.161298	0.278179
P-value HL-C test	0	0	0.204817

The DP-SVM model outperformd other methods in both indices and demonstrated its advantage.

5.6 Discussion

Discrimination and *calibration* are two major families of measurements in the evaluation of model performance. In clinical predictions, *discrimination* measures the ability of a model to separate patients with different outcomes. In the case of a binary outcome, good *discrimination* indicates adequate distinction in the distributions of predicted values, based on the model, between the two classes, defined by the binary outcome [41]. On the other hand, *calibration* measures the similarity between the predicted values and the observed outcomes. Usually, good *calibration* tends to correspond to good *discrimination*. However, a successful classifier in terms of *discrimination* ability or Area Under ROC Curve (AUC) is not necessarily calibrated. *Calibration* is crucial to biomedical decision making tasks, because only "calibrated" outputs can be used for individual risk assessment and further combined with other pieces of information in a consistent manner for personalized clinical decision making. Unfortunately, traditional supervised learning algorithms developed under the theory for a cohort study do not provide reliable individualized risk prediction to assist the caregivers.

Recent research in machine learning has advocated using a post-processing step to "recalibrate" uncalibrated models for better performance. However, both approaches are not sufficient. For example, Platt Scaling does not always fit the inputs; thus, approximation under such situations may lead to bad *calibration*. On the other hand, Isotonic Regression has another problem: it could easily over-fit training data due to the absence of a regularization term. In summary, both Platt Scaling and Isotonic Regression are post-steps applied to the outputs of a probabilistic model without considering the input space, which limits their performances.

Furthermore, a theoretical foundation for joint optimization was still missing and it is unclear if a global maximization of *discrimination* and *calibration* could impact the models in a positive way. After carefully analyzing the relations between these metrics, I suggested an integrated framework that combines *calibration* and *discrimination* under a global optimization objective. The framework incorporates, extends and improves existing approaches such as Platt Scaling and Isotonic Regression.

I also developed a novel technique, Doubly Penalized Support Vector Machine (DP-SVM), which instantiates the unified framework to provide more comprehensive joint optimization than previous methods

do. However, there is a limitation of this approach in which it requires additional steps to control the relative impact of *discrimination* and *calibration*. The user has to tune parameters, e.g., the trade-off factor, by cross validation.

Thus, although DP-SVM demonstrated good performance, it is not fully automatic and might not be practically attractive to clinicians. To minimize human intervention, I developed other *calibration* approaches that utilize model specific characteristics to improve *calibration* performance. The following chapters will discuss and demonstrate two alternative methods.

5.7 Conclusion

Empirical experiments in previous studies have demonstrated the benefits of calibrating probabilistic models for more effective individualized inspection. However, a theoretical foundation for understanding the relationship between *calibration* and a more traditional optimization criteria, *discrimination*, is still not available. It remained unclear whether the joint consideration of both could improve or harm the model performance. To address these issues, I investigated further in the direction of global optimization methods for both metrics.

I developed a systematic framework for *discrimination* and *calibration*. In a principled way, I integrated two important components of the probabilistic model: *discrimination* and *calibration*, which are traditionally considered separately. Through my investigation, I found that these two seemingly unrelated metrics are connected, and a well designed joint maximization algorithm can offer the best of both if they are optimized independently. This joint optimization using a combined objective function guards against learning degenerated model that performs well in one aspect but poorly in the other aspect.

My framework incorporates, extends, and improves state-of-the-art *calibration* approaches such as Platt Scaling (PS) and Isotonic Regression (IR). I also implemented a more generalized method under this integrated framework, which extends traditional Support Vector Machines to consider a *calibration*-related loss function in addition to a discriminative hinge loss function. The model is generalizable, and can be applied to any binary outcome prediction problem where a probabilistic outputs is preferred over a decision rule. Due to the joint contribution of *discrimination* and *calibration* specific loss function, it is capable of producing more accurate probabilities than models that consider a single aspect. I evaluated my approach using clinical related data and demonstrated improved performance.

Chapter 6

Smooth Isotonic Regression

[1]The diagnosis of human disease has witnessed steady improvement in the past few centuries has been in steady increase. Once considered a single entity, many diseases have been divided into finer categories based on the response to treatment (e.g., I type and type II diabetes), genetic (such as familial polyposis and non-familial), histology (such as small cell lung cancer) and the recent transcriptional profiling (such as leukemia, lymphoma) [80]. The next frontier in medicine seems to be the patient type, not the disease, i.e. disease X in person of type Y. This is also called "Personalized Medicine", which aims to predict the type of individual, along with the therapy that is most likely to benefit him or her. That is, personalized medicine needs "*calibrated*" probabilistic estimates about individuals.

Unfortunately, most popular supervised learning models focus on model *discrimination* ability while neglecting the *calibration* ability [12]. In the last chapter, I introduced a few recent attempts to address the *calibration* issue. However, they are not sufficiently developed to meet the needs of personalized clinical decision making [87]. For example, Devised by Platt [147], Platt scaling may lead to poor *calibration* when outputs cannot fit a sigmoid function. Isotonic Regression *calibration* approach, proposed by Zadrozny [207], suffers from monotonicity constraints and its results tend to overfit the training data. To this end, I developed a novel model DP-SVM to synthesizes both *discrimination* and *calibration* objectives for a global maximization. The method has not only demonstrated superior *calibration* and *discrimination* performance but also introduced additional parameters to be determined by users. This additional burden on users, most likely caregivers, makes the method unlikely to gain much attraction in the biomedical community.

[1]A version of this chapter is under review at 2010 AMIA Summit on Clinical Research Informatics [94].

To avoid human intervention and provide a useful *calibration* tool, I investigated an alternative approach to achieve both *discrimination* and *calibration* goals in an automatic manner. After a careful analysis of existing parametric and non-parametric models, my investigation indicated that these aspects could be complementary to each other. Specifically, I used outputs of isotonic function as inputs to a smoothing function minimized over the space of natural cubic splines with knots at the design points. Finally, I estimated unknown parameters of this Piecewise Cubic Hermite Interpolation Polynomial. The new regression function utilizes non-parametric outputs in a parametric way. The results are thus non decreasing in the whole domain and also has enough smoothness. Synthetic and real world experiments suggested my model performs well.

6.1 Data

To verify the efficacy of the developed method, I compared different *calibration* methods on eight binary classification problems. The datasets used in this chapter are: GSE2034, GSE2990, HOSPITAL, ADULT, BANKRUPTCY, HEIGHT_WEIGHT, MNISTALL, PIMATR. The first three data were introduced in Chapter 2. The other data were downloaded from UCI Repository [69]. The data are summarized in Table 6.1. The percentage of positive cases varied from 8% to 67%.

Table 6.1: Datasets for evaluating smooth Isotonic Regression. % POS indicates the percentage of positive cases.

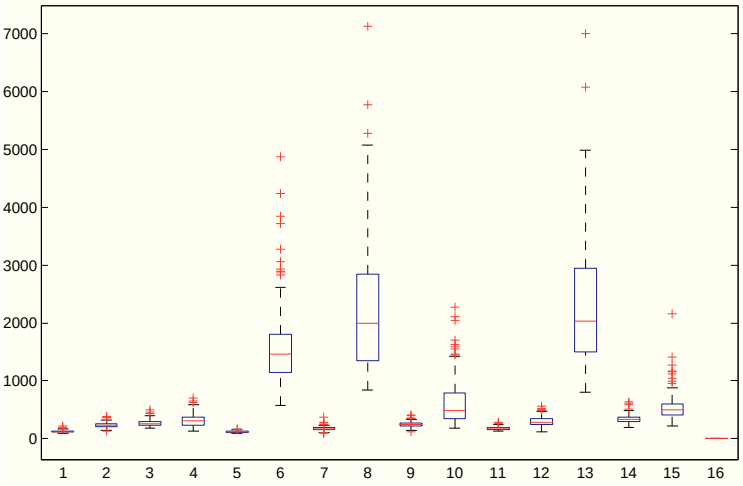
Data	# Attr	Train size	Test size	% POS
GSE2034	15	125	84	54
GSE2990	15	54	36	67
ADULT	14	4,000	41,222	25
BANKRUPTCY	2	40	26	48
HEIGHT_WEIGHT	7	126	84	64
HOSPITAL	22	2,891	1,927	8
MNISTALL	784	42,000	28,000	9.8
PIMATR	8	120	80	33

6.1.1 Breast Cancer Gene Expression data

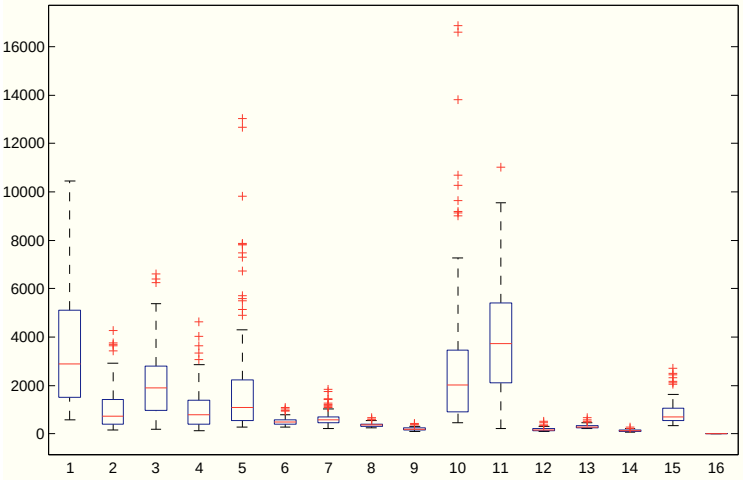
GSE2034 and GSE2990 are gene expression datasets related to breast cancer. The data were obtained from the NCBI Gene Expression Omnibus (GEO). The three individual downloaded datasets were previously studied by Wang et al. (GSE2034) [185] and Sotiriou et al. (GSE2990) [166], respectively.

To make my data comparable to the previous studies, I followed the criteria in [140] to select patient who

did not receive any treatment and had negative lymph node status. Among these pre-selected candidates, only patients with extreme outcomes, either poor outcomes (recurrence or metastasis within five years) or very good outcomes (neither recurrence nor metastasis within eight years) were selected. The number of samples after filtering are: 209 for GSE2034 (114 good/95 poor), and 90 for GSE2990 (60 good/30 poor).Both data have a feature size of 247,965, which corresponds to the gene expression results obtained from micro-array experiments. They were preprocessed to keep only the top 15 features ranked using t-test (see [140] for details). I showed boxplots and matrix plots for both data in Figure 6.1 and Figure 6.2, respectively.

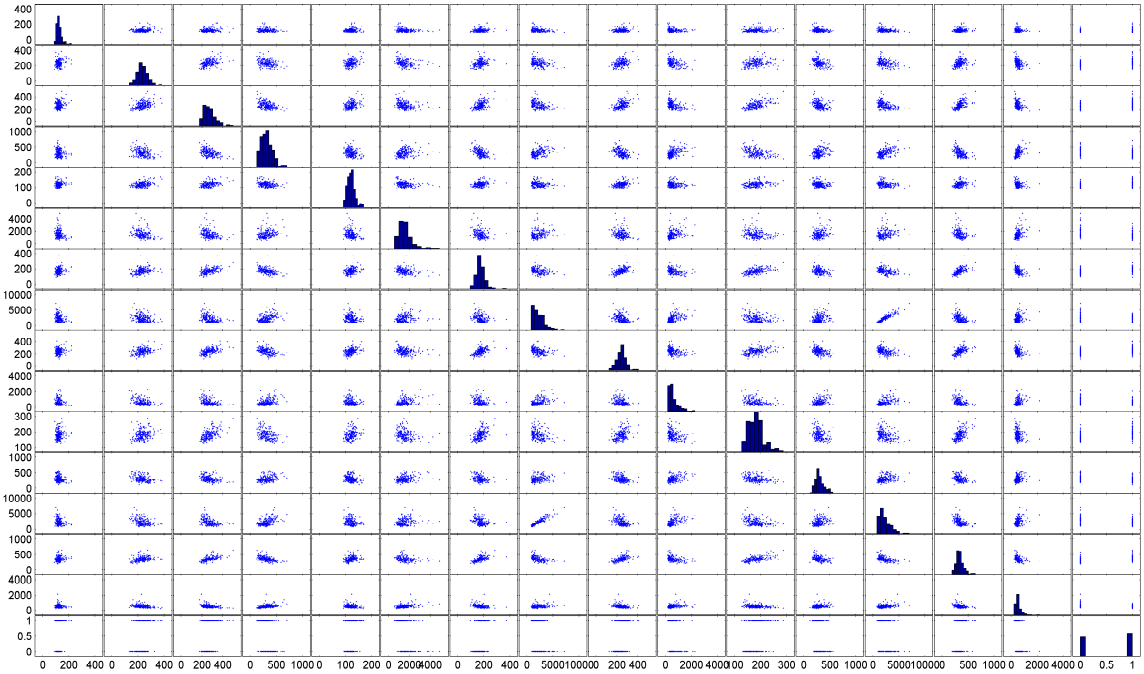


(a) GSE_2034

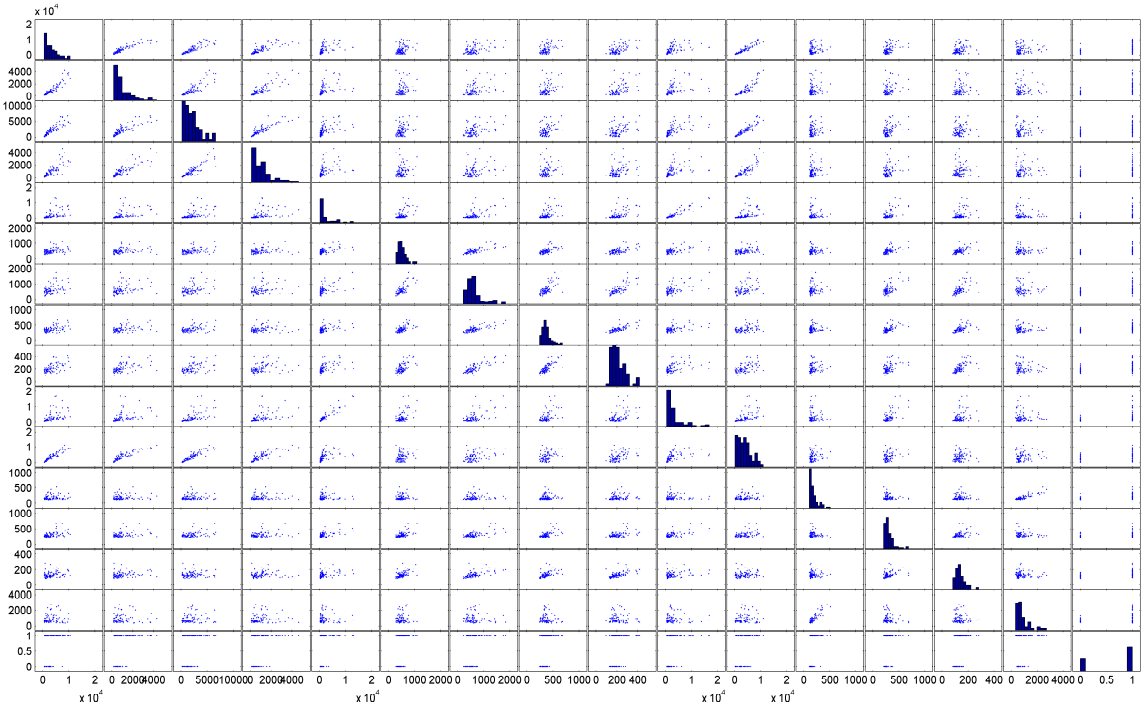


(b) GSE_2990

Figure 6.1: Boxplots of Breast Cancer Gene Expression Data. Each column corresponds to one feature vector and the last column indicates the outcome variable.



(a) GSE_2034



(b) GSE_2990

Figure 6.2: Matrix plots of Breast Cancer Gene Expression Data. Each subfigure corresponds to a matrix plot of one dataset.

6.1.2 Hospital data

The HOSPITAL data set consists of microbiology cultures and other variables related to hospital discharge errors [59]. For patient demographic data, this data contains age, gender, race, and insurance. Related to the hospital encounter, the dataset contains the visit type (admission, emergency room, procedure or outpatient) and admitting service, if applicable. Related to the microbiology result, the dataset contains the specimen type (blood, urine, sputum and cerebral spinal fluid), the hospital day number that the specimen was collected, whether the result was pending at the time of discharge from the hospital, whether the specimen was collected on a weekend, whether the preliminary results (for blood cultures) were reported on a weekend, and whether the final results were reported on a weekend. In addition to the data pulled directly from the hospital computer system, this dataset contains an additional outcome variable that indicates whether the case represents a potential post-discharge follow-up error using experts' knowledge. This variable is true if the following three criteria are met: 1) the result is considered clinically relevant; 2) the results return after the patient is discharged from the hospital; and 3) there is no antibiotic on the discharge medication list to which the organism is sensitive based on the microbiology results. The features are thus comprised of eight categorical variables and two numerical variables. The target is a Boolean variable (Pot_error) indicating the potential error.

The following table summarizes features and outcome variable with their descriptional statistic, i.e., min, 1st Qu., median, 3rd Qu. ,max. The clinical meaning for each column was explained in Chapter 2.

Table 6.2: Descriptional statistic for hospital discharge error data.

specimen	speciman days	collect week	final week	visit type	svc
0: 233	1 :1245	0:3755	0:3583	1:4818	1: 665
1:2564	0 :1030	1:1063	1:1235		2:1287
2:1467	2 : 682				3:1217
3: 554	3 : 391				4:1608
	4 : 327				5: 41
	5 : 227				
	(Other): 916				
age	female	race	insurance	pot error	
Min. : 0.00	0:2252	0:3360	0:1996	0:4449	
1st Qu.:43.28	1:2566	1: 577	1: 554	1: 369	
Median :57.76		2: 110	2:2152		
Mean :56.51		3: 405	3: 116		
3rd Qu.:71.24		4: 55			
Max. :99.71		5: 311			

There are 369 clinically important but highly suspicious observations out of 4819 returned post-discharge observations, which makes the data highly unbalanced and a challenge to *calibrate*.

I also drew the XY plot for various pairs of co-variates to show their co-occurrence patterns, as indicated by Figure 6.3.

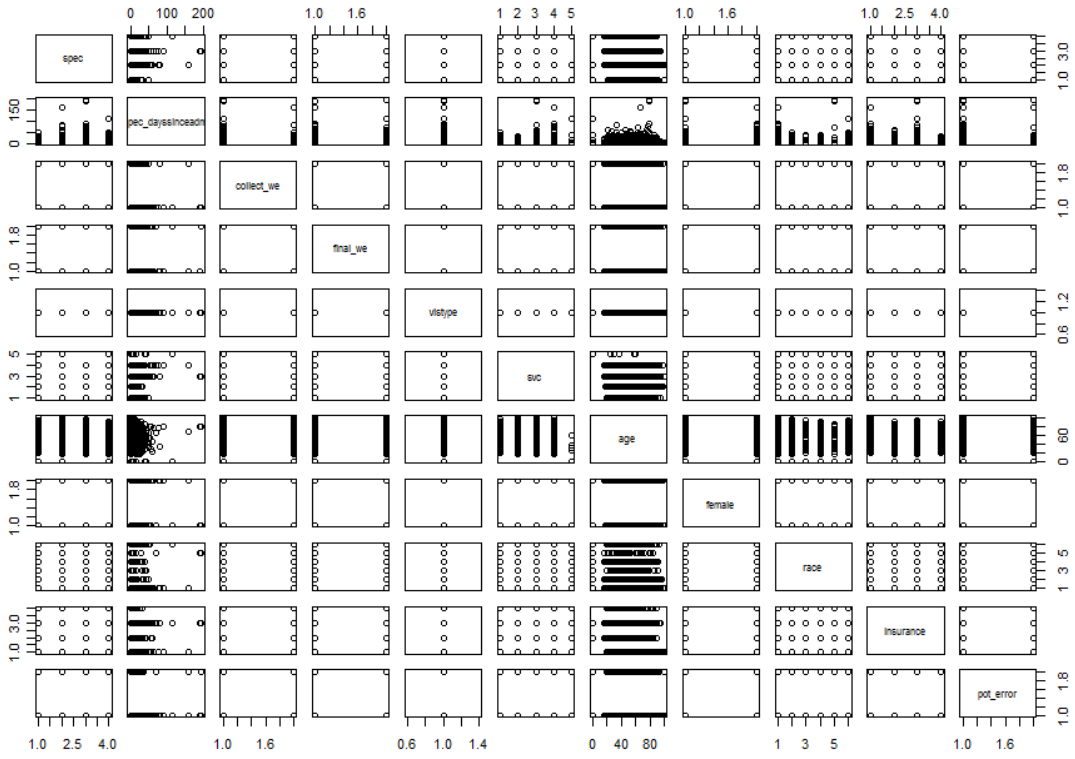


Figure 6.3: XY plots for hospital discharge error data.

6.1.3 Height and Weight data

The HEIGHT_WEIGHT data were downloaded from UCI Repository [69]. The data contain physiological measurements of different genders. The subjects are 213 students of an academic university. Seventy-three students are female and 140 are male. The data have the following features: height, weight, GPA, left arm length, right arm length, left foot size, and right foot size. These co-variates are self-explanatory by their names.

Table 6.3: Descriptional statistic for the HEIGHT_WEIGHT data.

Sex	Height	Weight	GPA
0: 73	Min. :55.00	Min. : 95.0	Min. :1.240
1: 140	1st Qu.:64.00	1st Qu.:125.0	1st Qu.:2.670
	Median :67.00	Median :140.0	Median :3.000
	Mean :67.31	Mean :145.5	Mean :3.004
	3rd Qu.:70.50	3rd Qu.:160.0	3rd Qu.:3.400
	Max. :79.00	Max. :280.0	Max. :3.910
LArm	RArm	LFoot	RFoot
Min. :20.50	Min. :20.50	Min. :19.50	Min. :20.00
1st Qu.:24.00	1st Qu.:24.00	1st Qu.:23.40	1st Qu.:23.00
Median :25.00	Median :25.00	Median :24.70	Median :25.00
Mean :25.17	Mean :25.31	Mean :25.16	Mean :25.20
3rd Qu.:26.50	3rd Qu.:27.00	3rd Qu.:27.00	3rd Qu.:27.00
Max. :31.00	Max. :31.00	Max. :32.00	Max. :32.00

To see these statistic variables visually, I box plotted 7 co-variables along with the outcome variable in Figure 6.4.

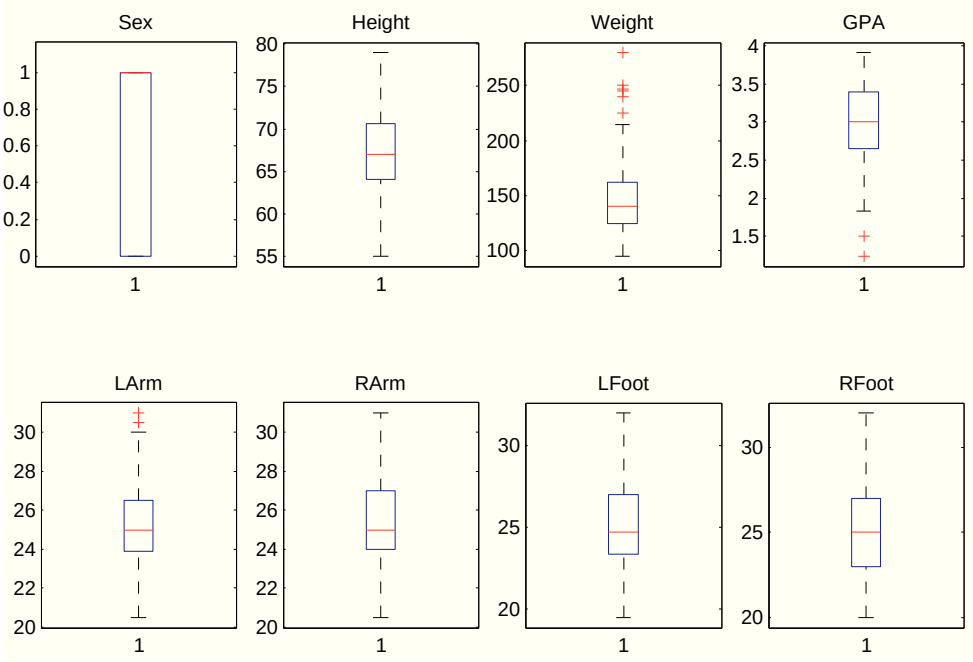


Figure 6.4: Boxplots for HEIGHT_WEIGHT data.

I also included XY plots for various co-variables for this data. The following figure contains 56 subplots, each indicates the co-occurrence of two variables. The diagonal cell corresponds to names of these variables.

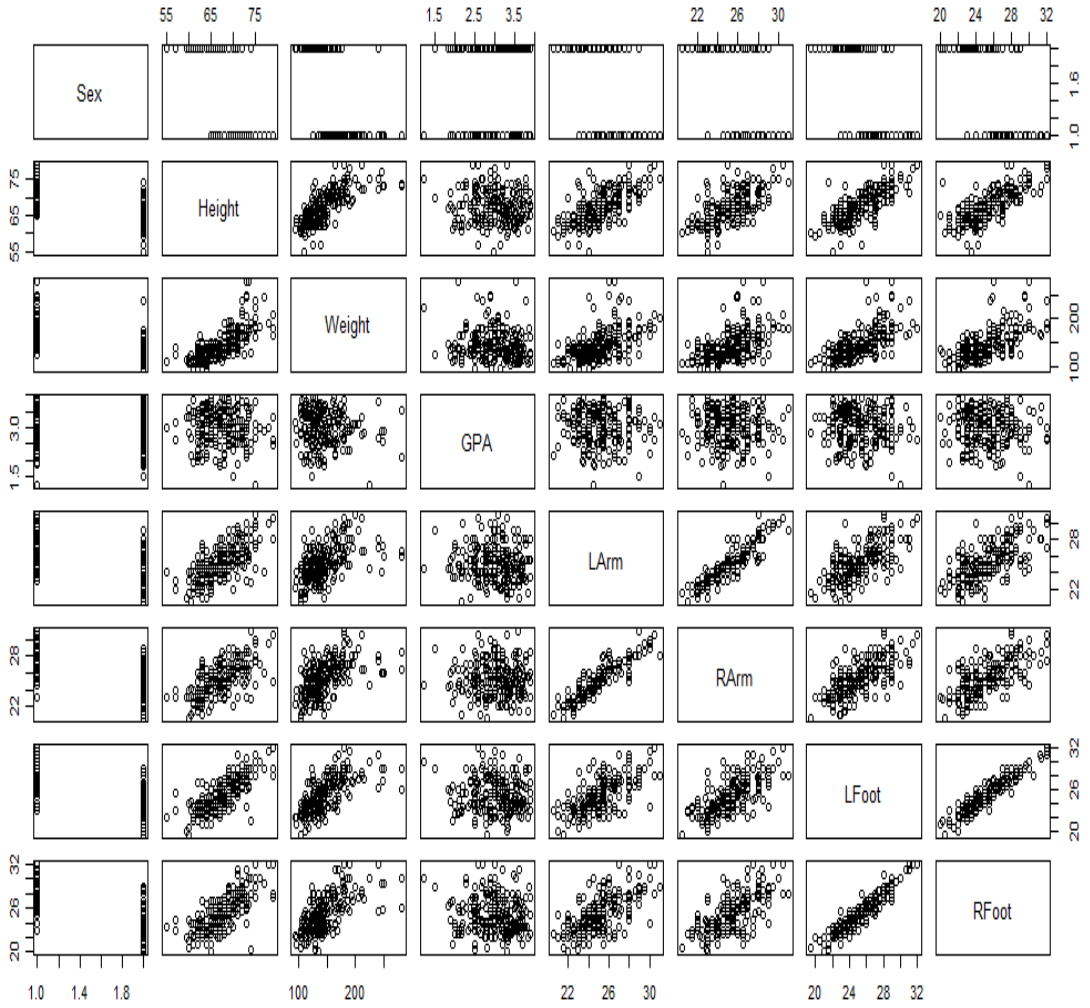


Figure 6.5: XY plots for the HEIGHT_WEIGHT data set.

6.1.4 Adult Census data

The extraction of the ADULT_CENSUS data was conducted by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted and the prediction task is to determine whether a person earns over 50K a year [101].

This data contains 14 co-variables and a binary outcome variable "income." The following table summarizes the basis statistics of these co-variables and the outcome variable.

Table 6.4: Descriptional statistic for the ADULT_CENSUS data.

age	workclass	fnlwgt	education	education num
Min. :17.00	Private :22696	Min. : 12285	HS-grad :10501	Min. : 1.00
1st Qu.:28.00	Self-emp-not-inc: 2541	1st Qu.: 117827	Some-college: 7291	1st Qu.: 9.00
Median :37.00	Local-gov : 2093	Median : 178356	Bachelors : 5355	Median :10.00
Mean :38.58	? : 1836	Mean : 189778	Masters : 1723	Mean :10.08
3rd Qu.:48.00	State-gov : 1298	3rd Qu.: 237051	Assoc-voc : 1382	3rd Qu.:12.00
Max. :90.00	Self-emp-inc : 1116	Max. :1484705	11th : 1175	Max. :16.00
NA	(Other) : 981		(Other) : 5134	
marital.status	occupation	relationship	race	sex
Divorced : 4443	Prof-specialty :4140	Husband :13193	Amer-Indian-Eskimo: 311	Female:10771
Married-AF : 23	Craft-repair :4099	Not-in-family : 8305	Asian-Pac-Islander: 1039	Male :21790
Married-civ :14976	Exec-managerial:4066	Other-relative: 981	Black : 3124	
Married-absent: 418	Adm-clerical :3770	Own-child : 5068	Other : 271	
Never-married :10683	Sales :3650	Unmarried : 3446	White :27816	
Separated : 1025	Other-service :3295	Wife : 1568		
Widowed : 993	(Other) :9541			
capital.gain	capital.loss	hours.per.week	native.country	income
Min. : 0	Min. : 0.0	Min. : 1.00	United-States:29170	<=50K:24720
1st Qu.: 0	1st Qu.: 0.0	1st Qu.:40.00	Mexico : 643	>50K : 7841
Median : 0	Median : 0.0	Median :40.00	? : 583	
Mean : 1078	Mean : 87.3	Mean :40.44	Philippines : 198	
3rd Qu.: 0	3rd Qu.: 0.0	3rd Qu.:45.00	Germany : 137	
Max. : 99999	Max. :4356.0	Max. :99.00	Canada : 121	
			(Other) : 1709	

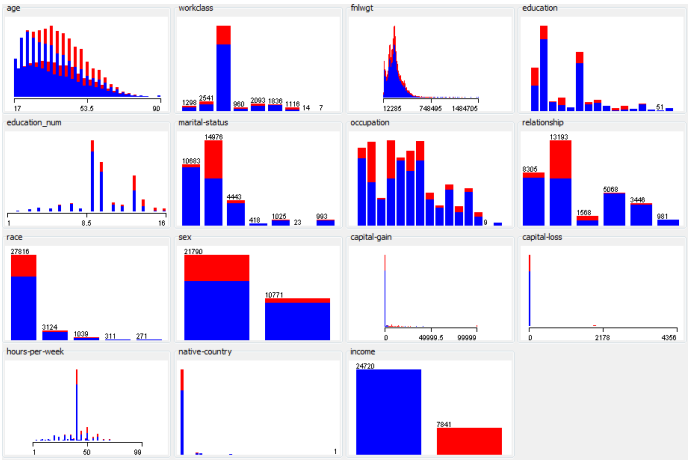


Figure 6.6: The co-variates are grouped by the outcome variable, Income.

To visually demonstrate the distribution of each variable, I plotted histograms for each of the 14 co-

variables grouped by the outcome variable in Figure 6.6. I also included XY plots for these co-variables to show their co-occurrence patterns. The following figure contains 110 subplots and each indicates the co-occurrence of two variables. The diagonal cell corresponds to variable names.

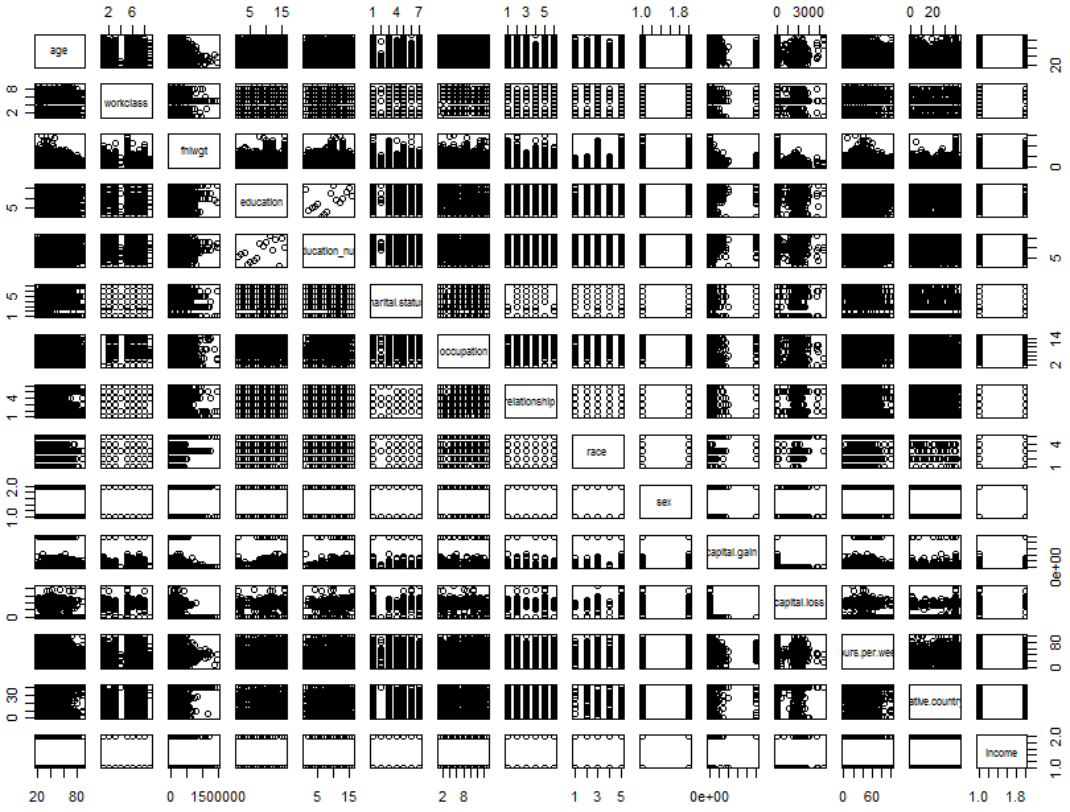


Figure 6.7: XY plots for the ADULT_CENSUS data.

6.1.5 Bankruptcy data

Bankruptcy data were also downloaded from UCI Repository [69]. The data contain two features: Return and EBIT (earnings before interest and taxes). The outcome variable "Bankruptcy" is binary. There are 66 samples in this data, where 33 samples correspond to observed bankruptcy and the others do not. The following table summarizes the co-variates and outcome variables with their description statistics, i.e., min, 1st Qu., median, 3rd Qu. ,max.

Table 6.5: Descriptional statistic for the BANKRUPTCY data.

Return	EBIT	Bankruptcy
Min. :-308.90	Min. :-280.000	0:33
1st Qu.: -39.05	1st Qu.: -17.675	1:33
Median : 7.85	Median : 4.100	
Mean : -13.63	Mean : -8.226	
3rd Qu.: 35.75	3rd Qu.: 14.400	
Max. : 68.60	Max. : 34.100	

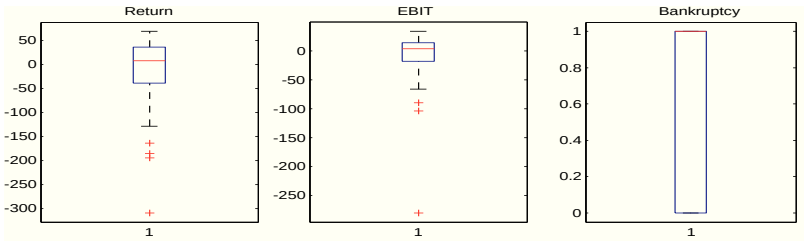


Figure 6.8: Boxplots for the BANKRUPTCY data set.

To visually demonstrate the distribution of each variable, I box plotted these variables in Figure 6.6. In addition, I included XY plots for co-variables and outcome variables to illustrate the co-occurrence patterns. The following figure contains 6 subplots, each indicates the co-occurrence of two variables. The diagonal cell corresponds to variable names.

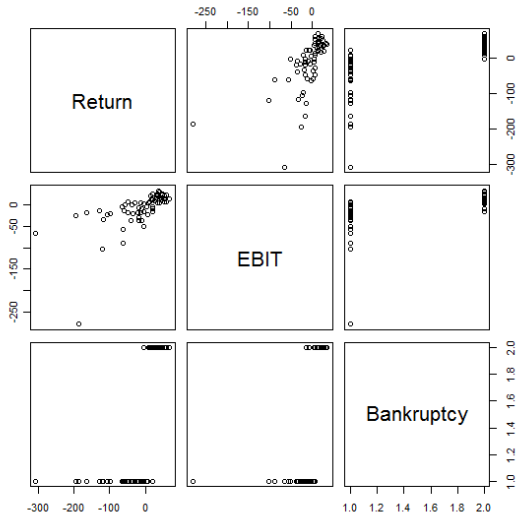


Figure 6.9: XY plot for the BANKRUPTCY data set.

6.1.6 Pimatr Indian Women data

The Pimatr Indian Women data were also downloaded from UCI Repository [69]. In this data, a population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data contains the 532 complete records after dropping the (mainly missing) data on serum insulin.

The data have the following co-variables: 'npreg' number of pregnancies; 'glu' plasma glucose concentration in an oral glucose tolerance test; 'bp' diastolic blood pressure (mm Hg); 'skin' triceps skin fold thickness (mm); 'bmi' body mass index (weight in kg/(height in m)^2); 'ped' diabetes pedigree function; 'age' age in years; finally, the outcome variable 'type' 1 indicates 'Yes' and 0 indicates 'No', for diabetic according to WHO criteria.

The following table summarizes the co-variates and outcome variables with their description statistics, i.e., min, 1st Qu., median, 3rd Qu. ,max for this data.

Table 6.6: Descriptional statistic for the PIMATR data.

obs	npreg	glu	bp	
Min. : 1.00	Min. : 0.00	Min. : 56.0	Min. : 38.00	
1st Qu.: 50.75	1st Qu.: 1.00	1st Qu.:100.0	1st Qu.: 64.00	
Median :100.50	Median : 2.00	Median :120.5	Median : 70.00	
Mean :100.50	Mean : 3.57	Mean :124.0	Mean : 71.26	
3rd Qu.:150.25	3rd Qu.: 6.00	3rd Qu.:144.0	3rd Qu.: 78.00	
Max. :200.00	Max. :14.00	Max. :199.0	Max. :110.00	
skin	bmi	pedigree	age	type
Min. : 7.00	Min. :18.20	Min. :0.0900	Min. :21.00	0: 132
1st Qu.:20.75	1st Qu.:27.57	1st Qu.:0.2500	1st Qu.:23.00	1: 68
Median :29.00	Median :32.80	Median :0.3700	Median :28.00	
Mean :29.21	Mean :32.31	Mean :0.4613	Mean :32.11	
3rd Qu.:36.00	3rd Qu.:36.50	3rd Qu.:0.6200	3rd Qu.:39.25	
Max. :99.00	Max. :47.90	Max. :2.2900	Max. :63.00	

To visually demonstrate the distribution of each variable, I box plotted these variables in Figure 6.10.

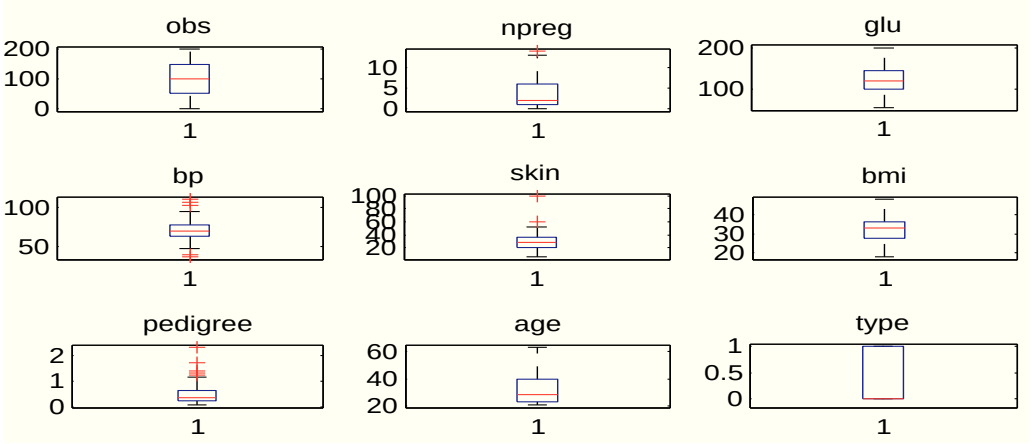


Figure 6.10: Boxplots for the Pimatr Indian Women data.

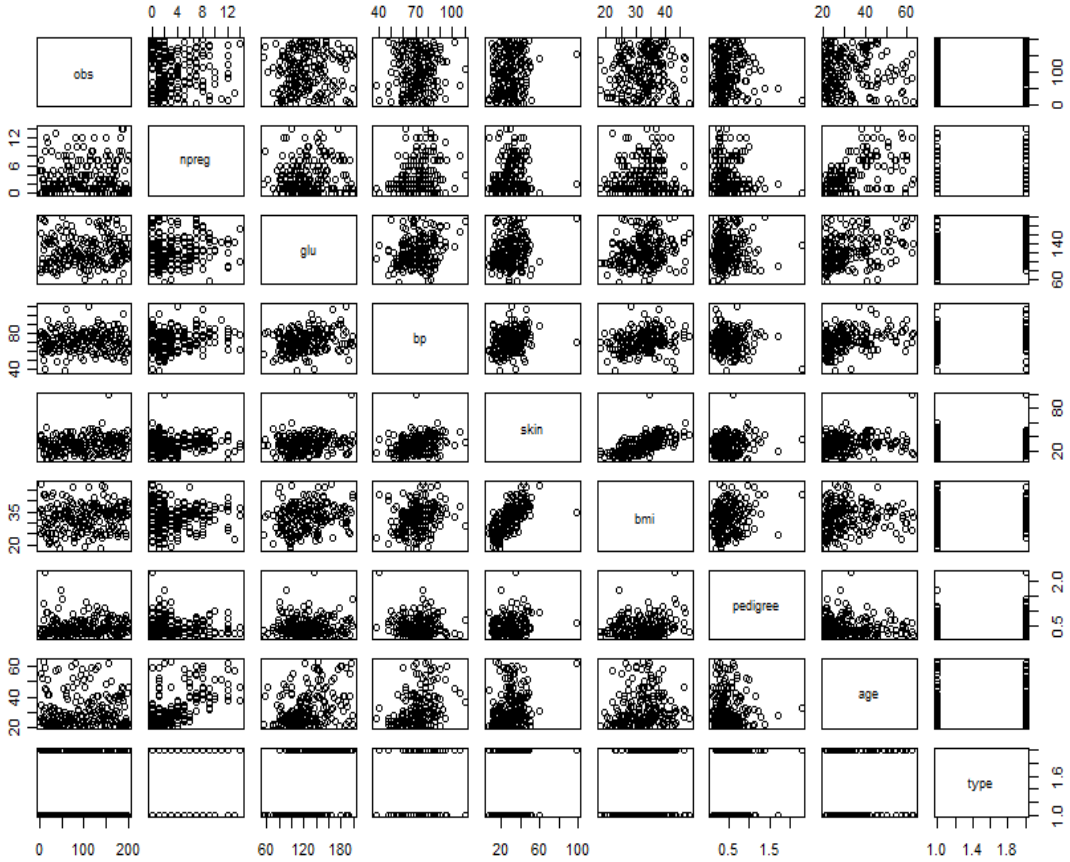


Figure 6.11: XY plots for the Pimatr Indian Women data.

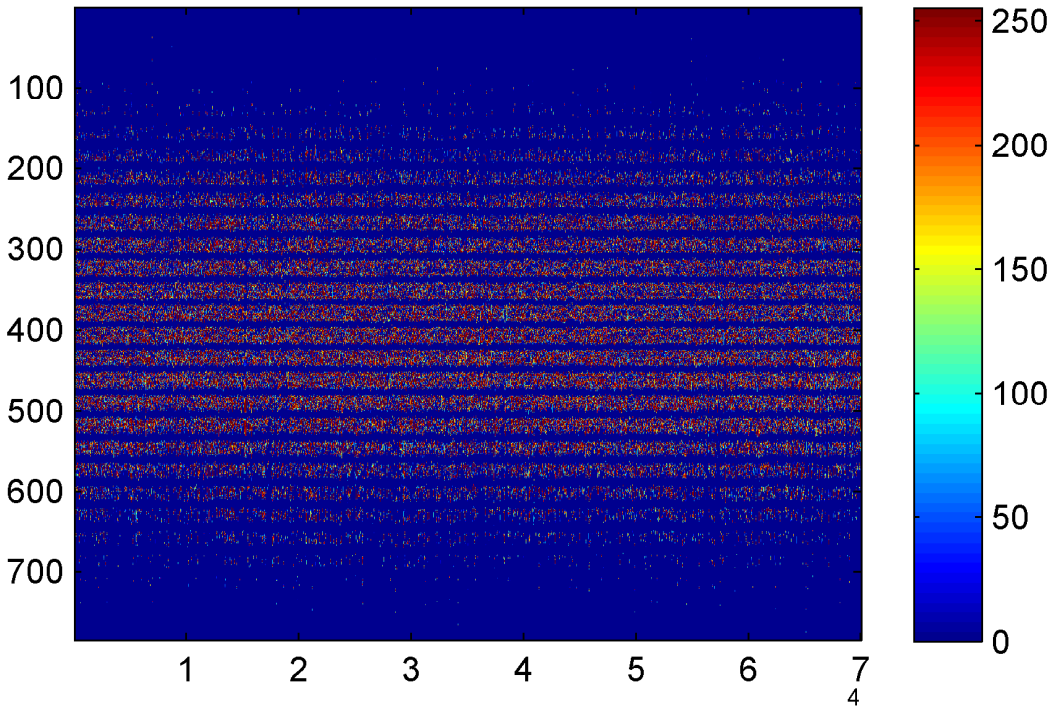


Figure 6.12: Image plot for the MNIST data. The size of the matrix is 784x70,000. Each row corresponds to one pixel and its values of all 70,000 samples and each column represents an extrapolated 28x28 image.

I also included XY plots for co-variables and the outcome variable to illustrate the co-occurrence patterns. The figure above contains 72 subplots, each indicates the co-occurrence of two variables. The diagonal cell corresponds to variable names.

6.1.7 MNISTALL data

MNISTALL data contain 70,000 handwritten numbers . Each character is represented by a black and white (bi-level) image, whose label is a number from '0-9'. The images were centered in a 28×28 image, and thus have a feature size of 784. I convert the multi-categorical prediction problem into a binary one by labeling all digits '0' as positive and others as negative. This conversion yields a very unbalanced set, which is a challenge to *calibration* models.

Due to the size of the data, I cannot display the boxplots or histograms for each co-variable as well as the matrix plot between them. In addition, the raw pixel feature does not has any semantic meaning.

Instead, I plotted an image of this gigantic 784×70,000 matrix in Figure 6.12. More detailed explanation and motivation for this data can be found at `http://yann.lecun.com/exdb/mnist/`.

6.2 Related Work

Two different approaches directly motivated my research in this chapter. One of them is a parametric method and the other is a non-parametric approach.

The parametric method is the Platt scaling developed by Platt [147], which suggested using a sigmoid function $P(c_i = \{+\}|p_i) + \frac{1}{1+exp(Ap_i+B)}$ to transform outputs of a "*uncalibrated*" model to probabilistic outputs. Here, c_i corresponds to the class label, p_i is the prediction probability and A, B are the model parameters. The parameters of the sigmoid function can be efficiently estimated by maximum likelihood criteria. The method was first developed to transform the score of Support Vector Machines to probabilistic outputs for decision support systems. But this approach is widely generalizable to any classifier that generates a score value to co-variate patterns. The advantage of using a model based approach to *recalibrate* outputs of non-probabilistic or uncalibrated models are expected to be smooth. However, this is a double-bladed sword that limits the model's flexibility of modeling more complex patterns. The Platt scaling approach cannot improve *calibration* when the inputs are non-monotonic or the inputs concentrate on narrow regions. In the first case, the method is not suitable because sigmoid function are strictly monotonic and continuous. In the later case, Platt scaling extrapolates outside the input dense regions with low confidence. Figure 6.13 illustrates the Platt Scaling Mapping Function together with "uncalibrated" scores of the Naive Bayes classifier. This is essentially a Sigmoid function with optimized parameters learned to fit the dichotomous class labels of training data.

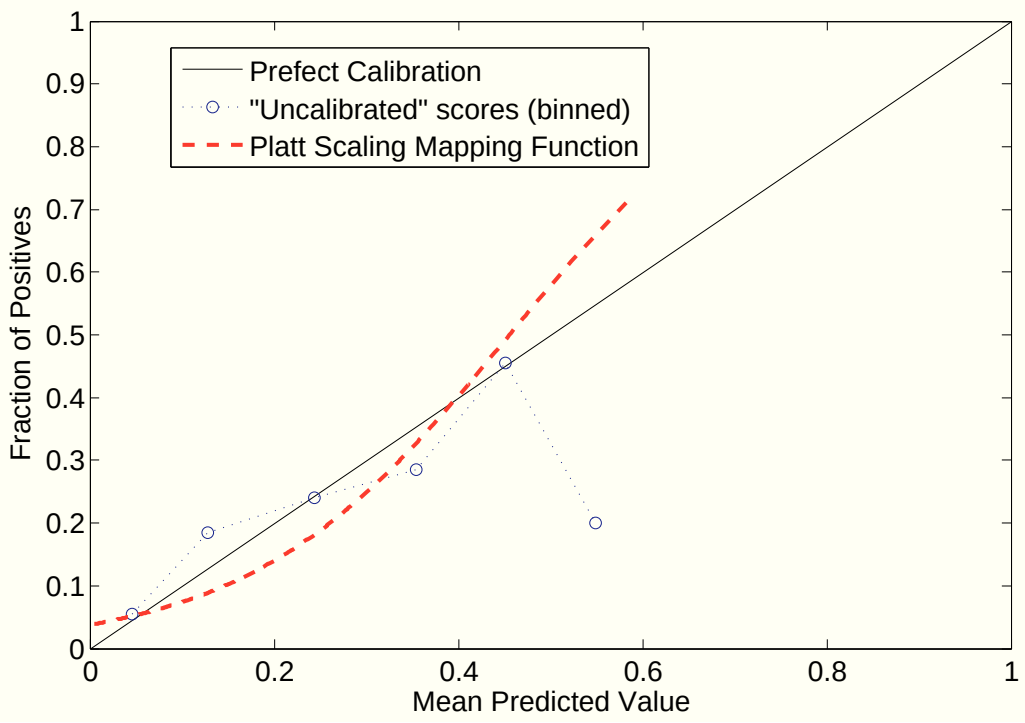


Figure 6.13: Illustration of Platt Scaling Mapping Function. The function is plotted along with "uncalibrated" scores of the Naive Bayes classifier.

The non-parametric method Isotonic Regression is not new to biomedical informatics [9, 130, 135, 203]. However, it was first used by Zadrozny et al. for *calibration* purpose. Their paper [206] applied Isotonic Regression to recalibrate scores of Naive Bayes and Support Vector Machine, which demonstrated good performance. The objective of Isotonic Regression is a least squares minimization to observed dichotomous class labels. It can be formulated as follows: $\min \sum_i w_i(p_i - y_i)^k$ subject to $p_i \geq p_j, \forall(i,j)$. Note the k is the order of the norm, w_i corresponds to the weight, y_i is the class label and p_i is the probability after *calibration*. The users assign the weights if they have prior information; otherwise, $w_i = 1, \forall i$. The formula is subject to a set of monotonicity constraints imposing simple or partial order over the variables. When $k = 2$, there is an efficient pair-adjacent violators (PAV) algorithm to solve the problem [197]. There are no regularization terms in this non-parametric recalibration approach, and its results tend to overfit the training data [186].

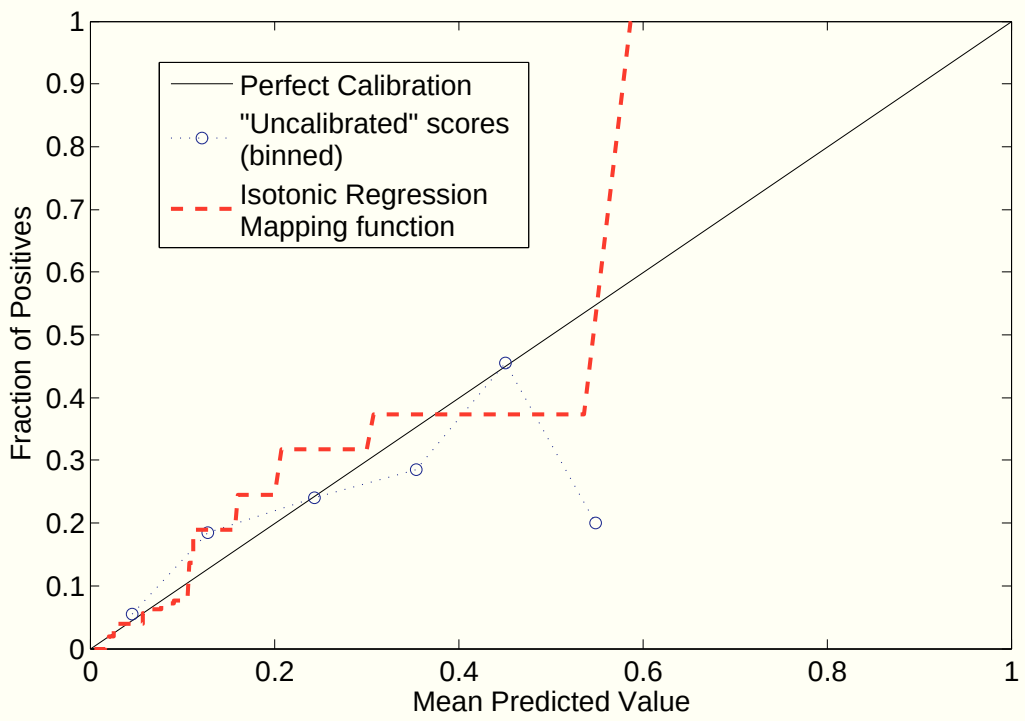


Figure 6.14: Illustration of Isotonic Regression Mapping Function. The mapping function is plotted along with "uncalibrated" scores of the Naive Bayes classifier.

Figure 6.14 illustrates the Isotonic Regression Mapping Function together with "uncalibrated" scores of the Naive Bayes classifier. The function minimized the least squares error between probabilistic outputs and observed dichotomous class labels, which resulted in a zigzag shape of the mapping function. The approach is non-parametric, that is data-driven, and provides more flexibility in modeling complex input pattens than Platt Scaling. However, there is no smoothing term in Isotonic Regression that could helps to improve the generalizability of method. As opposed to Platt Scaling, Isotonic Regression has much higher flexibility but limited generalizability as the method often overfits the training data.

6.3 Smooth Isotonic Regression

As neither the parametric model nor the non-parametric model provided satisfactory solution to the *calibration* problem, I investigated the possibilities of combining merits of both models to develop a smooth non-parametric method to avoid the over-fitting problem.

Coincidentally, Wang and Li conducted a similar but different research [186]. Their approach, called the

monotone smoothing spline estimator, aims to find a non-decreasing function $t()$ that minimizes:

$$\sum_i (y_i - t(p_i))^2 = \lambda \int_a^b t^{(k)}(z)\, dz, \tag{6.1}$$

where λ is a fixed smoothing parameter. The first term measures the fit to the data, and the second term controls the smoothness of the fitted function, where $k = 1$ corresponds to a piece-wise linear estimator, and $k = 2$ suggests a smoother monotone estimator. Unfortunately, minimizing Equation 6.1 when $k = 2$ over all smooth monotone functions is a difficult nonlinear optimization problem. The authors of [186] proposed a second-order cone programming (SOCP) approximation algorithm, which has an empirical loss of approximately 30%.

I observed that Isotonic Regression is a non-parametric method that collapses raw predictions into larger deciles; thus it uses only a few representative values. With only a finite number of these "representative" values, a parametric function can be used to smooth the model predictions. To maintain the AUC, I must ensure that such parametric function $G()$ is monotonically increasing, so that $t^*() = G(I())$ is also monotonically increasing, where $I()$ is the isotonic regression function.

I develop a novel approximation to the optimal smooth function $t^*()$ that minimizes Equation 6.1, in two steps. First, I apply isotonic regression to obtain a monotone non-parametric function t that minimizes $\sum_i (y_i - t(p_i))^2$. Second, I constructed a monotone smoothing spline to interpolate knots sampled from t to obtain a smoothed approximation (Algorithm 4).

Algorithm 4 Smooth Isotonic Regression.

Input: Original prediction probability $P = p_1, \ldots, p_n$, True and class labels $Y = y_1, \ldots, y_n$.
Output: Smoothed isotonic regression function H .
Parameters: α: number of samples at each iteration, k: dimension of hidden topics, N: an scale factor.

1: Obtain $I = argmax_i \sum_i (y_i - I(p_i))^2$, subject to $I(p_i) \leq I(p_{i+1}) \forall i$ (Isotonic Regression).
2: Sample S knots from $I(p), p \in (0, 1)$, one knot at the median of each step in $I()$. Denote these samples P^+, their corresponding class labels Y^+.
3: Construct a Piecewise Cubic Hermite Interpolating Polynomial with P^+, Y^+ to obtain a monotone smoothing spline H.

I leverage a special parametric model named Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) [70], which interpolates between the S knots p_i monotonically using cubic splines.

$$H(p) = \begin{cases} H_1(p) & if\ p_1 \leq p \leq P_2 \\ H_2(p) & if\ p_2 \leq p \leq P_3 \\ \dots & \dots \\ H_{N-1}(p) & if\ p_{N-1} \leq p \leq P_N \end{cases}, \tag{6.2}$$

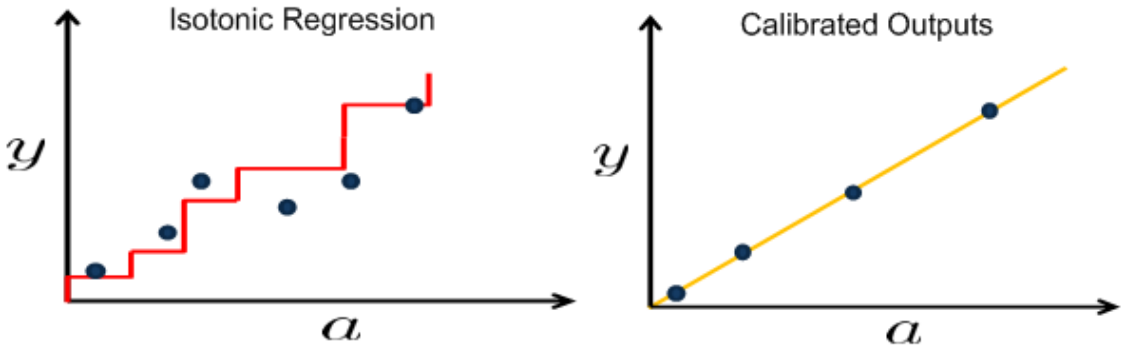
$$H_i(p) = a_i + b_i(p - p_i) + c_i(p - p_i)^2 + d_i(p - p_i)^2(p - p_{i+1}). \tag{6.3}$$

The PCHIP function above interpolates the values at intermediate points, such that:

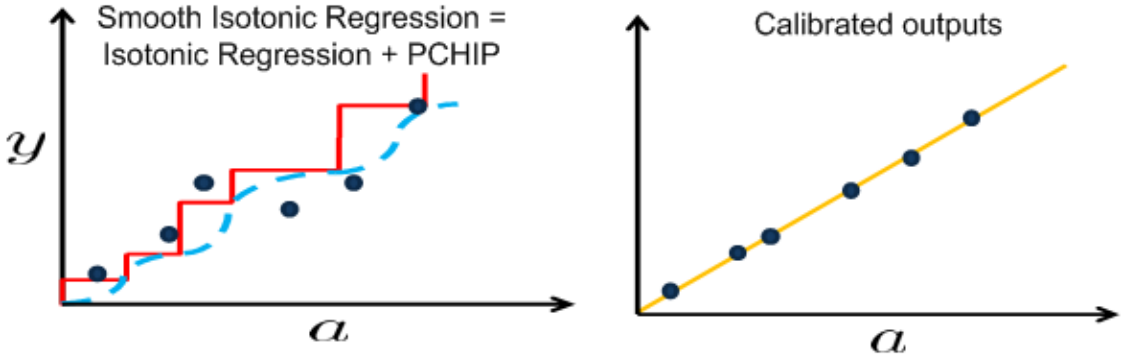
1. On each subinterval $p_i \leq p \leq p_{i+1}$, Cubic Spline interpolate the values in between these endpoints;
2. $H(p)$ interpolates y, i.e. $H(p_i) = y_i$, and the first derivative H' is continuous;
3. The slopes at p_j are chosen to respect monotonicity, which means that, on intervals where the data are monotonic, so is $H(p)$; at points where the data have a local extreme, so does $H(p)$.

Interpretation: Smooth Isotonic Regression interpolates Isotonic Regression (IR) monotonically so that outputs preserve the ranking induced by IR but is spread more smoothly.

Figure 6.15 gives an example of using both Isotonic Regression and Smooth Isotonic Regression to *calibrate* the same probabilistic model. The figures indicates that Isotonic Regression tends to overfit as it collapse multiple inputs to the same output value while Smooth Isotonic Regression does not have such problem.



(a) Isotonic Regression maps multiple inputs to the same value, and the model tends to overfit.



(b) Smooth Isotonic Regression maps every input get mapped to a unique output. The smoothness reduces overfitting.

Figure 6.15: Comparing Isotonic Regression with Smooth Isotonic Regression.

6.4 Experiments

I evaluated the method's performance using both synthetic and real world datasets. In both experiments, I used AUC [83] and HL-C test [90] as my evaluation metrics.

6.4.1 Synthetic Data

Random samples ($n = 1000$) were drawn from two Gaussian distributions with varied difference in means but fixed variances. The distance between μ_1 and μ_2 (shift) is set to increase from 0.5 to 2.0 at an interval of 0.5; $\Sigma_1 = 2.0, \Sigma_2 = 1.0$, respectively. The samples have features X and their class memberships are used as the ground-truth labels. The Logistic Regression (LR) model was fit on Y over X, where X is the sampled value Y is its class memberships, Y , the ground-truth label. I then compare the *calibration* of three

different recalibration models (sigmoid fitting, isotonic regression, and smooth isotonic regression) along with the *calibration* of the raw prediction probabilities of the LR model, as indicated in Figure 6.16. I use 80% of the data for training, and test on the remaining 20% for all models.

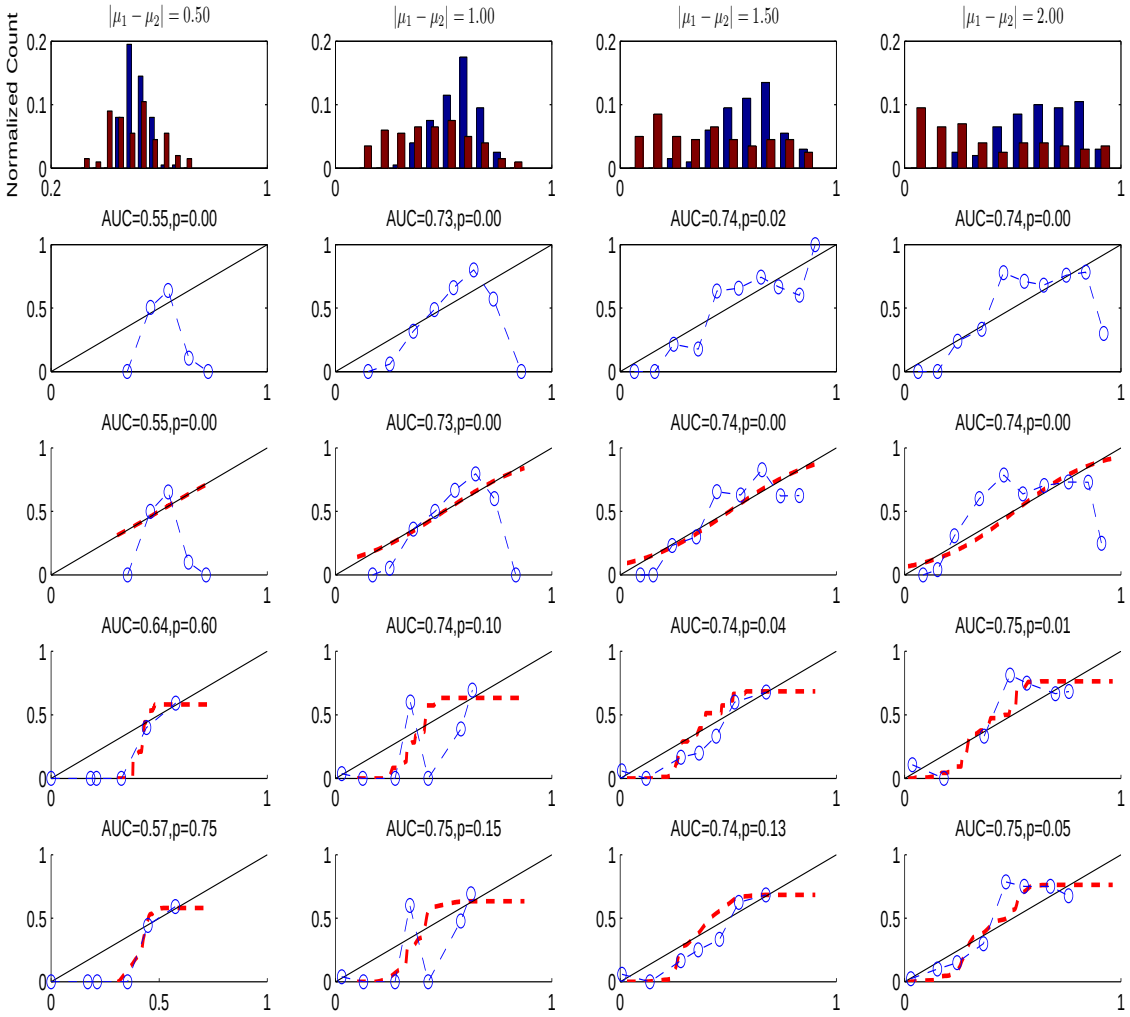


Figure 6.16: Comparison of different *calibration* methods. From rows one to five are 1) histograms of the original predicted probabilities (blue bars for class "1", red bars for class "0"), and reliability diagrams for 2) the original probabilities, 3) sigmoid fitting, 4) isotonic regression, and 5) smooth isotonic regression.

The blue circles are the original predicted probability on these reliability diagrams. The red dotted line corresponds to the transformations. While sigmoid fitting does not improve the *calibration* in all four cases, both isotonic regression and smooth isotonic regression follow the data pattern closely; the smooth version has less oscillation and shows $P > 0.05$ on the H-L test, indicating that the models are reasonably

well calibrated. As I observe in Figure 6.16, isotonic regression tends to over-fit, while smooth isotonic regression provides a continuous transformation and the highest p-values for the HL-test in most cases.

6.4.2 Real World Experiment

I quantitatively examined the *calibration* and *discrimination* using various methods. For each model, I built models on 60% random samples and tested on the remaining 40%, with the exception of the ADULT data-set, where I followed the split used in [136].

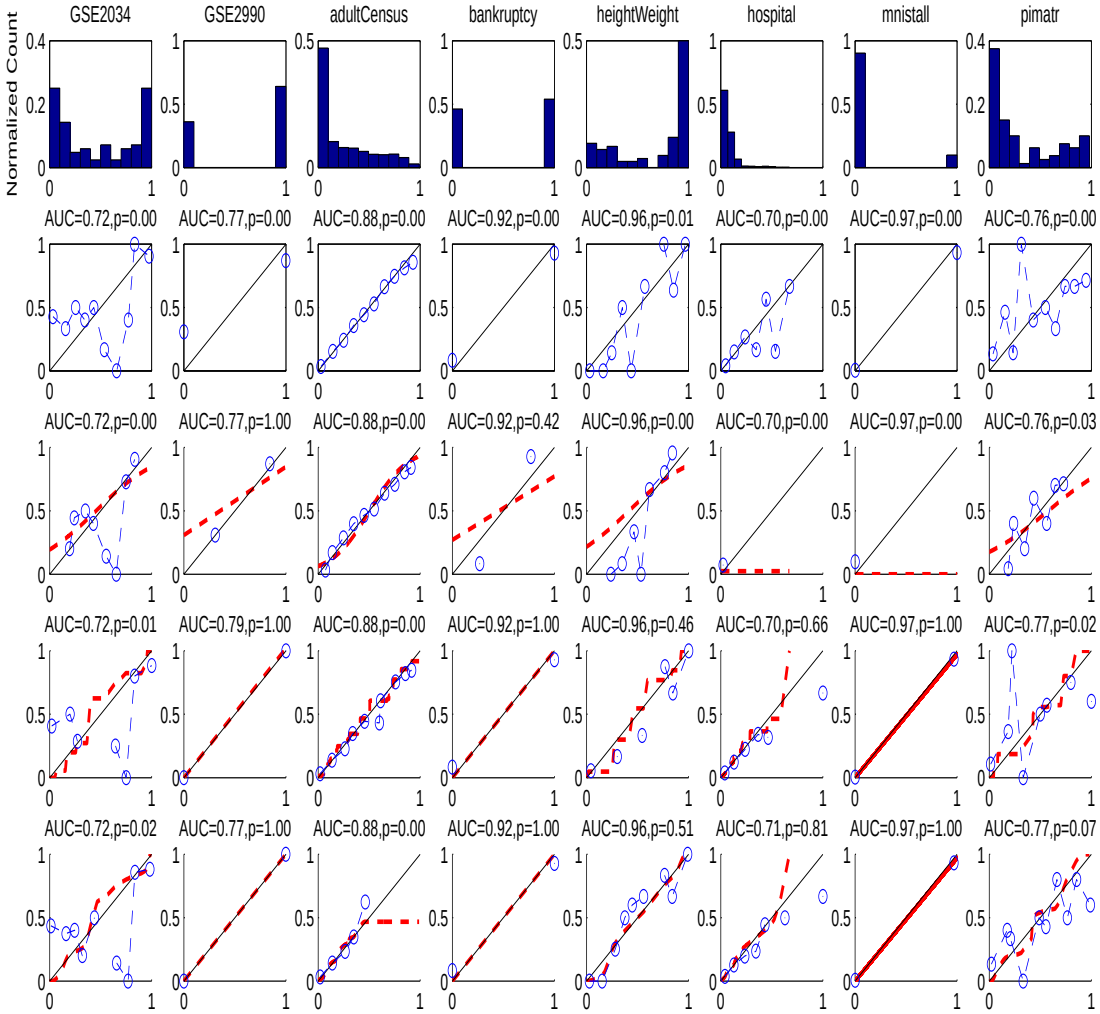


Figure 6.17: Comparison of different *calibration* methods on real world data. From rows one to five are (1) histograms of the original predicted values (no color *discrimination* for classes is used this time), and reliability diagrams for the (2) original predictions, (3) sigmoid fitting, (4) isotonic regression, and (5) smooth isotonic regression.

Figure 6.17 shows histograms of the predicted values (top row) and reliability diagrams for logistic regression, sigmoid fitting, isotonic regression, and smooth isotonic regression on all eight test sets used neither for training nor *calibration*. None of the *calibration* methods ever decreases the AUC, since the monotonic transformation functions preserve the orderings. Isotonic regression sometimes shows an increase in AUC because it introduces more ties into the ranking.

An interesting observation gathered from the reliability diagrams is that they seldom display a sigmoid shape for these problems, thus discouraging the use of a sigmoid to transform predictions into probabilities (see third row). The reliability diagrams in the fourth row of the figure show results for isotonic regression, which are not smooth and are unrealistically sharp at the corners. The reliability diagrams at the bottom of the figure show the functions fitted with my smooth isotonic regression, which has better performance than sigmoid fitting and less oscillation than isotonic regression. In all cases, smooth isotonic regression gives the highest p-value for the HL-test, suggesting a better fit than the sigmoid approach and less over-fit when compared to isotonic regression.

6.5 Discussion

There is an increasing interest to improve the *calibration* ability of predictive models, especially given their potential in personalized medicine. While *discrimination* is often optimized, *calibration* is sometimes neglected, potentially leading to models that are not adequate for use in practice. However, recent research in machine learning has shown the benefits of calibrating predictive models, especially in support decisions. Unfortunately, existing approaches like Platt Scaling and Isotonic Regression are limited in their capabilities to calibrate outputs of probabilistic models. The former method does not always fit observations, especially when the non-monotonic pattern presents. On the other hand, the latter method tends to over-fit training data and may lead to bad performance on testing.

To provide tools with an integrated *calibration* ability without human intervention, I investigated possibilities of achieving both *discrimination* and *calibration*. I carefully analyzed pros and cons of existing parametric model and non-parametric model. My investigation indicated that these two can be complementary to each other and save themselves from their own limitations. By extending the Isotonic Regression for recalibration to obtain a smoother fit in reliability diagrams, I developed a novel method, Smooth Isotonic Regression (SIR), which combines parametric and non-parametric approaches and utilizes non-

parametric outputs of Isotonic Regression in a parametric way. The method calibrates probabilistic outputs more smoothly than Isotonic Regression and showed better generalization ability than its ancestors (i.e., IR).

However, there is a major limitation of SIR that it is still a monotonic function. This constraint limits SIR's capability of rectifying probabilistic outputs. That is, SIR has to keep partial orderings of original probabilistic outputs, in which case only translation, rotation, and stretching operations are allowed. When there are a large number of observations, there is usually little degree-of-freedom to adjust outputs of a probabilistic model. To overcome this limitation, I considered alternative approaches that are capable of going beyond monotonicity but still optimizing the integrated framework (Equation 5.7). To achieve that, I decided to narrow the space of function so that specific model-based characteristics can be utilized. The next chapter introduces adaptive calibration for logistic regression (AC-LR) as another automated calibration approach.

6.6 Conclusion

I carefully analyzed a state-of-the-art *calibration* approach, Isotonic Regression. Under a monotonic transformation constraint, the objective function of Isotonic Regression is optimal for the *calibration* task. However, a major drawback of this approach is its inability to provide a smooth transformation function for calibrating probabilistic output, and its results tend to overfit the training data.

To alleviate such non-smoothness but preserve other merits of Isotonic Regression, I developed a new method, Smooth Isotonic Regression, that utilizes non-parametric outputs of Isotonic Regression in a parametric way. The method used Piecewise Cubic Hermite Interpolation Polynomials (PCHIP) to interpolate representative outcomes of Isotonic Regression. The PCHIP function is smooth and monotonic, thus providing the smoothness while retaining the shape constraints of Isotonic Regression. The method demonstrated better generalization ability than its ancestors (Isotonic Regression) in simulated and real world biomedical datasets. In the synthetic data experiments, SIR demonstrated superior performance over the rest model in comparison. Specifically, it calibrated raw inputs of logistic regression in all four cases while Isotonic Regression succeed two times and Platt Scaling failed in all cases. For the real world data, SIR also demonstrated better performance than IR and PS. SIR calibrated six out of a total of eight cases while IR and PS only succeeded in five cases and one case, respectively. Both experiments indicated that SIR are more preferable than existing approaches in *calibration*. In addition, the new model has good generalization abil-

ity. It can be directly used in calibrating outputs of binary outcomes from "uncalibrated" models like logistic regression, support vector machine and decision tree. SIR provides a better fitted model than PS and less overfitted results comparing to IR in most cases thanks to the consideration of model smoothness. It is capable of producing more reliable risk probabilities than existing approaches.

Chapter 7

Adaptive Calibration for Logistic Regression

[1]Most binary classifiers are used to provide not just class labels but also instance-based scores, which are often used in interpretation of the confidence the classifier has about its estimation, usually, the higher the score, the more likely an instance is assigned positive. In many situations, scores generated by classifiers are used to estimate the posterior class membership probabilities. These scores are necessary when the classifier is used for cost-sensitive applications, in which precise judgments about the cost of errors must be made [62]. However, raw scores are not always good estimates of true probabilities. Some model classes are notoriously poor at producing accurate estimates [136], so before using scores as posterior probability estimates, they must be calibrated.

In the previous chapter, I introduced a Smooth Isotonic Regression method for *calibration*. The method demonstrated superior performance than Platt Scaling and Isotonic Regression in a few synthetic and real world data. In addition, it does not need human intervention to adjust parameters. As a hybrid of non-parametric and parametric model, the method is applicable to cases where previous using Isotonic Regression to recalibrate the probabilistic outputs. However, there is a limitation of this approach, which is inherited by Isotonic Regression method. Smooth Isotonic Regression enforces smoothness using a Piecewise Cubic Hermite Interpolating Polynomials (PCHIP) to the *calibrate* outputs of Isotonic Regression (IR). Because

[1]A version of this chapter is under review at Nature Biotechnology. [93].

both functions (PCHIP, IR) are monotonic, the combined mapping function are still monotonic. The results are largely constrained by the objective of Isotonic Regression. However, in many real world cases, the probabilistic outputs do not necessarily have monotonicity and Smooth Isotonic Regression fails to fit these cases. In addition, Smooth Isotonic Regression cannot improve models' discrimination performance because monotonic functions do not change the partial ranking of the observed cases.

A well calibrated probabilistic classifier produces outputs that represent "true probability" of underlying events (class membership). To optimize both aspects of a probabilistic model: *discrimination* and *calibration*, I have to break the global monotonic constraints, but instead, calibrate probabilistic outputs adaptively. The reason for this is "true probability" of underlying events are unknown but may be estimated by class membership of similar patterns. That is, from a frequentist perspective, a reliable estimator of the "true probability" is the fraction of positive labeled cases of statistically similar cases. A key insight is that good *calibration* should be performed locally as opposed to that good *discrimination* can be examined globally.

In this chapter, I focused on a specific model: Logistic Regression because it is popular and widely used in various biomedical tasks [14, 71, 110, 133, 142, 165]. But my method can be easily extended to any probabilistic models that output confidence interval of predictions. As opposed to previous approaches that are input irrelevant, my method adaptively includes model-specific input information. The method is data-driven and only relevant information are considered for each prediction.

7.1 Data

In this chapter, I used three clinical data: hospital discharge error (HOSPITAL), Myocardial_Infarction (MI) and Breast Cancer (BREAST_CANCER). All these data have binary outputs, thus I can test model's *discrimination* and *calibration* ability against the ground-truth.

7.1.1 Hospital data

The HOSPITAL data set consists of microbiology cultures and other variables related to hospital discharge errors [59]. For patient demographic data, this data contains age, gender, race and insurance. Related to the hospital encounter, the dataset contains the visit type (admission, emergency room, procedure or outpatient) and admitting service, if applicable. Related to the microbiology result, the dataset contains the specimen type (blood, urine, sputum and cerebral spinal fluid), the hospital day number that the specimen was col-

lected, whether the result was pending at the time of discharge from the hospital, whether the specimen was collected on a weekend, whether the preliminary results (for blood cultures) were reported on a weekend, and whether the final results were reported on a weekend. In addition to the data pulled directly from the hospital computer system, this dataset contains an additional outcome variable, which indicates whether the case represents a potential post-discharge follow-up error using expert's knowledge. This variable is true if the following three criteria are met: (1) the result is considered clinically relevant; (2) the results return after the patient is discharged from the hospital; and (3) there is not an antibiotic on the discharge medication list to which the organism is sensitive based on the microbiology results. The features are thus consisted of eight categorical variables and two numerical variables. The target is a Boolean variable (Pot_error) indicating the potential error.

The following table defines various features and outcome variables for this data.

Table 7.1: Details of co-variables and the target variable in the hospital discharge error data. Eight out of ten explanatory variables are categorical and two are numerical.

Name	**Details**
	Features
Specimen:	0=blood, 1=urine, 2=sputum, 3=csf
Spec_days:	Number of days between admission date and specimen collection date.
Collect_week:	0=specimen collected on weekday, 1=specimen collected on weekend
Final_week:	0=final result on weekday, 1=final result on weekend
Vistyp:	1=admission, 0=non-admission
Svc:	0=<blank> (patient not admitted), 1=ONC, 2=MED, 3=Medical Sub-specialties, 4=Surgery and Surgical Sub-specialties, 5=Other
Age:	Age in years
Female:	0=male, 1=female
Race:	0=white, 1=black, 2=Asian, 3=Hispanic, 4=other, 5=unknown/declined
Insurance:	0=medicare, 1=medicaid, 2=commercial, 3=other
	Target Variable
Pot_error:	0=not a potential follow-up error, 1=a potential follow-up error

I also summarized co-variables and the outcome variable with their description statistic, i.e., min, 1st Qu., median, 3rd Qu. ,max.

Table 7.2: Descriptional statistic for the hospital discharge error data set.

spec	spec dayssinceadm	collect we	final we	vistype	svc
0.161806	Min. : 0.000	2.607639	2.488194	3.3875	0.503472
1.822222	1st Qu.: 1.000	0.779861	0.899306		0.977083
1.102083	Median : 2.000				0.970139
0.509722	Mean : 4.355				1.283333
	3rd Qu.: 4.000				5:41
	Max. :195.000				
age	female	race	insurance	pot error	
Min. : 0.00	1.563889	2.333333	1.386111	3.089583	
1st Qu.:43.28	1.823611	0.442361	0.426389	0.297917	
Median :57.76		0.159722	1.577778		
Mean :56.51		0.40625	0.205556		
3rd Qu.:71.24		4:55			
Max. :99.71		0.424306			

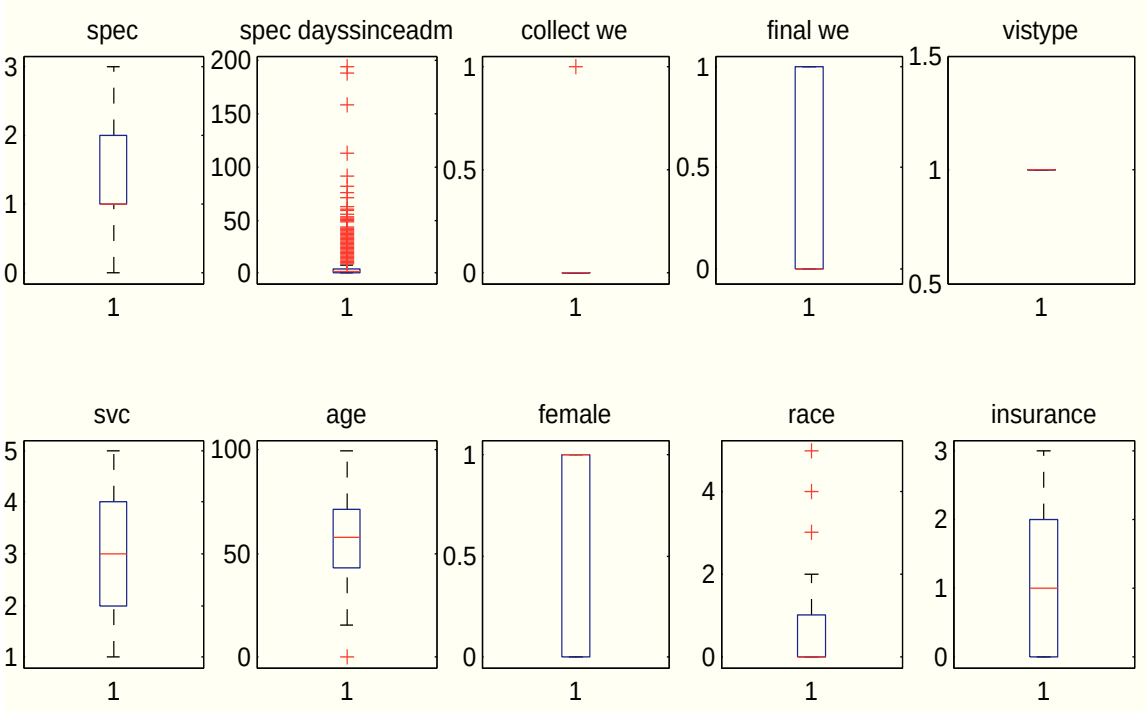


Figure 7.1: Boxplots for ten co-variates of the hospital discharge error data.

To see these visually, I box plotted the co-variates in Figure 7.1. It is easy to observe that there are two numerical co-variables and eight categorical ones in hospital discharge error data. I expressly represented each categorical co-variable with a set of binary co-variables. The total number of expanded co-variables

are 20.

There are 369 clinically important but highly suspicious observations out of 4819 returned post-discharge observations, which makes the data highly unbalanced and challenge to calibrate. More details about this data can be found in Chapter 2.

7.1.2 Myocardial Infarction data

Myocardial Infarction data contain clinical Myocardial infarction (MI) patient records[98]. The goal of this study is to determine which, and how many data items are required to construct a decision support algorithm for early diagnosis of acute myocardial infarction using clinical and electrocardiographic data available at presentation [98].

These data are collected from patients admitted and discharged on a regimen. The data contains patient records of two medical centers in the Great Britain; among these, 500 patients admitted to the emergency department with chest pain are observed in Sheffield, England, and 1,353 patients with the same symptoms are observed in Edinburgh, Scotland.

The total number of the patients is 1,853, the feature size is 54 and the target is a binary variable indicating whether a patient has myocardial infarction (MI). Table 7.3 summarizes co-variables and their clinical meanings. Note the last six features $(49 - 54)$ correspond to electrocardiograph readings that are highly correlated to the target, should not be included for prediction. I represent every categorical feature by a set of binary features to preserve the categorical information, in order to be applicable to some machine learning algorithms.

Table 7.3: Explanations for different variables of the Myocardial_Infarction data.

ID	Abbreviation	Clinical Explainations
1-7	age	Age in years (under 30, 30-39, 40-49, 50-59, 60-69, 70-79, 80 and over)
8	Smokes	Smoker
9	Exsmoker	Ex-smoker
10	Fhistory	Family history of ischaemic heart disease
11	Diebetes	Diabetes mellitus
12	BP	Hypertension
13	Lipids	Hyperlipidaemia
14	CPmajorSymp	Is chest pain the major symptom?
15	Restrostern	Central chest pain
16	Lchest	Pain in left side of chest
17	Rchest	Pain in right side of chest
18	Back	Pain radiates to back
19	Larm	Pain radiates to left arm, neck or jaw
20	Rarm	Pain radiates to right arm
21	breath	Worse on inspiration
22	postural	Pain related to posture
23	Cwtender	Chest wall tenderness
24	Sharp	Pain described as sharp or stabbing
25	Tight	Pain described as tight, heavy, gripping or crushing
26	Sweating	Sweating
27	SOB	Short of breath
28	Nausea	Nausea
29	Vomiting	Vomiting
30	Syncope	Syncope
31	Episodic	Episodic pain
32-36	Worsening	Hours since 1st symptom (0-5, 6-10, 11-20, 21-40, over 40)
37-42	Duration	Hours of pain at presentation (0-5, 6-10, 11-20, 21-40, 41-80, over 80)
43	prev-ang	History of angina
44	Prev-MI	Previous myocardial infarction
45	Worse	Worse than usual angina/similar to previous acute myocardial infarction
46	Crackles	Fine crackles suggestive of pulmonary oedema
47	Added-HS	Added heart sounds
48	Hypoperfusion	Signs of hypoperfusion
49	Stelve	New ST-segment elevation
50	NewQ	New pathological Q waves
51	STorT-abnorm	ST segment or T-wave changes suggestive of ischaemia
52	LBBBorRBBB	Bundle branch block
53	Old-MI	Old electrocardiogram features of myocardial infarction
54	Old-isch	Electrocardiogram signs of ischaemia known to be old

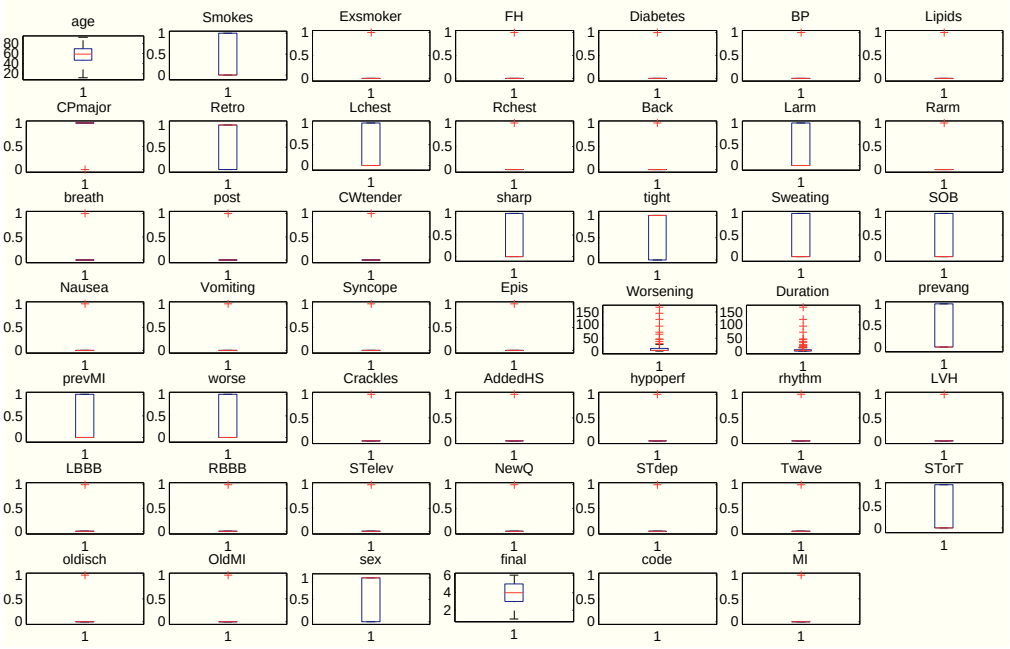
The tables below list the descriptional statistics of Edinburgh and Sheffield datasets. I also included boxplots of these datasets to illustrates their co-variables' distribution visually.

Table 7.4: Descriptional statistic for the Edinburgh data.

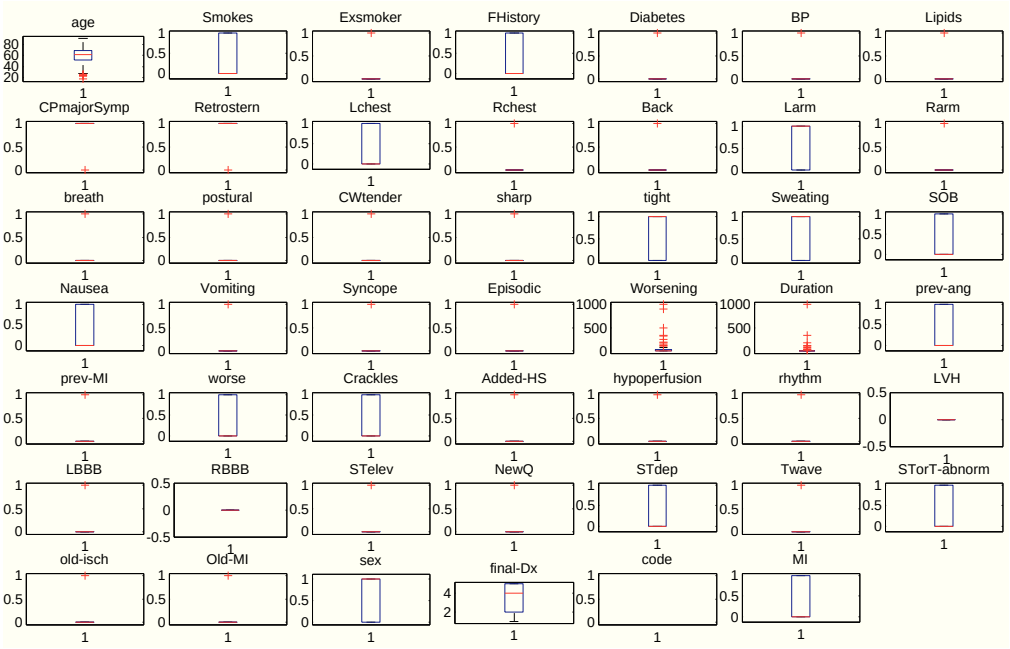
Abbreviation				
age	min: 13.0	median:59	mean:57.6	max: 92
Smokes	0: 785	1: 468		
Exsmoker	0: 959	1: 294		
Fhistory	0: 967	1: 286		
Diebetes	0: 1165	1: 88		
BP	0: 1053	1: 200		
Lipids	0: 1215	1: 38		
CPmajorSymp	0: 62	1: 1191		
Restrostern	0: 331	1: 922		
Lchest	0: 907	1: 346		
Rchest	0: 1109	1: 144		
Back	0: 1122	1: 131		
Larm	0: 670	1: 583		
Rarm	0: 1042	1: 211		
breath	0: 1031	1: 222		
postural	0: 1017	1: 236		
Cwtender	0: 1201	1: 52		
Sharp	0: 1208	1: 45		
Tight	0: 572	1: 681		
Sweating	0: 739	1: 514		
SOB	0: 731	1: 522		
Nausea	0: 1124	1: 129		
Vomiting	0: 1124	1: 129		
Syncope	0: 1208	1: 45		
Episodic	0: 1161	1: 92		
Worsening	min: 0.0	median: 4.0	mean: 17.4	max: 168
Duration	min: 0.0	median: 3.0	mean: 8.84	max: 168
prev-ang	0: 699	1: 554		
prev-MI	0: 836	1: 361		
Worse	0: 892	1: 361		
Crackles	0: 1106	1: 147		
Added-HS	0: 1247	1: 6		
Hypoperfusion	0: 1203	1: 50		
Stelve	0: 1199	1: 54		
NewQ	0: 1240	1: 13		
STorT-abnorm	0: 1240	1: 13		
LBBBorRBBB	0: 1203	1: 50		
Old-MI	0: 1101	0: 152		
Old-isch	0: 1141	1: 112		
MI	0: 979	1: 274		

Table 7.5: Descriptional statistic for the Sheffield data.

Abbreviation				
age	min: 17.0	median:61	mean:59.9	max: 91
Smokes	0: 318	1: 182		
Exsmoker	0: 388	1: 112		
Fhistory	0: 373	1: 127		
Diebetes	0: 451	1: 49		
BP	0: 403	1: 97		
Lipids	0: 482	1: 18		
CPmajorSymp	0: 37	1: 463		
Restrostern	0: 110	1: 390		
Lchest	0: 373	1: 127		
Rchest	0: 438	1: 62		
Back	0: 426	1: 74		
Larm	0: 237	1: 263		
Rarm	0: 418	1: 82		
breath	0: 422	1: 78		
postural	0: 455	1: 45		
Cwtender	0: 491	1: 9		
Sharp	0: 400	1: 100		
Tight	0: 246	1: 254		
Sweating	0: 235	1: 265		
SOB	0: 281	1: 219		
Nausea	0: 341	1: 159		
Vomiting	0: 449	1: 51		
Syncope	0: 467	1: 33		
Episodic	0: 417	1: 83		
Worsening	min: 0.0	median: 6.0	mean: 50.37	max: 1000
Duration	min: 0.0	median: 4.0	mean: 12.34	max: 1000
prev-ang	0: 281	1: 219		
prev-MI	0: 377	1: 123		
Worse	0: 338	1: 162		
Crackles	0: 373	1: 127		
Added-HS	0: 476	1: 24		
Hypoperfusion	0: 441	1: 59		
Stelve	0: 403	1: 97		
NewQ	0: 470	1: 30		
STorT-abnorm	0: 403	1: 97		
LBBBorRBBB	0: 474	1: 26		
Old-MI	0: 454	1: 46		
Old-isch	0: 473	1: 27		
MI	0: 346	1: 154		



(a) Edinburgh MI data.



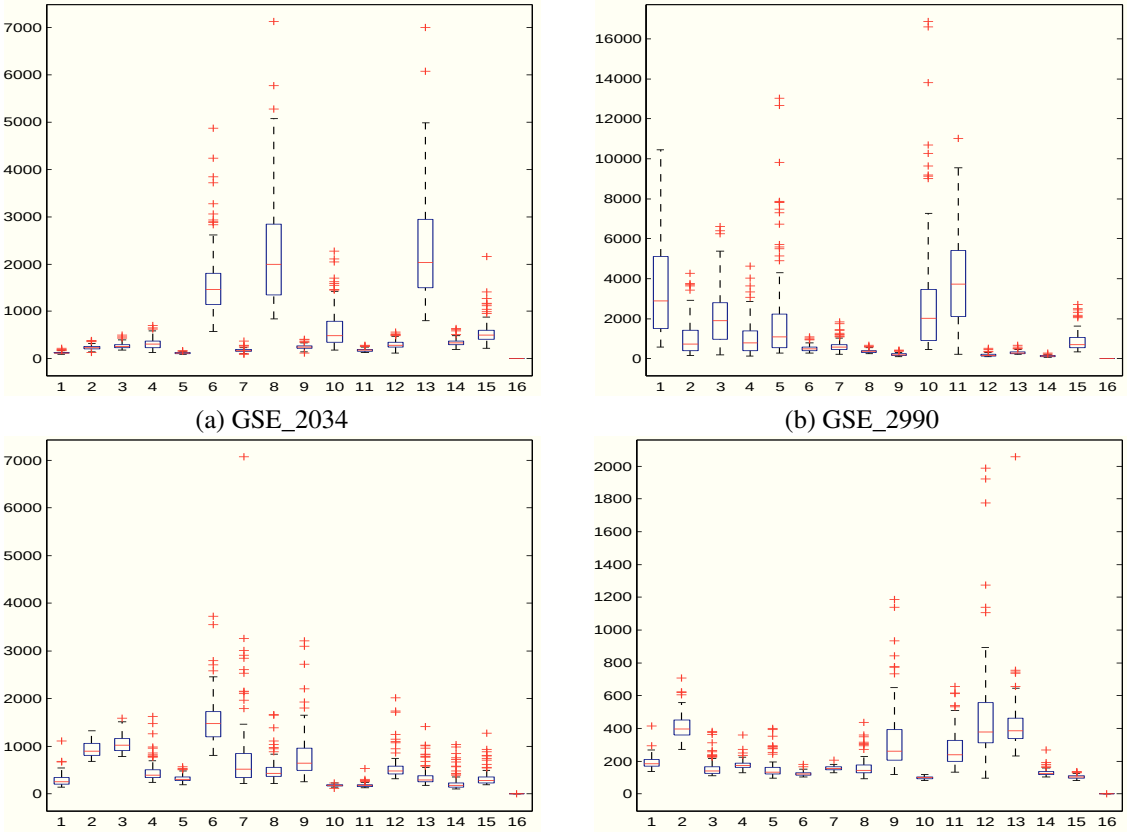
(b) Sheffield MI data.

Figure 7.2: Boxplots of Myocardial Infarction data.

7.1.3 Breast Cancer Gene Expression data

The Breast Cancer Gene Expression data were obtained from the NCBI Gene Expression Omnibus (GEO). Three individual data downloaded are previously studied by Wang et al. (GSE2034) [187], Sotiriou et al. (GSE2990) [166], and Miller et al. (GSE3494) [128], respectively.

To make my data comparable to the previous studies, I followed the criteria in [140] to select patients, who did not receive any treatment and had negative lymph node status. Among these pre-selected candidates, only patients with extreme outcomes, either poor outcomes (recurrence or metastasis within five years) or good outcomes (neither recurrence nor metastasis within eight years) are selected. The number of samples after filtering are: 209 for GSE2034 (114 good/95 poor), 90 for GSE2990 (60 good/30 poor), and 242 for GSE2034 (224 good/18 poor).

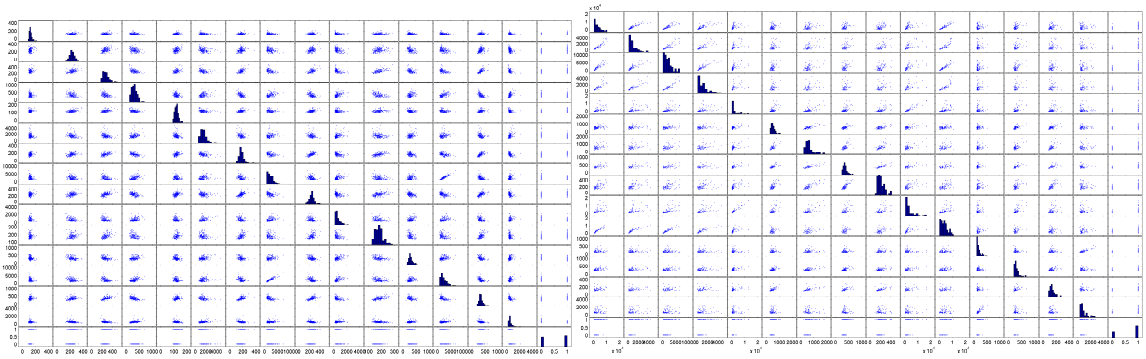


(a) GSE_2034 (b) GSE_2990

(c) GSE3494_U133A (d) GSE3494_U133B

Figure 7.3: Boxplots of GSE_2034, GSE_2990 and GSE_3493. Each column corresponds to one feature vector and the last column indicates the outcome variable.

I applied a split to divide GSE3494 into two groups, as suggested by [140], GSE3494-A and GSE3493-B, according to the sample's Affymetrix platform. Thus, the breast cancer data-set has four separate data: GSE2034, GSE2990, GSE3494-A and GSE3494-B. All of these data have a feature size of $247,965$, corresponds to the gene expression results obtained from micro-array experiments. They were preprocessed to keep only the top 15 features ranked using t-test (see [140] for details). Figure 7.3 shows boxplots of these selected gene features. It can be observed in figures below that effective gene feature are different from each other in different population groups.



(a) GSE_2034 (b) GSE_2990

(c) GSE3494_U133A (d) GSE3494_U133B

Figure 7.4: Matrix plots of GSE_2034, GSE_2990 and GSE_3493. Each subfigure corresponds a matrix plot of one data set.

I also plotted co-variable occurrence of breast cancer data in Figure 7.4 to investigate feature correlations visually.

7.2 Background

A well calibrated probabilistic classifier produces outputs that represent the true probability of underlying events (class membership). Oftentimes, the outputs of classifiers are combined with other sources of information, e.g. domain knowledge, mis-classification cost. Probabilistic models for classification aim to maximize the likelihood of observing the training data and assign to each test case a continuous output between 0 and 1, which are interpreted as class membership probability estimates. It is well know that the results of several machine learning approaches, e.g. naive Bayes and decision trees, are not well-calibrated [58], and thus may not always be reliably trusted as a proxy for the absolute risk in clinical decision making. To address this problem, several *calibration* methods have been independently proposed to adjust the outputs of popular machine learning models [45, 205, 206, 207].

These *calibration* methods can be generally divided into two main categories: parametric methods and non-parametric methods. Platt suggested a parametric approach that transforms the probabilistic outputs into posterior probabilities [147] by re-fitting these outputs to a sigmoid function. Thus, the KL-divergence from post-processed outputs to true class labels is minimized. The parameters of the sigmoid function are estimated using maximum likelihood estimation (MLE). Essentially, this approach adjusts the outputs of any probabilistic model (including LR) by an independent one-dimension logistic function. However, this approach may have difficulties when the outputs cannot be sufficiently modeled by a sigmoid function (e.g., patterns that demonstrate non-monotonicity); in those cases, the MLE can lead to poor *calibration* results.

Zadrozny and Elkan [206] proposed a non-parametric approach that utilizes Isotonic Regression (IR), which involves finding a weighted least square fit with the following form: $\min \sum_i (\hat{y}_i - y_i)^p$ subject to $\hat{y}_i \geq \hat{y}_j, \forall (i,j)$, where p is the order of the norm, y_i is the binomial class label and $\hat{y}_i$ is the "calibrated" output. The formula is subject to a set of monotonicity constraints that enforces a partial order over the variables. When $p = 2$, there is an efficient pair-adjacent violators (PAV) algorithm to solve it [26]. However, the result of the *calibration* is non-continuous and tends to over-fit [186].

Osl et al. [141] suggested a novel approach to improve *calibration* of logistic regression models by including additional information from the input space. The idea is to adjust outputs of a fitted logistic regression model within local regions in the input space. These local regions are determined in a pre-processing step using Gaussian Mixture Models. This approach improved previous methods that only focus on the output space. However, this approach requires an additional pre-processing step to cluster cases in the

input space, and the user has to tune more parameters, e.g., number of clusters and threshold used to exclude outliers.

In contrast to the approaches listed above, I showed that including model-specific input information helps to improve *calibration* without increasing the modeling complexity, or introducing new parameters.

My approach extends Logistic Regression model in a non-parametric way. I calibrate the outputs of a logistic regression locally by using model-specific confidence intervals (C.I.) for each individual prediction. Intuitively, I use a smaller neighborhood to calibrate the outputs when the region ("locality") of a test point is dense and I use a larger neighborhood to calibrate the outputs when the region ("locality") of a test point is sparse. In addition, my method is capable of handling non-linear separable patterns and offering increased *discrimination* ability. I introduce my method in the next Section 7.3. In Section 7.4, I evaluate my method on both synthetic and real medical data; finally, I conclude my findings in Section 7.5.

7.3 Method

In this section, I started with a review of the Logistic Regression model, followed by the estimation of logistic regression parameters and the estimation of confidence intervals (C.I.) for each prediction. Finally, I introduced my adaptive *calibration* approach.

7.3.1 Logistic Regression Review

Logistic Regression (LR) aims to learn a function of the form $f : X \rightarrow Y$, or $P(Y|X)$, where $Y \in \{0, 1\}$ is the true class label, and $X =< x_1, \ldots, x_n >$ is a vector of discrete or continuous values. Let $\mathbf{X} = \{X^l\}_{l=1}^{L}$ indicates the corpus of features and $\mathbf{Y} = \{Y^l\}_{l=1}^{L}$, where l represents the index of data points and L is the size of training samples. The model assumes a parametric form for the distribution $P(Y|X)$, whose parameters can be estimated from the training data. Let's denote "1" as the event $Y = 1|X$ and "0" as the event $Y = 0|X$. The parametric model of Logistic Regression is thus defined as:

$$P(\text{"1"}) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i x_i)}, \tag{7.1}$$

$$P(\text{"0"}) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i x_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i x_i)}, \tag{7.2}$$

which can be also written as:

$$Logit(P) = w_0 + \sum_{i=1}^{n} w_i X_i, \tag{7.3}$$

where $Logit(P) = \ln(\frac{P}{1-P})$.

> **Interpretation:** Logistic Regression is also known as a linked function of Generalized Linear Model because it links a Logit function to a linear function.

7.3.2 Parameter Estimation

Before discussing how to infer the confidence interval for each prediction, I introduce estimation procedures for mean and standard deviation of model parameters W.

7.3.2.1 Mean of the Weight Parameters

The parameters W of a LR model is optimized using maximum likelihood estimation (MLE):

$$W \leftarrow \text{argmax}_W \prod_l P(Y^l|X^l, W), \tag{7.4}$$

where $W =< w_0, w_1, \ldots, w_n >$ is the vector of parameters to be estimated, X^l and Y^l represent the l-th training example and class label, respectively. Because the nature logarithm does not change the parameters of the objective function (Equation 7.4) at its optimal, it is more convenient to work with the log-sum (Equation 7.5) instead of the products in Equation 7.4.

$$W \leftarrow \text{argmax}_W \sum_l \ln P(Y^l|X^l, W). \tag{7.5}$$

> **Interpretation:** Log-likelihood objectives are optimized rather than the original likelihood function for two reasons: 1. it alliviates the problem of floating point error due to the production of tiny probabilities; 2. it induces factorizable objective function which can be optimized easily.

I can explicitly expand the log-likelihood $l(W)$ as,

$$l(W) = \sum_l Y^l \ln P(Y^l = 1|X^l, W) + \\ + (1 - Y^l) \ln P(Y^l = 0|X^l, W),$$

which can be rewritten as:

$$l(W) = \sum_l Y^l (w_0 + \sum_{i=1}^{n} w_i x_i^l) \\ - \ln(1 + \exp(w_0 + \sum_{i=1}^{n} w_i x_i^l)), \tag{7.6}$$

where x_i^l indicates the i-th feature of the l-th training point.

> **Interpretation:** Logistic Regression got its "ambiguous" name "regression" for some historic reasons. But it is a discriminative classification method for binary outcomes.

I briefly review how to use maximum likelihood criteria in estimating the model parameters. As the closed form of W cannot be obtained from Equation 7.6, I work with the gradient, which is the partial derivative of W. Note the i-th component of this partial derivative has the following form:

$$\frac{\partial l(W)}{\partial w_i} = \sum_l x_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W)), \tag{7.7}$$

where $\hat{P}(Y^l = 1|X^l, W)$ is the Logistic Regression prediction with the old W value. To consider w_0 (e.g., the intercept of a linear model in two-dimension) in the derivatives, I include an illusory $X_0 = 1$ for all data instance$l$. As the log-likelihood function 7.6 is concave, it is guaranteed to obtain an optimal value if I keep moving W towards the direction of the gradients:

$$w_i \leftarrow w_i + \eta \sum_l x_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W)), \tag{7.8}$$

where η is constant step size.

Interpretation: There is no closed form solution to weight parameters W. A common approach to obtain W is to use gradient descent algorithm.

7.3.2.2 Standard Deviation of the Weight Parameters

I can calculate the second derivative of the likelihood to get the Hessian matrix H,

$$H_{ij} = \frac{\partial l(W)}{\partial w_i \partial w_j} = \mathbf{X}^T \mathcal{V} \mathbf{X}, \tag{7.9}$$

where $\mathcal{V}$ is a $L \times L$ diagonal matrix of weights with l-th element $\hat{P}(Y^l = 1|X^l, W)(1 - \hat{P}(Y^l = 1|X^l, W))$. Under the delta method and asymptotic efficiency theory, I can approximate the parameter covariance matrix by the inversion of the Hessian,

$$\Sigma \approx \frac{1}{E_W(-\frac{\partial^2 l(W)}{\partial W^2})} = -H^{-1}, \tag{7.10}$$

where $E_W(-\frac{\partial^2 l(W)}{\partial W^2})$ is also known as the fisher information matrix; the standard errors of each w_i are thus estimated by the square root of the diagonal elements,

$$s.e. = \sqrt{Diag(\Sigma)}. \tag{7.11}$$

Interpretation: Standard deviation of the weight parameters can be approximated using the second derivative of the likelihood.

7.3.3 Confidence Interval of the Estimate Prediction

Given the mean and standard deviation of weight parameters, I can estimate the confidence interval for each prediction. Let's rewrite LR in an alternative way as a general linear model (GLM) that is linked to an inverse logit function:

$$\ln(\frac{P_X}{1-P_X}) = w_0 + \sum_{i=1}^{n} w_i x_i, \tag{7.12}$$

where P_X is the estimated probability of data point X. Let $X_0 = 1$, so that

$$Z(P_X) = \ln(\frac{P_X}{1-P_X}) = \sum_{i=0}^{n} w_i x_i. \tag{7.13}$$

I can estimate the variance of $Z(P_X)$ first, and then use the Delta method to estimate the true variance of P. If I treat W as random variables and X_i as the fixed value, its standard deviation is estimated in Equation 7.11:

$$\begin{aligned} var\left(Z(P_X)\right) &= var(\sum_{i=0}^{n} w_i x_i) \\ &= (\Sigma^{\frac{1}{2}})' X^2 (\Sigma^{\frac{1}{2}}), \end{aligned} \tag{7.14}$$

where Σ corresponds to the covariance (Equation 7.10) of the W and $\Sigma = (\Sigma^{1/2})'(\Sigma^{1/2})$.

> **Interpretation:** The variance of each prediction can be calculated by wrapping the variance of weight parameters to the outputs.

Because $Z(P_X) \sim N(\cdot, (\Sigma^{\frac{1}{2}})' X^2 (\Sigma^{\frac{1}{2}}))$ and $P_X = h(Z) = \frac{e^Z}{1+e^Z}$, I can easily verify $P_X \sim N(\cdot, var(Z)\left(h'(Z)\right)^2)$. I explicitly write $var(P_X) = var(Z)\left(h'(Z)\right)^2$, where $h'(Z) = \frac{e^Z}{(1+e^Z)^2}$, as follows:

$$\begin{aligned} var(P_X) = &\left((\Sigma^{\frac{1}{2}})' X^2 (\Sigma^{\frac{1}{2}})\right) \\ &* \left(P_X^2 (1-P_X)^2\right). \end{aligned} \tag{7.15}$$

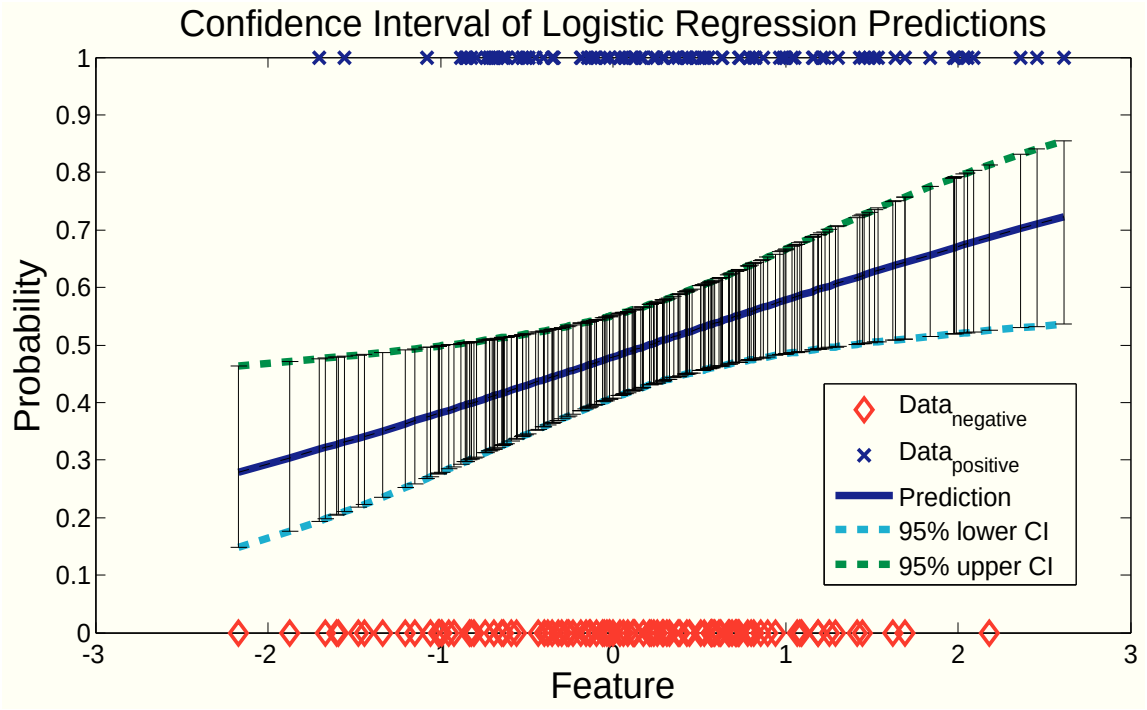


Figure 7.5: Logistic regression predictions and associated estimated confidence intervals. Data are sampled from two Gaussian distributions $N(0,1)$ and $N(0.5,1)$; the former is labeled as class 0 points (bottom) and the later labeled as class 1 points (top).

The confidence interval ($\mathcal{CI}$) of P_X is $P_X \pm 1.96 * \sqrt{var(P_X)}$. Figure 7.5 illustrates the predictions and their 95% confidence intervals on a simulated data-set. As indicated by this figure, the dense regions are associated with narrow Confidence Intervals, and vice verse.

7.3.4 Adaptive Calibration

It is believed that calibrated estimations $p(X)$ are smooth, in other words, data points that are close should have approximately the same probability [4]. Ideally, if I want to estimate $p(X^*)$ for a novel data point X^*, I should select a neighborhood of X^* and calibrate the raw probabilistic output of X^* as if the data point were actually a sample taken in this neighborhood. An intuitive estimator is thus:

$$p(X^*) = \frac{1}{|\mathcal{N}(X^*)|} \sum_{X^j \in \mathcal{N}(X^*)} Y^j, \tag{7.16}$$

where X^j and Y^j correspond to a neighboring point of X^* and its class label, respectively. Here $\mathcal{N}(X^*)$ denotes the neighborhood of X^*. Depending on the construction criteria for this neighborhood, Equation 7.16 could be a nearest neighbor estimator if I select a fixed number of K points; or a Parzen window estimator if I choose a fixed bandwidth $\epsilon \geq \max(|X^* - X^j|)$ s.t. $\forall X^j \in \mathcal{N}(X^*)$. Given a reasonable K or ϵ, the estimator induced by Equation 7.16 represents a local fraction of positives.

However, it is non-trivial to select the best K or ϵ. First, computational complexity could be demanding because these estimators need to find the neighborhood for every novel X^* at run-time. Furthermore, there might be no single K or ϵ that works best for all the testing data points. Thus, I propose a new approach, Adaptive Calibration for Logistic Regression (AC-LR), to close the performance gap.

My method adaptively calibrates LR (indicated in Figure 7.6) with a varying bandwidth (Equation 7.17) as follows:

$$p(X^*) = \frac{1}{|\mathcal{CI}(X^*) \otimes r(P)|} \sum_{X:P(X) \in \mathcal{CI}(P^*) \otimes r(\mathbf{P})} Y^j, \tag{7.17}$$

where $\mathcal{CI}(X^*)$ corresponds to the confidence interval of a prediction for a novel data point X^*, $r(\mathbf{P}) = |\max(\mathbf{P}) - \min(\mathbf{P})|$ indicates the range of the training predictions, where $\mathbf{P} = \{P : X \in \mathcal{D}\}$, $\mathcal{D}$ is the training data corpus.

> **Interpretation:** The adaptive calibration approach induced by Equation 7.17 uses a smaller bandwidth (fewer samples) when Logistic Regression predicts with a high confidence and a wider bandwidth (more points) when Logistic Regression is less confident.

Although $0 \leq r(\mathbf{P}) \leq 1$, this scaling factor is typically close to 1 when the size of training date is large enough. $P(X)$ represents the prediction probability of some training data point within the bandwidth of $\mathcal{CI}(X^*) \otimes r(P)$, where $\otimes$ denotes the Kronecker product. Note $|\mathcal{CI}(X^*) \otimes r(P)|$ indicates the cardinality of the P^* induced neighborhood (number of samples within this locality) instead of a fixed bandwidth.

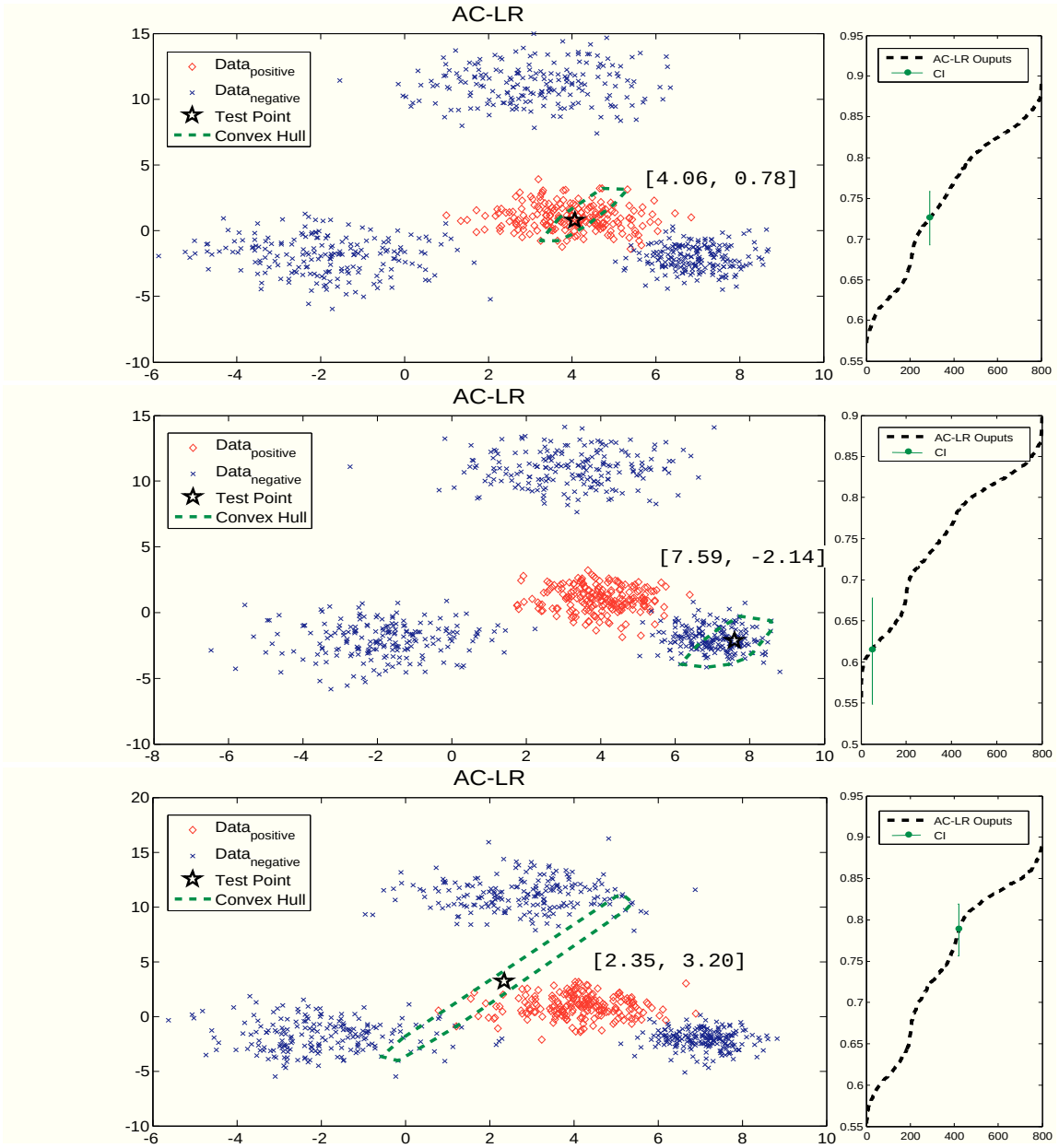


Figure 7.6: Each sub-figure illustrates a test point, its neighborhood convex hull and confidence interval of proposed AC-LR approach for *calibration*.

This adaptive range incorporates the feature correlations induced in the input space to guide local *calibration*. The probability of each test case is adjusted locally using training cases with similar probability. As a result, the AC-LR model is capable of handling many nonlinear separable cases where the original LR model fails. Figure 7.7 gives an example.

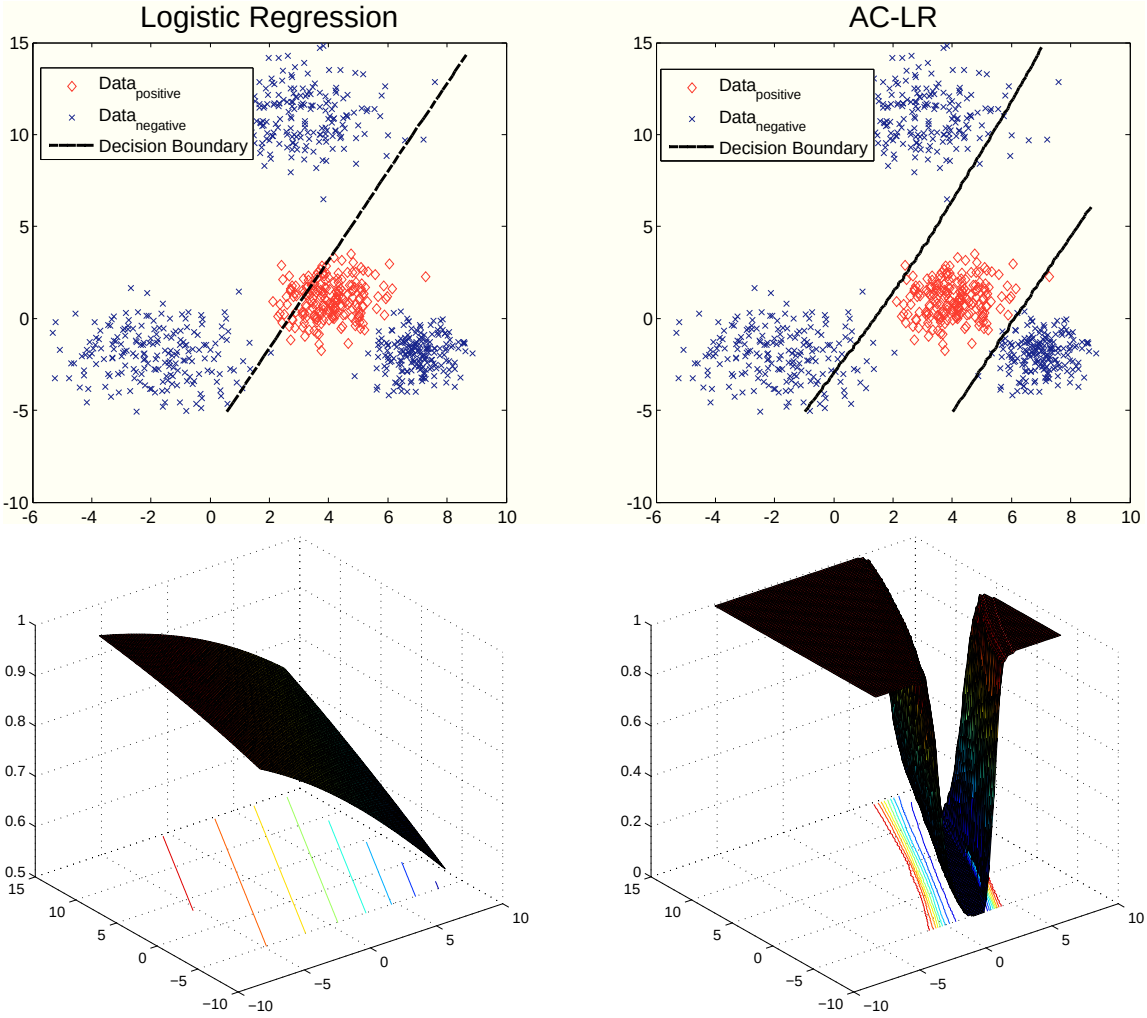


Figure 7.7: Visual comparison of AC-LR and LR models on a simulated 2D data. In the first row, the blue crosses correspond to negative cases, the red diamonds correspond to the positive cases. The black dotted lines indicates the decision boundaries of LR and AC-LR models at threshold p=0.5. In the second row, the surface plots indicate the probability values indicated by both models.

The adaptive calibration approach uses a smaller bandwidth (fewer points) to calibrate the output when LR is sure about a certain prediction $p(X^*)$ and a wider bandwidth (more points) to calibrate the output when LR is less confident. Equation 7.17 suggests a local method that takes both LR predictions and associated confidence intervals into consideration. The method is thus capable of choosing, for each novel testing point X^*, an adaptive bandwidth ϵ^* that maximizes the likelihood. As this *calibration* process takes place in one dimension, it can be optimized in time complexity $O(1)$ using a hash function.

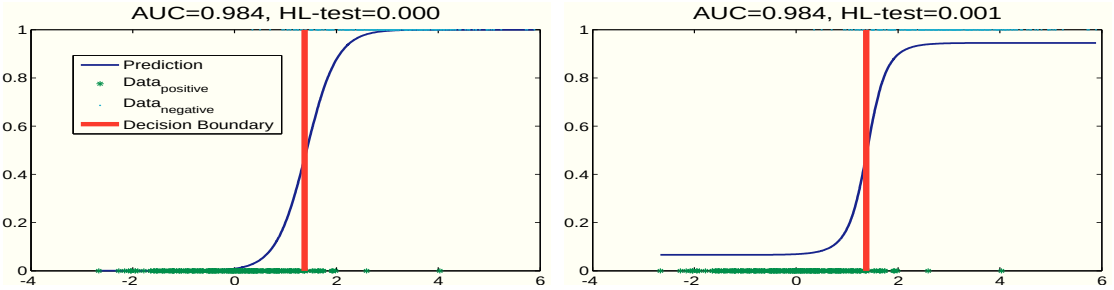
7.4 Results

I evaluate the performance of various models on synthetic and clinical-related data. For the comparison, I use two metrics, the Area Under the ROC Curve (AUC) [83] and the Hosmer-Lemeshow (HL) goodness of fit test [90]. These two are independent measurements of a probabilistic model's performance. AUC is a one number summary of a model's *discrimination*. It can be expressed by the following integration: $AUC = \int_0^1 \frac{TP}{pos} d\frac{FP}{neg}$ where TP and FP correspond to the true positive rate and the false positive rate, respectively. pos and neg corresponds to the cardinality of the positive and negative data points. Thus, AUC counts the concordant pairs out of all positive/negative pairs.

The Hosmer-Lemeshow goodness-of-fit statistic measures the *calibration* of a probabilistic model. It can be written as: $H = \sum_{g=1}^{10} \frac{(O_g - E_g)^2}{N_g \pi_g (1-\pi_g)}$, where O_g, E_g, N_g and π_g correspond to observed positive events, expected positive events, number of total observations and predicted risk for the g^{th} risk deciles, respectively. H is called the Hosmer-Lemeshow H test statistic if deciles are defined as equal-length subgroups of fitted risk values; otherwise, H is called the Hosmer-Lemeshow C test statistic if deciles are defined as equal-size subgroups of fitted risk values. I use the latter definition in our experiments. The distribution of the statistics H is approximated by a chi-square with 8 degrees of freedom. I set $\sigma = 0.05$ to be the significant level to reject null hypothesis that predictions are calibrated. At extreme cases, H statistic could be infinity when $\pi_g = 0$ or $\pi_g = 1$ and Hosmer-Lemeshow goodness-of-fit test cannot handle such cases. In the following sections, I compare four different approaches: Logistic Regression (LR), LR+Platt Scaling (LR-PS), LR+Isotonic Regression (LR-IR), Adaptive Calibration for Logistic Regression (AC-LR).

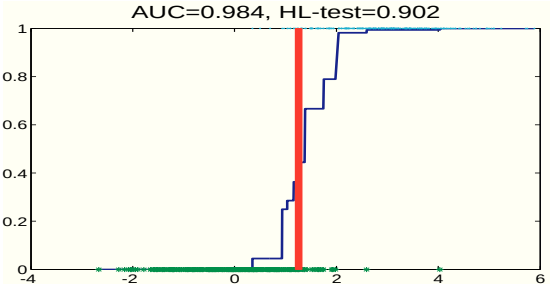
7.4.1 Synthetic Data-set

I first use synthetic data-set to verify our concept in the previous section. For illustration purpose, I sample one-dimensional data so that the probabilistic outputs of all four different approaches (LR, LR-PS, LR-IR, AC-LR) can be demonstrated in figures. The first data-set is generated by sampling from two Gaussian distributions with unit variance but different means $X^0 \in N(0,1)$, $X^1 \in N(3,1)$ and $X = X^1 \cup X^0$, where X^1 and X^0 correspond to data with class label "1" and "0". The results are illustrated in Figure 7.8.

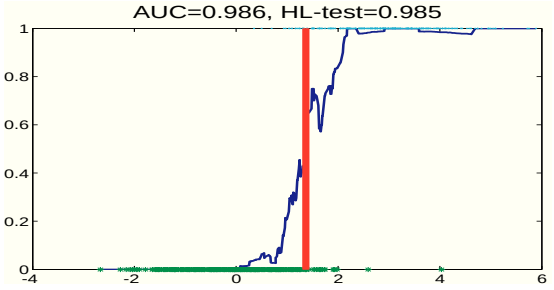


(a) Uncalibrated LR predictions

(b) sigmoid fitting *calibration*

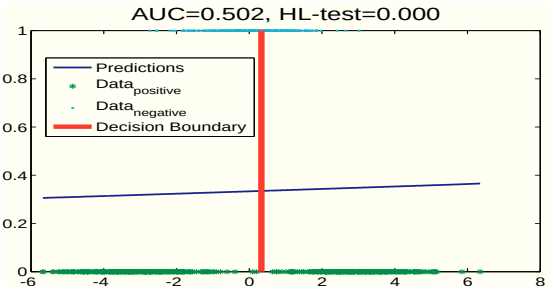


(c) Isotonic regression *calibration*

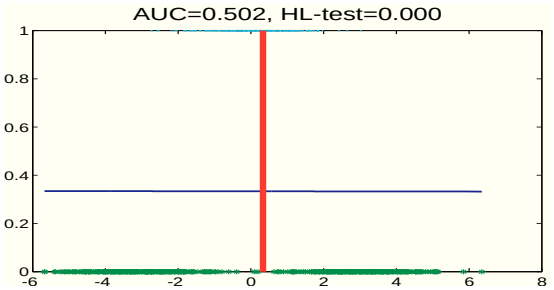


(d) AC-LR *calibration*

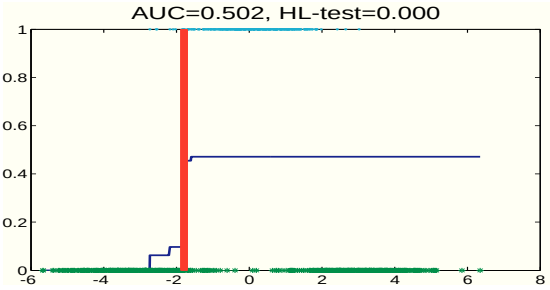
Figure 7.8: In this data, LR and LR-PS does not pass the *calibration* test at 0.05 significance level. LR-IR and AC-LR are both able to calibrate the outputs but the latter has a higher AUC.



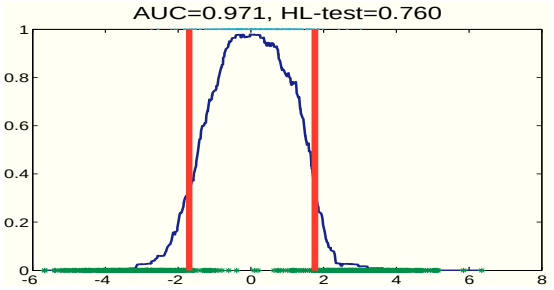
(a) Uncalibrated LR predictions



(b) sigmoid fitting *calibration*



(c) Isotonic regression *calibration*



(d) AC-LR *calibration*

Figure 7.9: This data is linearly non-separable. LR, LR-PS and LR-IR failed in the *calibration* test and have poor AUC. AC-LR is the only approach generates a well-calibrated probabilistic outputs with a large AUC.

For the first data-set, both LR and the LR-PS do not pass the HL-test at significance level $\alpha = 0.05$. LR-IR and AC-LR generate calibrated outputs but AC-LR has a higher AUC. The second data-set is generated to be extreme linearly non-separable, such that $X^0 \in \mathcal{N}(-3,1) \cup \mathcal{N}(3,1)$, $X^1 \in \mathcal{N}(0,1)$ and $X = X^1 \cup X^0$. Figure 7.9 shows the results of various methods. LR and sigmoid fitting have AUCs around 0.5, which are close to the performance of a random classifier. The Isotonic Regression calibrates the output but does not perform well in terms of the AUC. AC-LR demonstrates superior performance, indicating it is capable of handling linear non-separable data (higher AUC) while it calibrates the output ($p > 0.760$). Note LR-IR in Figure 7.9 has thicken decision boundaries due to the presence of multiple testing points at the threshold.

7.4.2 Clinical Related Experiments

7.4.2.1 Hospital Discharge Data

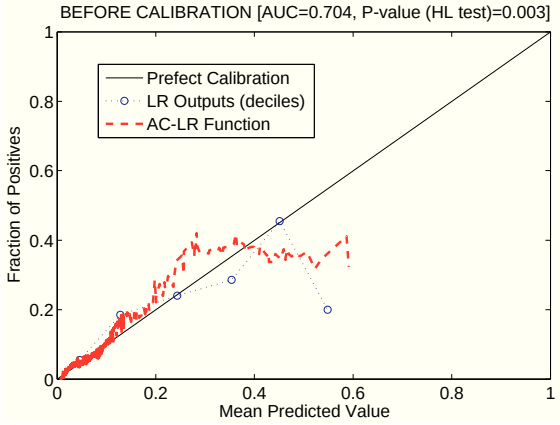
This experiment is conducted on a data used for predicting follow up errors on microbiology cultures. The data-set was created through a retrospective analysis of microbiology cultures performed at Brigham and Women's Hospital in 2007. Specifically, the features consisted of eight categorical variables and two numerical variables.

The target is a Boolean variable indicating the potential error. Table 2.8 explains the clinical meaning of the features. Figure 2.9 illustrates the distribution of each feature variable and the target variable. From a total number of 4,819 hospital discharged cases, 369 are considered as "potential errors" by experts.

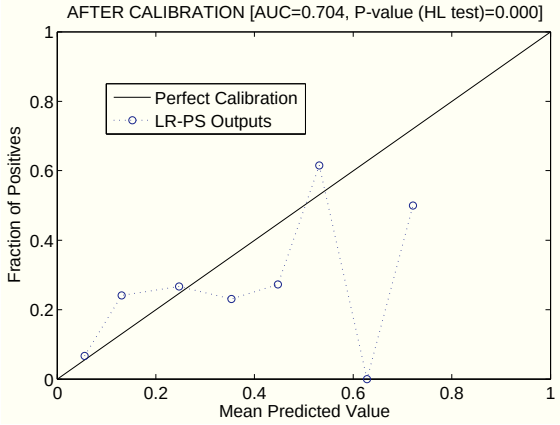
To train a LR model properly, I explicitly expressed each categorical variable as a set of Boolean variables so that different categories are treated fairly, e.g., "sputum=2" does not impact the target more than that of "blood=0" by default. Thus, *spec* was replaced by three Boolean variables. The fully expanded feature space has 22 dimensions. As in the synthetic experiment, I apply different *calibration* models to compare their performance. To generate fair comparison, I randomly split the data into training (66%) and testing (34%) for evaluation. The results are listed in Table 7.6.

Table 7.6: Performance of LR, LR-PS, LR-IR and AC-LR using hospital discharge error data.

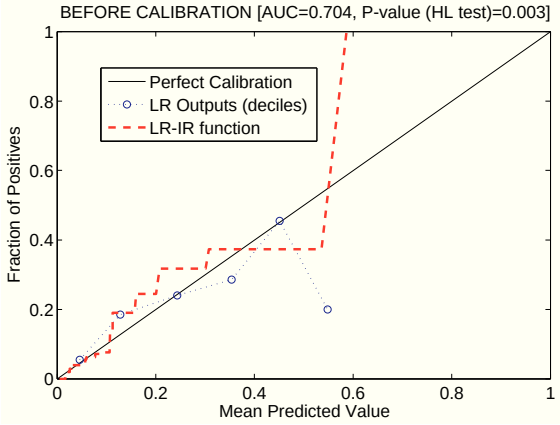
Model	AUC	P-value (HL test)
LR	0.704	0.003
LR-PS	0.704	0.000
LR-IR	0.704	0.000
AC-LR	0.717	0.349



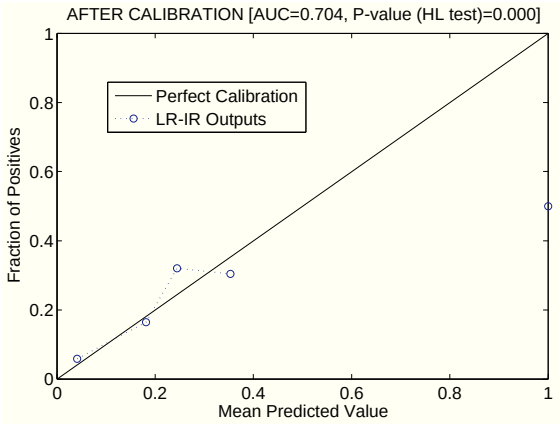
(a) LR-PS *calibration* method.



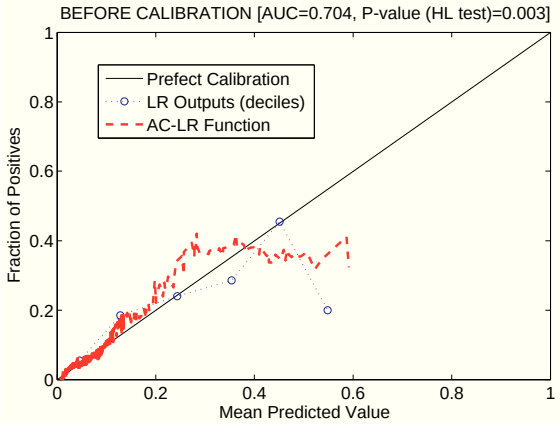
(b) LR-PS rectified outputs.



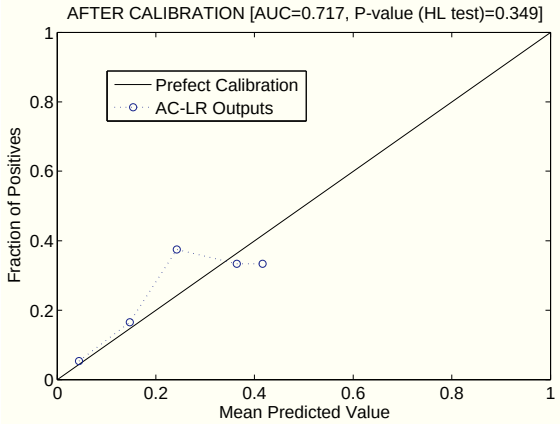
(c) LR-PS *calibration* method.



(d) LR-PS rectified outputs.



(e) AC-LR *calibration* method.



(f) AC-LR rectified outputs.

Figure 7.10: Visual results of various *calibration* methods being applied to the hospital discharge data. Left: the blue dots indicate the outputs of LR model, aggregated in deciles; the red curve corresponds to *calibration* functions. Right: rectified outputs.

As it is easy to see, LR, LR-PS and LR-IR failed to generate calibrated outputs. AC-LR was the only approach that calibrates LR outputs, and even improves the AUC. As indicated in Figure 7.10, the calibrate transformation function (the red curve in (e)) of our model is adaptive; as opposed to LR-PS and LR-IR that enforce a global monotonicity constraint, AC-LR calibrates the outputs locally and adaptively.

7.4.2.2 Myocardial Infarction Data

The Myocardial Infarction (MI) data correspond to results of patient with and without Myocardial Infarction [98]. The original motivation of this study is to determine which, and how many, data items are required to construct a decision support algorithm for early diagnosis of acute myocardial infarction using clinical and electrocardiograph data available at presentation. These data items were collected from patients who were admitted and patients who were discharged. The data contain patient records from two medical centers in Britain; among these, 600 patients attending at the emergency room (ER) with chest pain are observed in Sheffield, England, and $1,253$ patients with the same symptoms are observed in Edinburgh, Scotland.

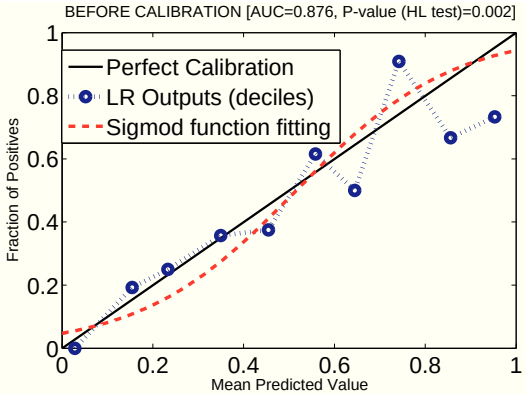
The total number of patients is $1,853$, the feature size is 48 and the target is a binary variable indicating whether or not a patient has MI. Note that I represent every categorical feature by a number of binary ones to preserve the categorical information.

I use a random split to divide the Edinburgh data into training (60%) and testing (40%). Similarly, Sheffield is divided into (60%) and (40%) for training and testing, respectively.

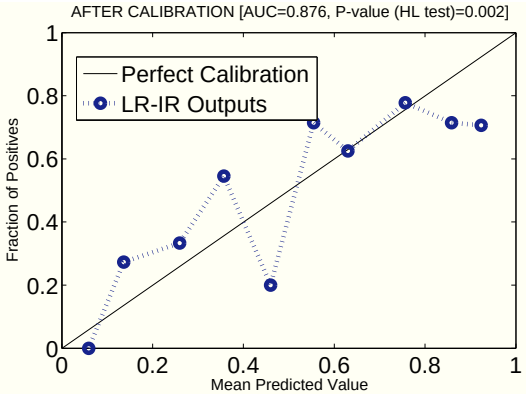
Table 7.7: AUC and HL-test of various *calibration* methods using Myocardial Infarction (MI) data.

Data	Edinburgh		Sheffield	
	AUC	P-value	AUC	P-value
LR	0.876	0.002	0.845	0.023
LR-PS	0.876	0.002	0.845	0.000
LR-IR	0.876	0.000	0.845	0.002
AC-LR	0.880	0.645	0.863	0.246

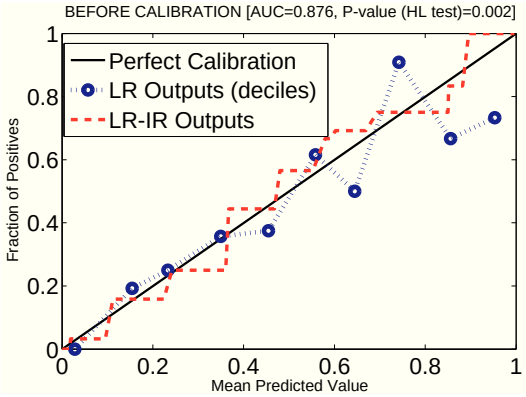
In this experiment, none of the previous methods: LR-PS and LR-IR, is capable of calibrating the raw LR outputs, as indicated in Table 7.7. AC-LR does a good job on both data and demonstrated superior *calibration* and AUC. I visually compared these approaches in Figure 7.11 and Figure 7.12 to show the difference.



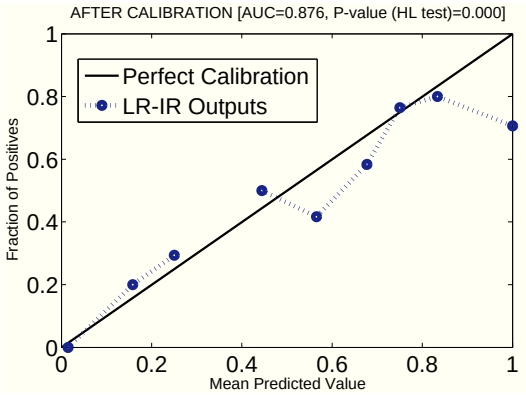
(a) LR outputs in deciles and the LR-PS *calibration* function.



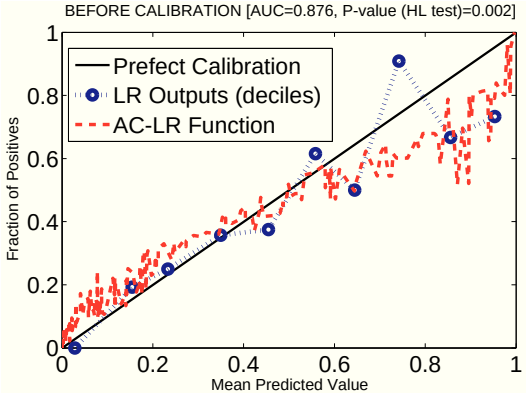
(b) LR-PS rectified outputs in deciles.



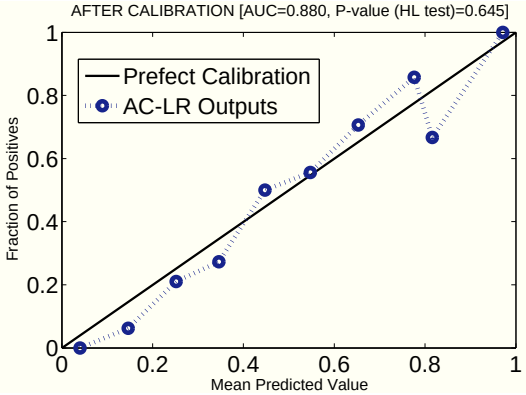
(c) LR outputs in deciles and the LR-IR *calibration* function.



(d) LR-IR rectified outputs in deciles.

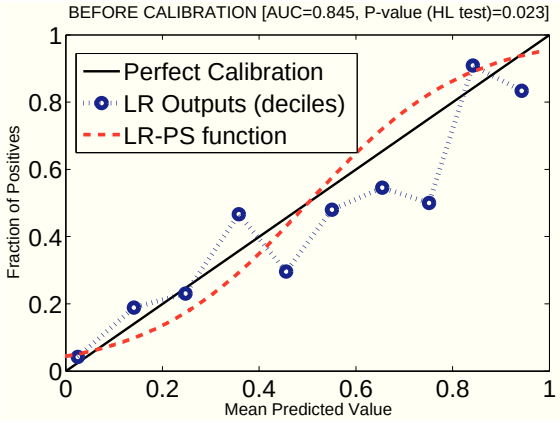


(e) LR outputs in deciles and the AC-LR *calibration* function.

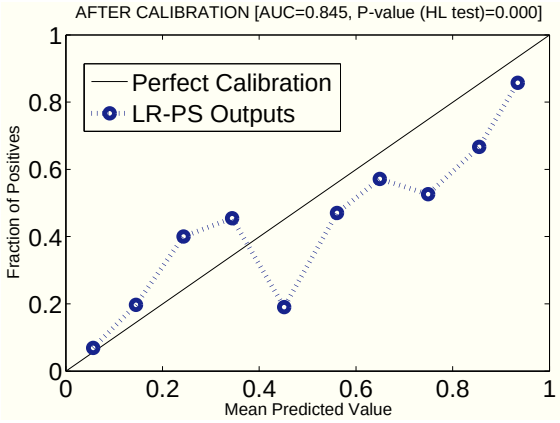


(f) AC-LR rectified outputs in deciles.

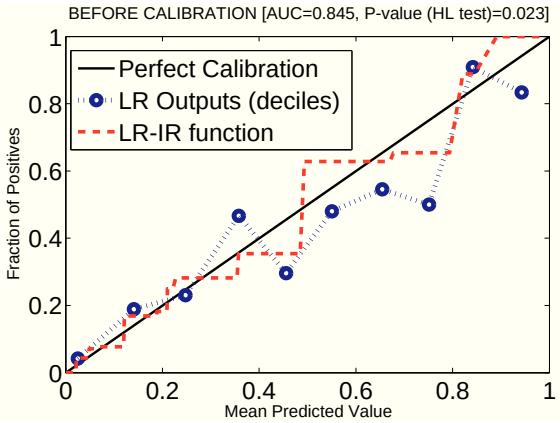
Figure 7.11: Visual results of various *calibration* approaches being applied to the Sheffield data. One can see easily that LR, LR-PS, LR-IR outputs deviate from the perfect *calibration* line. AC-LR method provides the best *calibration* yet the highest AUC among all three approaches.



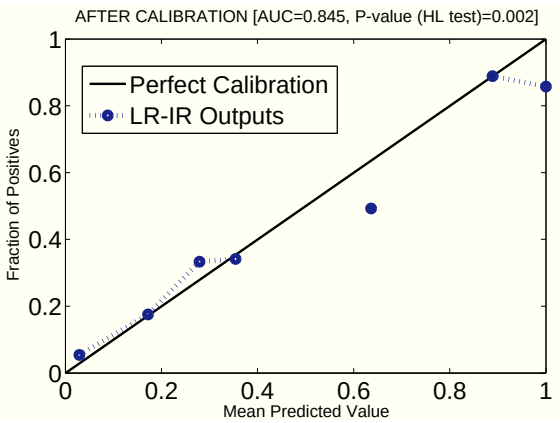
(a) LR outputs in deciles and the LR-IR *calibration* function.



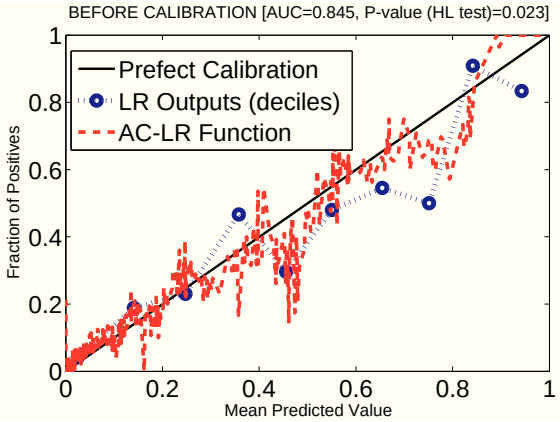
(b) LR-IR rectified outputs in deciles.



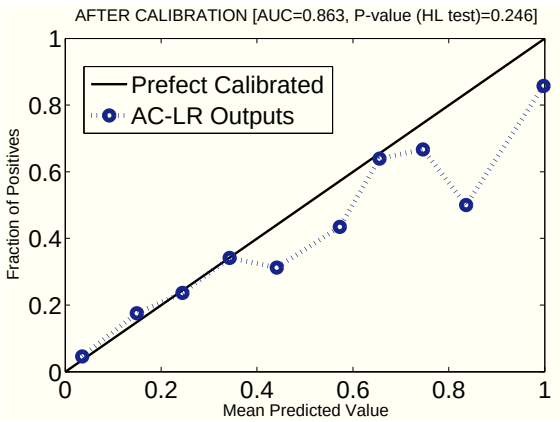
(c) LR outputs in deciles and the LR-IR *calibration* function.



(d) LR-IR rectified outputs in deciles.



(e) LR outputs in deciles and the AC-LR *calibration* function.



(f) AC-LR rectified outputs in deciles.

Figure 7.12: Visual results of various *calibration* approaches being applied to the Edinburgh data. Similar to results of the Sheffield data, the outputs of LR, LR-PS and LR-IR deviate from the perfect *calibration* line. AC-LR method provides the best *calibration* yet the highest AUC among all three approaches.

7.4.2.3 Breast Cancer Gene Expression Data

This experiment was conducted on four data sets (GSE2034, GSE2990, GSE3494A/B) obtained from the NCBI Gene Expression Omnibus (GEO). The number of samples in the data sets are: 209 for GSE2034 (114 good/95 poor), 90 for GSE2990 (60 good/30 poor), and 242 for GSE2034 (224 good/18 poor). All these data-sets have a feature size of $247,965$, which is formidable for most machine learning algorithms.

Table 7.8: Performance comparison of different models using the Breast Cancer Gene Expression Data.

(a) 5-CV result in GSE2034: AUC and HL-test of various *calibration* methods.

	AUC					P-value (HL test)				
Folders	#1	#2	#3	#4	#5	#1	#2	#3	#4	#5
LR	0.87	0.86	0.85	0.73	0.72	0.01	0.00	0.02	0.00	0.00
LR-PS	0.87	0.86	0.85	0.73	0.72	0.00	0.00	0.00	0.13	0.94
LR-IR	0.87	0.86	0.85	0.73	0.72	0.04	0.00	0.16	0.00	0.00
AC-LR	0.87	0.85	0.85	0.75	0.69	0.12	0.01	0.30	0.06	0.21

(b) 5-CV result in GSE2990: AUC and HL-test of various *calibration* methods.

	AUC					P-value (HL test)				
Folders	#1	#2	#3	#4	#5	#1	#2	#3	#4	#5
LR	0.76	0.88	0.98	0.72	0.66	0.00	0.00	0.99	0.00	0.00
LR-PS	0.76	0.88	0.98	0.72	0.66	0.99	0.74	0.00	0.36	0.52
LR-IR	0.76	0.88	0.98	0.72	0.66	0.00	0.01	0.99	0.00	0.00
AC-LR	0.83	0.87	0.99	0.74	0.72	0.07	0.30	0.46	0.06	0.08

(c) 5-CV result in GSE3494_u133A: AUC and HL-test of various methods.

	AUC					P-value (HL test)				
Folders	#1	#2	#3	#4	#5	#1	#2	#3	#4	#5
LR	1	1	1	0.97	0.95	1	1	1	0.00	0.00
LR-PS	1	1	1	0.97	0.95	0.03	0.08	0.03	0.37	0.31
LR-IR	1	1	1	0.97	0.95	1	1	1	1	1
AC-LR	1	1	1	0.93	0.98	1	1	1	0.99	1

(d) 5-CV result in GSE3494_u133B: AUC and HL-test of various methods.

	AUC					P-value (HL test)				
Folders	#1	#2	#3	#4	#5	#1	#2	#3	#4	#5
LR	1	1	0.97	0.82	1	0.00	1	0.00	0.00	0.00
LR-PS	1	1	0.97	0.82	1	0.17	0.02	0.17	0.13	0.34
LR-IR	0.97	1	0.97	0.82	1	1	1	1	1	1
AC-LR	1	1	0.97	0.83	1	0.98	1	0.98	0.85	0.96

To be compatible with other models, I apply a feature selection pre-process via Student's t-test, refer to [140] for more details. The P-value was used to select the most relevant features. I used the top 15

features from each data-set. The final evaluation is conducted on a 5-fold cross-validation to compare the AUC (*discrimination*) and the P-value of the HL test (*calibration*) on LR, LR-PS, LR-IR and AC-LR.

Table 7.8 displays the results. Note the red numbers corresponds to when AC-LR, the proposed model, achieves the best AUC among the *calibration* models, the blue numbers represents when AC-LR passes the HL test at 0.05 significant level. The proposed AC-LR outperforms the other *calibration* approaches in this data-set.

7.5 Discussion

To obtain two achievements of probabilistic models: *discrimination* and *calibration* concurrently, I broke the global monotonic constraints, but instead, calibrate probabilistic outputs adaptively. The reason for this is "true probability" of underlying events are unknown but may be estimated by class membership of similar patterns. That is, from a frequentist perspective, a reliable estimator of the "true probability" is the fraction of positive labeled cases of statistically similar cases. An important lesson I learned through my investigation is that good *calibration* should be performed locally as opposed to that good *discrimination* can be examined globally. Thus previous approaches tend to adjust probability irrelevant of inputs and model characteristics do not achieve good *calibration*.

To this end, I developed an approach that considers model-specific information to calibrate the logistic regression predictions locally. Without increasing computational complexity, my approach shows performance advantage on various synthetic. For these synthetic data experiments, I showed intuitively how AC-LR is superior to existing approaches. I visualized 1D and 2D non-linear separable cases, which be handled by AC-LR but not the others. Furthermore, the method went even beyond the capability of linear classifiers to handle linear non-separable situations thanks to its adaptive nature.

I also conducted real world experiments using Hospital Discharge Error, Myocardial Infarction and Breast Cancer Gene Expression Data. In Hospital Discharge Experiment, AC-LR passed HL-test at 0.05 significance level with a p-value of 0.349 while all the other methods failed. In addition, AC-LR even improved AUC from 0.704 (the best of other methods in comparison) to 0.717 showing joint optimization of *calibration* and *discrimination* improved single-target models in both perspectives. For the Myocardial Infarction dataset, AC-LR demonstrated its performance advantage over conventional methods again. AC-LR passed HL-test at 0.05 significance level with p-values of 0.645 and 0.246 for Sheffield data and Edin-

burgh data while LR, LR-PS and LR-IS failed. Improvements for *discrimination* are also prominent, AC-LR achieved an AUC of 0.880 and 0.863 comparing to 0.876 and 0.845 of LR, LR-PS and LR-IS for Sheffield data and Edinburgh data, respectively. Similarly, the performance of AC-LR led the competition of *discrimination* and *calibration* in the Breast Cancer Gene Expression data. Finally, I evaluated model using the Breast Cancer Gene Expression data. Results of five cross-validation of four different Gene data consistently demonstrated the performance advantage of AC-LR.

In conclusion, AC-LR is an automatic method that calibrates probabilistic outputs of Logistic Regression adaptively but its formulation can be easily extended to other machine learning algorithm. The synthetic experiments and clinical related evaluations confirmed my assertion that *calibration* should be computed locally, including only relevant information.

7.6 Conclusion

In this chapter, I developed a new *calibration* approach, Adaptive Calibration for Logistics Regression (AC-LR), a completely automatic tool which bridges a widely used learning model (Logistic Regression) to *calibration* for personalized medicine. Note that Logistic Regression was developed under theories for cohort studies but AC-LR went beyond its capacity.

The AC-LR approach is tailored and targeted to benefit individuals in more specific groups, based on more relevant information. As opposed to conventional methods constructed on the entire population of patients, my model used confidence intervals for individual predictions to construct a dynamic neighborhood for each patient. That is, the predictions are based on more relevant information about the patient. Experiments on multiple clinical data demonstrated improved *calibration* and *discrimination* ability of this new model. Yet another advantage of AC-LR is that its computational cost is much lower compared to other methods considering using dynamic neighborhood. This data-driven approach has demonstrated significant improvement to previous methods like Platt Scaling and Isotonic Regression in experiments using synthetic data and clinically related data.

The AC-LR model is ready for other cases where a probabilistic outputs is preferred over a decision rule. The model can be directly applied to calibrate probabilistic estimates of binary outcomes. Thanks to the joint consideration of input characteristics and output values, AC-LR produces more reliable probabilities of positive events than existing approaches.

Chapter 8

Data Scalability Issue

The data scalability issue is common in biomedical informatics and is often used as a touchstone for the applicability of machine learning models in the real world. The performance of a simple model can deteriorate when the amount of data increases as the number of little errors accumulate. On the other hand, a sophisticated model might retain its performance at the cost of exponential growth in computational complexity. Both situations might indicate that the particular predictive model is not appropriate for a large scale dataset even if it is theoretically sound. In addition, the affects of data scalability impacts are also useful in determining which and how many data are required to construct reliable decision support predictive models.

This chapter is dedicated to investigating the impact of data scalability on models developed in the thesis. Specifically, I used data from various sources to compare different models, including existing approaches and these developed in this thesis. At increasing ratios of training data size vs. testing data size, I evaluated these models' performance in terms of *discrimination*, *calibration* and *computational cost*. The results indicated that model developed in this thesis, i.e., Smooth Isotonic Regression, Adaptive Calibration for Logistic Regression and Temporal Maximum Margin Markov Network fit well to large datasets as their performances are superior to existing models. In addition, their computational complexity at increasing size of the data is growing at a comparable rate to the computational complexity of previous approaches.

Because models developed for large-scale disease outbreak prediction are different from those designed to improve personalized clinical risk estimation, I used different data to evaluate these two categories of models. The chapter is thus divided into two parts: the first part concentrates on evaluating the scalability of models developed for predicting a single target variable; the second part focuses on the scalability of the

multiple variable co-prediction problem.

8.1 Data

I gathered a broad range of real world data to evaluate the scalability impact on different models. Among these real world data, I used data with binary outcomes to evaluate models developed to improve personalized clinical risk estimation. The datasets included the following: HOSPITAL, MI, HEIGHT_WEIGHT, ADULT_CENSUS, BREAST_CANCER and BANKRUPTCY. Regarding the multiple variable co-estimation model I developed for large scale disease outbreak prediction, I used BioWar-II, which contains a set of correlated manifestations.

8.1.1 Hospital Discharge Error data

The HOSPITAL data set consists of microbiology cultures and other variables related to hospital discharge errors [59]. The following table defines various features and outcome variables for this data.

Table 8.1: Details of co-variates and the outcome variables in the hospital discharge error data. Eight out of ten explanatory variables are categorical and two are numerical.

Name	**Details**
	Features
Specimen:	0=blood, 1=urine, 2=sputum, 3=csf
Spec_days:	Number of days between admission date and specimen collection date.
Collect_week:	0=specimen collected on weekday, 1=specimen collected on weekend
Final_week:	0=final result on weekday, 1=final result on weekend
Vistyp:	1=admission, 0=non-admission
Svc:	0=<blank> (patient not admitted), 1=ONC, 2=MED, 3=Medical Sub-specialties, 4=Surgery and Surgical Sub-specialties, 5=Other
Age:	Age in years
Female:	0=male, 1=female
Race:	0=white, 1=black, 2=Asian, 3=Hispanic, 4=other, 5=unknown/declined
Insurance:	0=medicare, 1=medicaid, 2=commercial, 3=other
	Target Variable
Pot_error:	0=not a potential follow-up error, 1=a potential follow-up error

I also summarize features and the outcome variable by their description statistics, i.e., min, 1st Qu.,

median, 3rd Qu. , and max. The clinical meaning for each column was explained in Chapter 2.

Table 8.2: Descriptional statistics for the hospital discharge error data.

specimen	specimen days	collect week	final week	visit type	svc
0: 233	1 :1245	0:3755	0:3583	1:4818	1: 665
1:2564	0 :1030	1:1063	1:1235		2:1287
2:1467	2 : 682				3:1217
3: 554	3 : 391				4:1608
	4 : 327				5: 41
	5 : 227				
	(Other): 916				
age	female	race	insurance	pot error	
Min. : 0.00	0:2252	0:3360	0:1996	0:4449	
1st Qu.:43.28	1:2566	1: 577	1: 554	1: 369	
Median :57.76		2: 110	2:2152		
Mean :56.51		3: 405	3: 116		
3rd Qu.:71.24		4: 55			
Max. :99.71		5: 311			

There are 369 clinically important but highly suspicious observations out of 4819 returned post-discharge observations, which makes the data highly unbalanced and a challenge to calibrate.

8.1.2 Myocardial Infarction data

Table 8.3: Descriptional statistic for the Edinburgh data set.

Abbreviation				
age	min: 13.0	median:59	mean:57.6	max: 92
Smokes	0: 785	1: 468		
Exsmoker	0: 959	1: 294		
Fhistory	0: 967	1: 286		
Diebetes	0: 1165	1: 88		
BP	0: 1053	1: 200		
Lipids	0: 1215	1: 38		
CPmajorSymp	0: 62	1: 1191		
Restrostern	0: 331	1: 922		
Lchest	0: 907	1: 346		
Rchest	0: 1109	1: 144		
Back	0: 1122	1: 131		
Larm	0: 670	1: 583		
Rarm	0: 1042	1: 211		
breath	0: 1031	1: 222		
postural	0: 1017	1: 236		
Cwtender	0: 1201	1: 52		
Sharp	0: 1208	1: 45		
Tight	0: 572	1: 681		
Sweating	0: 739	1: 514		
SOB	0: 731	1: 522		
Nausea	0: 1124	1: 129		
Vomiting	0: 1124	1: 129		
Syncope	0: 1208	1: 45		
Episodic	0: 1161	1: 92		
Worsening	min: 0.0	median: 4.0	mean: 17.4	max: 168
Duration	min: 0.0	median: 3.0	mean: 8.84	max: 168
prev-ang	0: 699	1: 554		
prev-MI	0: 836	1: 361		
Worse	0: 892	1: 361		
Crackles	0: 1106	1: 147		
Added-HS	0: 1247	1: 6		
Hypoperfusion	0: 1203	1: 50		
Stelve	0: 1199	1: 54		
NewQ	0: 1240	1: 13		
STorT-abnorm	0: 1240	1: 13		
LBBBorRBBB	0: 1203	1: 50		
Old-MI	0: 1101	0: 152		
Old-isch	0: 1141	1: 112		
MI	0: 979	1: 274		

Table 8.4: Descriptional statistics for the Sheffield data set.

Abbreviation				
age	min: 17.0	median:61	mean:59.9	max: 91
Smokes	0: 318	1: 182		
Exsmoker	0: 388	1: 112		
Fhistory	0: 373	1: 127		
Diebetes	0: 451	1: 49		
BP	0: 403	1: 97		
Lipids	0: 482	1: 18		
CPmajorSymp	0: 37	1: 463		
Restrostern	0: 110	1: 390		
Lchest	0: 373	1: 127		
Rchest	0: 438	1: 62		
Back	0: 426	1: 74		
Larm	0: 237	1: 263		
Rarm	0: 418	1: 82		
breath	0: 422	1: 78		
postural	0: 455	1: 45		
Cwtender	0: 491	1: 9		
Sharp	0: 400	1: 100		
Tight	0: 246	1: 254		
Sweating	0: 235	1: 265		
SOB	0: 281	1: 219		
Nausea	0: 341	1: 159		
Vomiting	0: 449	1: 51		
Syncope	0: 467	1: 33		
Episodic	0: 417	1: 83		
Worsening	min: 0.0	median: 6.0	mean: 50.37	max: 1000
Duration	min: 0.0	median: 4.0	mean: 12.34	max: 1000
prev-ang	0: 281	1: 219		
prev-MI	0: 377	1: 123		
Worse	0: 338	1: 162		
Crackles	0: 373	1: 127		
Added-HS	0: 476	1: 24		
Hypoperfusion	0: 441	1: 59		
Stelve	0: 403	1: 97		
NewQ	0: 470	1: 30		
STorT-abnorm	0: 403	1: 97		
LBBBorRBBB	0: 474	1: 26		
Old-MI	0: 454	1: 46		
Old-isch	0: 473	1: 27		
MI	0: 346	1: 154		

The Myocardial Infarction (MI) data correspond to results of patients with and without Myocardial Infarction who were observed at emergency department in UK [98]. The data contain patient records from two medical

centers in Britain; among these, 600 patients attending at the emergency room (ER) with chest pain were observed in Sheffield, England, and $1,253$ patients with the same symptoms were observed in Edinburgh, Scotland. The following tables summarize feature and outcome variables for both Sheffield and Edinburgh data. More details about the MI data set were provided in Chapter 2.

8.1.3 Height and Weight data

The date for height and weight pertained to groups of both men and women. The subjects are 213 students of an academic university. Of these students, 73 were female and 140 were male. The data contains the following features: height, weight, GPA, left arm length, right arm length, left foot size, and right foot size.

Table 8.5: Descriptional statistics for the HEIGHT_WEIGHT dataset.

Sex	Height	Weight	GPA
0: 73	Min. :55.00	Min. : 95.0	Min. :1.240
1: 140	1st Qu.:64.00	1st Qu.:125.0	1st Qu.:2.670
	Median :67.00	Median :140.0	Median :3.000
	Mean :67.31	Mean :145.5	Mean :3.004
	3rd Qu.:70.50	3rd Qu.:160.0	3rd Qu.:3.400
	Max. :79.00	Max. :280.0	Max. :3.910
LArm	RArm	LFoot	RFoot
Min. :20.50	Min. :20.50	Min. :19.50	Min. :20.00
1st Qu.:24.00	1st Qu.:24.00	1st Qu.:23.40	1st Qu.:23.00
Median :25.00	Median :25.00	Median :24.70	Median :25.00
Mean :25.17	Mean :25.31	Mean :25.16	Mean :25.20
3rd Qu.:26.50	3rd Qu.:27.00	3rd Qu.:27.00	3rd Qu.:27.00
Max. :31.00	Max. :31.00	Max. :32.00	Max. :32.00

8.1.4 AdultCensus data

The extraction of this ADULT_CENSUS data was conducted by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted and the prediction task was to determine whether a person earns over $50K$ a year [101]. This data contained 14 explanatory variables and a binary outcome variable "income."

Table 8.6: Descriptional statistics for the ADULT_CENSUS data set.

age	workclass	fnlwgt	education	education num
Min. :17.00	Private :22696	Min. : 12285	HS-grad :10501	Min. : 1.00
1st Qu.:28.00	Self-emp-not-inc: 2541	1st Qu.: 117827	Some-college: 7291	1st Qu.: 9.00
Median :37.00	Local-gov : 2093	Median : 178356	Bachelors : 5355	Median :10.00
Mean :38.58	? : 1836	Mean : 189778	Masters : 1723	Mean :10.08
3rd Qu.:48.00	State-gov : 1298	3rd Qu.: 237051	Assoc-voc : 1382	3rd Qu.:12.00
Max. :90.00	Self-emp-inc : 1116	Max. :1484705	11th : 1175	Max. :16.00
NA	(Other) : 981		(Other) : 5134	
marital.status	occupation	relationship	race	sex
Divorced : 4443	Prof-specialty :4140	Husband :13193	Amer-Indian-Eskimo: 311	Female:10771
Married-AF : 23	Craft-repair :4099	Not-in-family : 8305	Asian-Pac-Islander: 1039	Male :21790
Married-civ :14976	Exec-managerial:4066	Other-relative: 981	Black : 3124	
Married-absent: 418	Adm-clerical :3770	Own-child : 5068	Other : 271	
Never-married :10683	Sales :3650	Unmarried : 3446	White :27816	
Separated : 1025	Other-service :3295	Wife : 1568		
Widowed : 993	(Other) :9541			
capital.gain	capital.loss	hours.per.week	native.country	income
Min. : 0	Min. : 0.0	Min. : 1.00	United-States:29170	<=50K:24720
1st Qu.: 0	1st Qu.: 0.0	1st Qu.:40.00	Mexico : 643	>50K : 7841
Median : 0	Median : 0.0	Median :40.00	? : 583	
Mean : 1078	Mean : 87.3	Mean :40.44	Philippines : 198	
3rd Qu.: 0	3rd Qu.: 0.0	3rd Qu.:45.00	Germany : 137	
Max. : 99999	Max. :4356.0	Max. :99.00	Canada : 121	
			(Other) : 1709	

8.1.5 Breast Cancer Gene Expression data

These data were obtained from the NCBI Gene Expression Omnibus (GEO). Three individual data that were downloaded were previously studied by Wang et al. (GSE2034) [185], Sotiriou et al. (GSE2990) [166], and Miller et al. (GSE3494) [128], respectively.

To make my data compatible with previous studies, I followed the criteria in [140] to select patients who did not receive any treatment and had negative lymph node status. Among these pre-selected candidates, only patients with extreme outcomes, either poor outcomes (recurrence or metastasis within five years) or very good outcomes (neither recurrence nor metastasis within eight years) were selected. The number of samples after filtering were: 209 for GSE2034 (114 good/95 poor), 90 for GSE2990 (60 good/30 poor), and 242 for GSE2034 (224 good/18 poor).

I also applied a split to divide GSE3494 into two groups, as suggested by [140], GSE3494-A and GSE3493-B, according to the sample's Affymetrix platform. Thus, the breast cancer data-set has four separate data. All these data have a feature size of 247,965, which corresponds to the gene expression results obtained from micro-array experiments.

8.1.6 Bankruptcy data

The Bankruptcy data contain two features: Return and EBIT (earnings before interest and taxes). The outcome variable "Bankruptcy" is binary. There are 66 samples in this data, where 33 samples correspond to observed bankruptcy and the others do not. The following table summarizes their description statistics, i.e., min, 1st Qu., median, 3rd Qu., and max.

Table 8.7: Descriptional statistics for the BANKRUPTCY data set.

Return	EBIT	Bankruptcy
Min. :-308.90	Min. :-280.000	0:33
1st Qu.: -39.05	1st Qu.: -17.675	1:33
Median : 7.85	Median : 4.100	
Mean : -13.63	Mean : -8.226	
3rd Qu.: 35.75	3rd Qu.: 14.400	
Max. : 68.60	Max. : 34.100	

8.1.7 BioWar II data

The BioWar II data contain multiple five-year-period observations from 9/1/2002 to 8/30/2007. The number of simulated agents are set to vary from 153,090 to 1,224,726, at an approximately equal scale ($150k$); specifically, the number of simulated agents varies from 10% (153,090) to 100% (1,224,726). The city of simulation is Norfolk, VA. There was one outbreak of airborne diseases for every year during the simulated period. The data incorporate both relational and temporal information.

In this data, the simulated agents interact and transmit airborne diseases (avian influenza) over time. There are six time ticks everyday; thus $365*6$ time ticks are observed for each year. I used BioWar simulation engine to generate ten five-year periods rather than a single 50-year period to avoid the birth and death factors to impact the disease modeling. The following table summarizes features and outcome variables of all 1,224,736 agents that are simulated. More details and motivation about this data were introduced in Chapter 2.

Table 8.8: Descriptional statistics for BioWar II.

tick	dayOfWeek	month	day	dead	is.er
Min. : 0	Fri:1560	Dec : 930	Min. : 1.00	Min. :0	Min. : 0.00
1st Qu.: 2737	Mon:1566	Jan : 930	1st Qu.: 8.00	1st Qu.:0	1st Qu.: 0.00
Median : 5474	Sat:1560	Jul : 930	Median :16.00	Median :0	Median : 36.00
Mean : 5474	Sun:1566	Mar : 930	Mean :15.72	Mean :0	Mean : 38.94
3rd Qu.: 8212	Thu:1566	May : 930	3rd Qu.:23.00	3rd Qu.:0	3rd Qu.: 49.00
Max. :10949	Tue:1566	Oct : 930	Max. :31.00	Max. :0	Max. :368.00
	Wed:1566	(Other):5370			
kidsAtHome	adultsAtHome	at.work	weblookup	medcalls	num.exchanges
Min. : 85069	Min. :154198	Min. : 0	Min. : 0.00	Min. : 0	Min. : 0.0000
1st Qu.:126708	1st Qu.:362493	1st Qu.: 0	1st Qu.: 0.00	1st Qu.: 0	1st Qu.: 0.0000
Median :316864	Median :907737	Median : 0	Median : 42.00	Median : 0	Median : 0.0000
Mean :240999	Mean :622832	Mean :175130	Mean : 45.76	Mean :102464	Mean : 0.8463
3rd Qu.:316867	3rd Qu.:907859	3rd Qu.: 0	3rd Qu.: 57.00	3rd Qu.: 0	3rd Qu.: 1.0000
Max. :316867	Max. :907859	Max. :753630	Max. :375.00	Max. :595796	Max. :46.0000
in.hospital	is.home	is.work	is.school	is.pharmacy	is.doctor
Min. : 0.00	Min. : 239387	Min. :0	Min. : 0	Min. : 0.00	Min. : 0.000
1st Qu.: 0.00	1st Qu.: 489130	1st Qu.:0	1st Qu.: 0	1st Qu.: 0.00	1st Qu.: 0.000
Median : 36.00	Median :1224600	Median :0	Median : 0	Median : 0.00	Median : 0.000
Mean : 38.94	Mean : 863832	Mean :0	Mean : 37529	Mean : 37.51	Mean : 9.815
3rd Qu.: 49.00	3rd Qu.:1224726	3rd Qu.:0	3rd Qu.: 0	3rd Qu.: 56.00	3rd Qu.: 15.000
Max. :368.00	Max. :1224726	Max. :0	Max. :231797	Max. :542.00	Max. :237.000
is.stadium	is.theater	is.store	is.restaurant	is.university	is.military
Min. : 0	Min. : 0	Min. :0	Min. : 0	Min. :0	Min. :0
1st Qu.: 0	1st Qu.: 0	1st Qu.:0	1st Qu.: 0	1st Qu.:0	1st Qu.:0
Median : 0	Median : 0	Median :0	Median : 0	Median :0	Median :0
Mean : 4988	Mean :15666	Mean :0	Mean :127495	Mean :0	Mean :0
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.:0
Max. :25269	Max. :78581	Max. :0	Max. :634041	Max. :0	Max. :0

8.2 Single Target Variable Prediction Models

In this section, I evaluate the impact of varying data size to the performance of various models, including Logistic Regression, Platt Scaling, Isotonic Regression, Smooth Isotonic Regression and Adaptive Calibrated Logistic Regression. The first three models were proposed by previous papers and the latter two are models I developed.

The data used to access the model's performances are: Breast Cancer Gene Expression Data (GSE2034, GSE2990, GSE3494), Myocardial_Infarction (Edin, Shef), Hospital_discharge and three additional UCI

machine learning repository data (Bankruptcy, PIMATR and HeightWeight) [69]. All data are randomly split into training and testing according to a ratio factor that varies from 0.3 to 0.7.

8.2.1 Discrimination and Computational Cost

To evaluate the performance at increasing amount of training data more systematically, I sample the data of the ten training/testing ratios (from 0.1 to 0.9) for 60 time and demonstrate the model's *discrimination* v.s. training data size and computational cost v.s. training data size use all nine data sets. The models accessed are Logistic Regression reliability diagram (LR), Platt Scaling reliability diagram (PLATT), Isotonic Regression reliability diagram (IR), Smooth Isotonic Regression reliability diagram (SIR) and Adaptive Calibrated Logistic Regression reliability diagram (ACLR), respectively.

For each of the following figures, the AUC and time cost (in seconds) are computed as the average of the 60 random experiments.

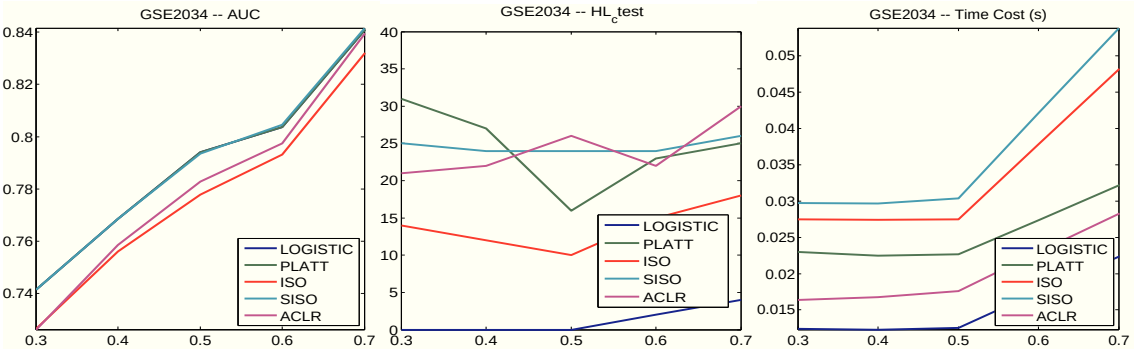


Figure 8.1: Scalability performance evaluation for GSE2034. Left: *Discrimination* vs. increased training/testing data ratio, Middle: How many times the model's outputs pass a HL-test out of the 60 random experiments, Right: Computational cost vs. increased training/testing data ratio.

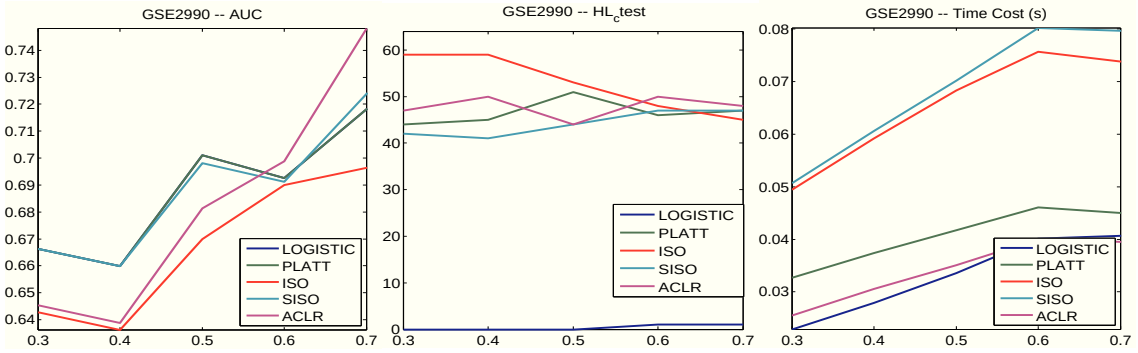


Figure 8.2: Scalability performance evaluation for GSE2990. Left: *Discrimination* vs. increased training/testing data ratio, Middle: How many times the model's outputs pass a HL-test out of the 60 random experiments, Right: Computational cost vs. increased training/testing data ratio.

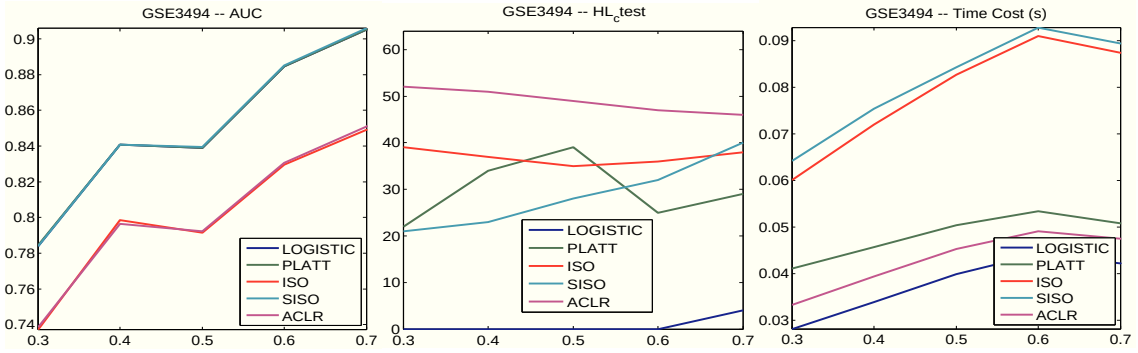


Figure 8.3: Scalability performance evaluation for GSE3494. Left: *Discrimination* vs. increased training/testing data ratio, Middle: How many times the model's outputs pass a HL-test out of the 60 random experiments, Right: Computational cost vs. increased training/testing data ratio.

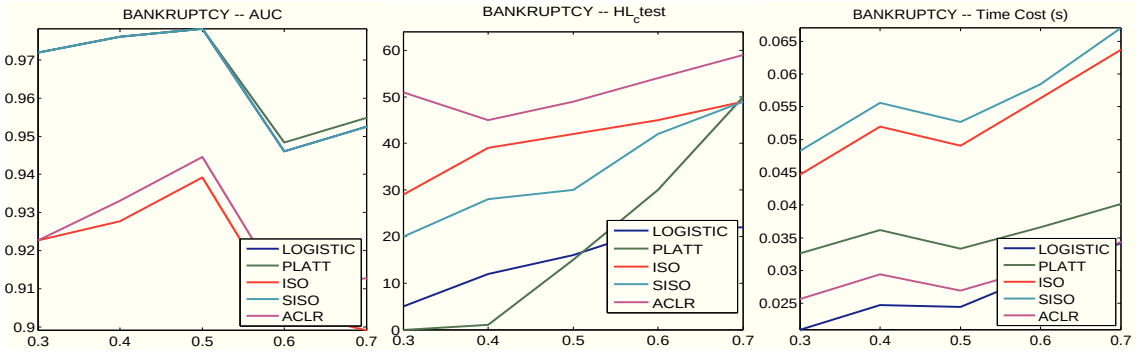


Figure 8.4: Scalability performance evaluation for BANKRUPTCY. Left: *Discrimination* vs. increased training/testing data ratio, Middle: How many times the model's outputs pass a HL-test out of the 60 random experiments, Right: Computational cost vs. increased training/testing data ratio.

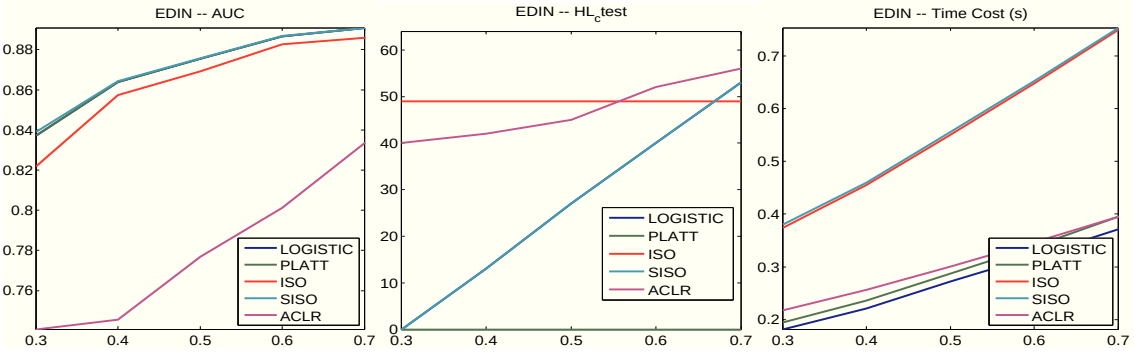


Figure 8.5: Scalability performance evaluation for EDINBURGH. Left: *Discrimination* vs. increased training/testing data ratio, Middle: How many times the model's outputs pass a HL-test out of the 60 random experiments, Right: Computational cost vs. increased training/testing data ratio.

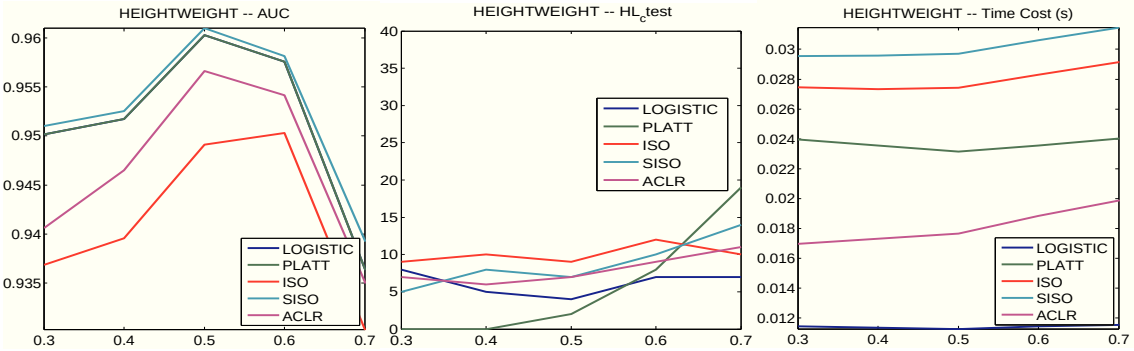


Figure 8.6: Scalability performance evaluation for HEIGHTWEIGHT. Left: *Discrimination* vs. increased training/testing data ratio, Middle: How many times the model's outputs pass a HL-test out of the 60 random experiments, Right: Computational cost vs. increased training/testing data ratio.

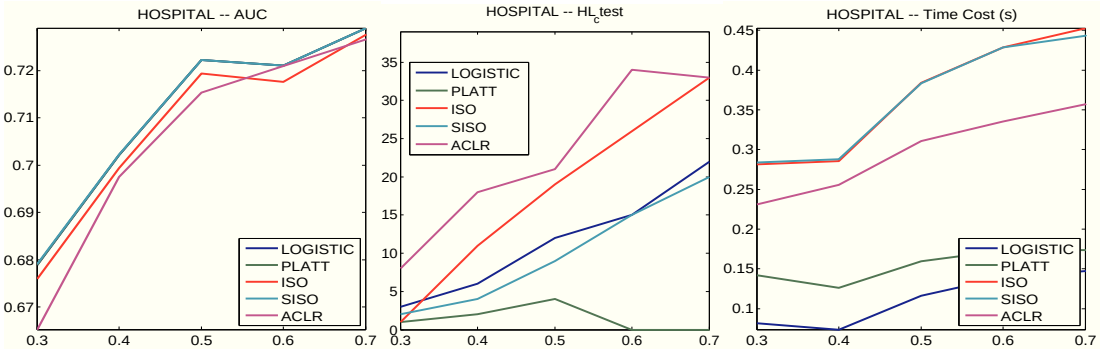


Figure 8.7: Scalability performance evaluation for HOSPITAL_DISCHARGE. Left: *Discrimination* vs. increased training/testing data ratio, Middle: How many times the model's outputs pass a HL-test out of the 60 random experiments, Right: Computational cost vs. increased training/testing data ratio .

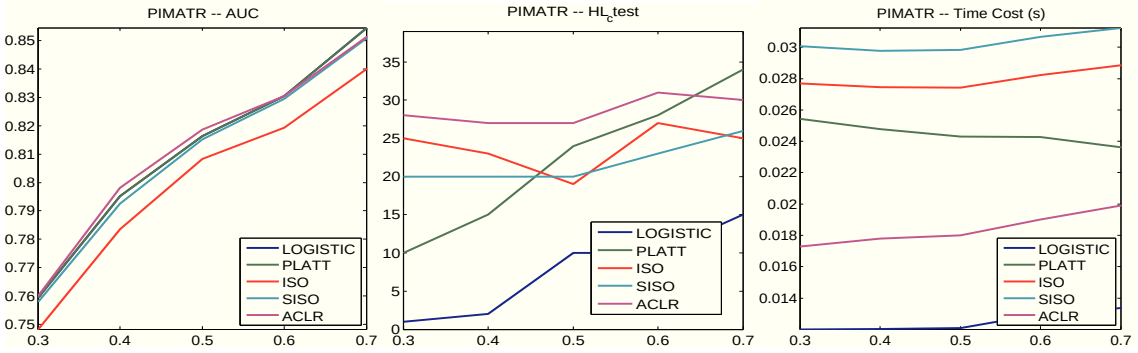


Figure 8.8: Scalability performance evaluation for PIMATR. Left: *Discrimination* vs. increased training/testing data ratio, Middle: How many times the model's outputs pass a HL-test out of the 60 random experiments, Right: Computational cost vs. increased training/testing data ratio .

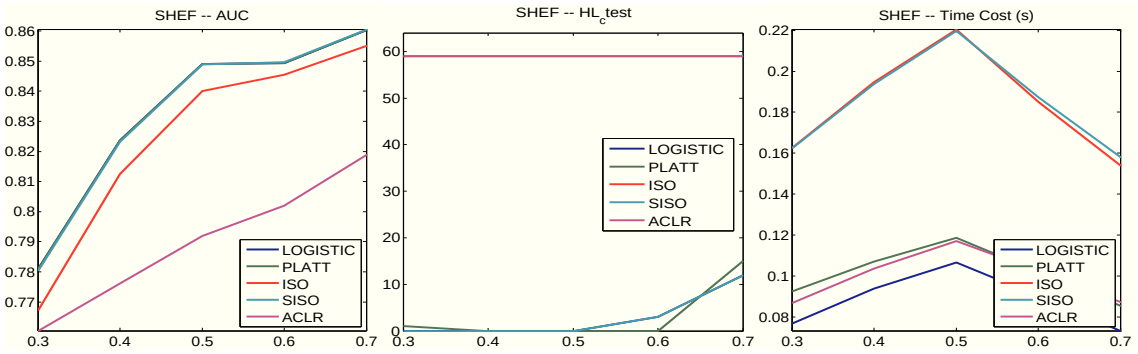


Figure 8.9: Scalability performance evaluation for SHEFFIELD. Left: *Discrimination* vs. increased training/testing data ratio, Middle: How many times the model's outputs pass a HL-test out of the 60 random experiments, Right: Computational cost vs. increased training/testing data ratio.

Figures from 8.1 to 8.9 demonstrated models' *discrimination* and *calibration* at increasing amount of training data size. In most cases, the AUC starts at a low value (due to under-fit) and keeps increasing until a saddle point is hit, and then the performance goes down or stay flat. The reason of such decreasing in *discrimination* is due to over-fitting the training data. A reasonable trade-off between performance and computational cost lies in the training/testing ratio 6:4, where the performance tends to stabilize and the computational cost has not increased tremendously.

In general, AC-LR and SISO demonstrated superior performance in term of *calibration*. For most cases, these two approaches take the lead in the total number of models that pass the HL-test. The last column demonstrated the computational cost of these models, indicating that SISO did not increase the computational cost exponentially, and AC-LR is often faster than ISO and Platt approaches.

8.3 Multiple Target Variable Co-Estimation Models

Next, I evaluate the impact of data scale to structured models that predict multiple variables. I include following models: Hidden Markov Model (HMM), Conditional Random Field (CRF), Maximum Margin Markov Network (M3N), Temporal Maximum Margin Markov Network (TM3N) for performance assessment.

All four models are evaluated at an increasing amount of simulated agents, generated by the BioWar simulation engine. The simulation data are multiple five-year-period observations from 9/1/2002 to 8/30/2007. The number of simulated agents are set to vary from 153,090 to 1,224,726 at approximately equal scales ($150k$), specifically, the number of simulated agents varies from 10% to 100% . Please refer to Chapter 2 for details of the variables.

Figure 8.10 illustrates the accuracy of HMM, CRF, M3N and TM3N. All four models stabilize around the scale of 80%. TM3N's accuracy increases 10%, which outperforms HMM(6%), CRF(2%) and M3N(4%).

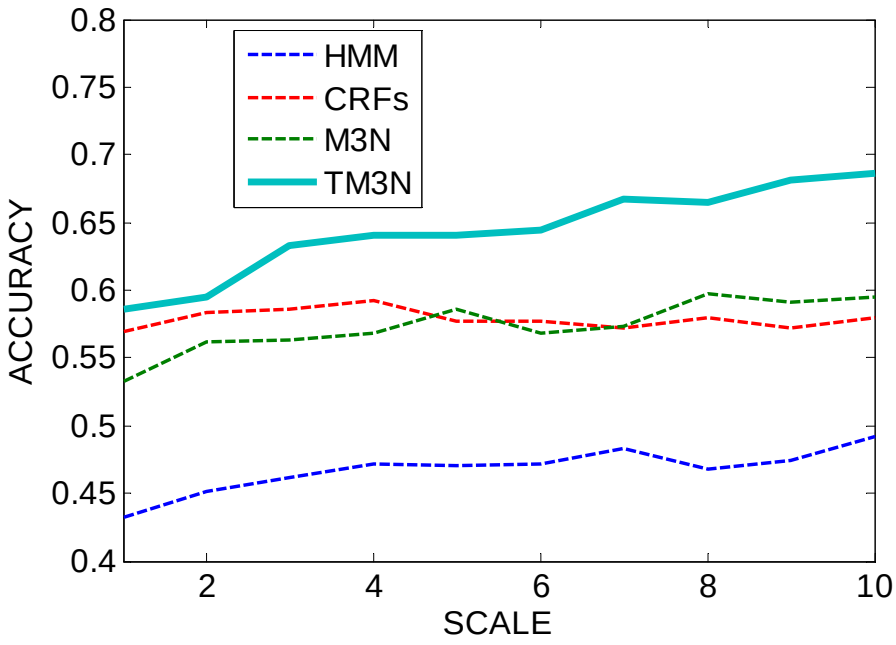


Figure 8.10: Model comparison using BioWar-II results. The X axis corresponds to the scale of agents at ten different levels (10%-100%); and Y axis represents the accuracy. The HMM is not able to capture the feature correlations, which performs poorly due to the ignorance of the context. Both M3N and CRF have fairly good generalization at increasing scales. TM3N model's accuracy begins with 0.5865 at the 10% scale, which is quite close to 0.5701 and 0.5431 of CRF and M3N, respectively. But TM3N's performance increases much rapidly than the other models in comparison with more observations.

Interpretation: The scalability experiment on TM3N method using BioWar-II data confirmed model's ability to handle large amount of training data. The results indicated TM3N led the averaged accuracy at various scales of training data, and the model scaled well.

8.4 Discussion

The data scalability issue is critical to biomedical informatics, and is often used as a touchstone for evaluating the applicability of machine learning models in the real worlds. There is a common dilemma for performance deterioration and computational complexity. That is, the performance of simple model can deteriorate when the size of data grow up as the little errors due to assumption violations are accumulated quickly. On the other hand, a sophisticated model might retain its performance at the cost of exponential growth in computational complexity. Both situations might indicate that the particular predictive model is not appropriate for large scale dataset even if it is theoretically sound. In addition, the data scalability impacts are also useful in determining which and how many data are required to construct reliable decision support predictive models.

This chapter investigated the impact of data scalability on models developed in the thesis. I used data from different sources to compare different models, including existing approaches and methods developed in this thesis. At increasing ratios of training data to testing data, I evaluated these models' performance in terms of *discrimination*, *calibration* and *computational cost*. The results indicated that the models developed in this thesis, i.e., Smooth Isotonic Regression (SIR), Adaptive Calibration for Logistic Regression (AC-LR) and Temporal Maximum Margin Markov Network (TM3N) fits well to large datasets.

Specifically, SIR and AC-LR demonstrated better *calibration* ability comparing to the other methods without sacrificing their *discrimination* ability. In terms of computational cost, all methods are scalable but SIR is slightly more expensive because it inserted an additional smoothing procedure LR-IR approach. The second and third most expensive methods are LR-IR and LR-PS, respectively. Interestingly, the most balanced method, AC-LR turned out to be the least computational expensive. TM3N, on the other hand, was evaluated on simulated BioWar-II data with an increasing amount of training data. The model demonstrated superior performance (averaged prediction accuracy) to existing models at different scales of simulated agents. Furthermore, its computational complexity at an increasing amount of the data grows at a comparable rate to the computational complexity of other state-of-the-art approaches.

8.5 Conclusion

We all know that data scalability could have a significant impact on the reliability and feasibility to machine learning methods developed for solving specific problems. Because my purpose is to provide generalized prediction methods for clinical decision support, I decided to verify such impacts to models developed in this thesis. Although the available data do not support exhaustive study of various aspects of biomedical research, I intended to make my evaluation comprehensive. To this end, I included data from different sources to compare models, including existing approaches and methods developed in this thesis. At increasing ratios of training data against testing data, I evaluated these models' performance in terms of *discrimination*, *calibration* and *computational cost*. The results indicated that the models developed in this thesis, i.e., Smooth Isotonic Regression (SIR), Adaptive Calibration for Logistic Regression (AC-LR) and Temporal Maximum Margin Markov Network (TM3N) scale well to large datasets.

Chapter 9

Data Unbalance Issue

I discussed the data scalability issue to prediction in biomedical learning. This issue is not unique to biomedical research. Thus most methods, including those developed earlier, demonstrated reasonable addictiveness in increasing amounts of the data.

However, another difficult issue in biomedical learning is class imbalance[21, 44, 195, 211, 212], which is not common in other machine learning tasks. In typical biomedical applications, e.g., classification and knowledge discovery in protein databases [153], it is hard to obtain enough labels of observations to train a reliable probabilistic model. The labeling procedure requires the experts' knowledge, staff time, and even expensive laboratory test. Oftentimes, the care providers confirm a tiny fraction of the highly suspicious observations (positive) and leave a large amount of observations unlabeled. The lack of negative labeled samples makes it difficult to apply traditional supervised learning approaches. Usually, they require a large amount of dichotomy training samples. On one hand, there is a very limited amount of usable labeled data that supervised learning algorithms can use. On the other hand, the unlabeled samples are many but cannot be leveraged by these supervised learning algorithms.

To handle a situation of a tiny amount of positive labeled data and a large amount of unlabeled data, I developed a method, called Structured Biased Support Vector Machine (SB-SVM), by considering: 1) the sample unbalance between labeled and unlabeled observations; 2) the feature correlation embedded in unlabeled observations. This additional information offer more granularity to model data faithfully. To show the usefulness of my model, I evaluate its performance with simulation data and a real clinical data for hospital discharge errors prediction. Both experiments demonstrate the advantage of the method.

9.1 Data

9.1.1 Hospital Discharge Error Data

To verify my model, I used HOSPITAL data, which was introduced earlier, and is highly unbalanced. The data consist of microbiology cultures and other variables related to hospital discharge errors [59]. There are 369 clinically important but highly suspicious observations out of 4819 returned post-discharge observations. The following table defines various features and outcome variables for this data.

Table 9.1: Details of co-variables and the outcome variable in the hospital discharge error data. Eight out of ten explanatory variables are categorical and two are numerical.

Name	Details
	Features
Specimen:	0=blood, 1=urine, 2=sputum, 3=csf
Spec_days:	Number of days between admission date and specimen collection date.
Collect_week:	0=specimen collected on weekday, 1=specimen collected on weekend
Final_week:	0=final result on weekday, 1=final result on weekend
Vistyp:	1=admission, 0=non-admission
Svc:	0=<blank> (patient not admitted), 1=ONC, 2=MED, 3=Medical Sub-specialties, 4=Surgery and Surgical Sub-specialties, 5=Other
Age:	Age in years
Female:	0=male, 1=female
Race:	0=white, 1=black, 2=Asian, 3=Hispanic, 4=other, 5=unknown/declined
Insurance:	0=medicare, 1=medicaid, 2=commercial, 3=other
	Target Variable
Pot_error:	0=not a potential follow-up error, 1=a potential follow-up error

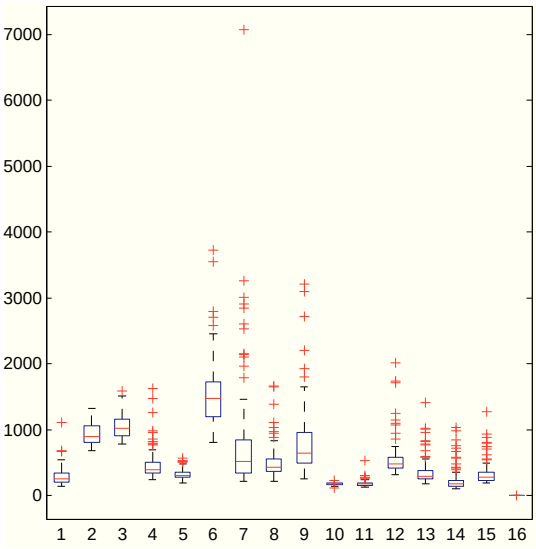
I also summarize features and the outcome variable by their description statistics, i.e., min, 1st Qu., median, 3rd Qu., and max. The clinical meaning for each column was explained in Chapter 2.

Table 9.2: Descriptional statistics for the hospital discharge error dataset.

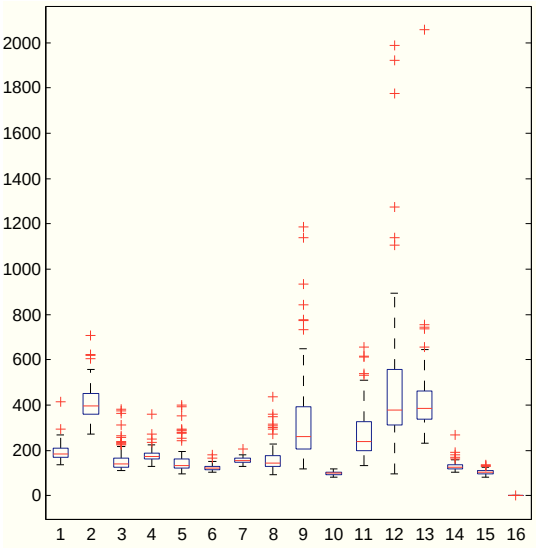
spec	spec dayssinceadm	collect we	final we	vistype	svc
0.161806	Min. : 0.000	2.607639	2.488194	3.3875	0.503472
1.822222	1st Qu.: 1.000	0.779861	0.899306		0.977083
1.102083	Median : 2.000				0.970139
0.509722	Mean : 4.355				1.283333
	3rd Qu.: 4.000				5:41
	Max. :195.000				
age	female	race	insurance	pot error	
Min. : 0.00	1.563889	2.333333	1.386111	3.089583	
1st Qu.:43.28	1.823611	0.442361	0.426389	0.297917	
Median :57.76		0.159722	1.577778		
Mean :56.51		0.40625	0.205556		
3rd Qu.:71.24		4:55			
Max. :99.71		0.424306			

9.1.2 Breast Cancer Gene Expression Data

The second data I used is the Breast Cancer Gene Expression (GSE3494), which was obtained from the NCBI Gene Expression Omnibus (GEO) and studied by Miller et al. [128]. To make my data comparable to previous studies, I followed the criteria in [140] to select patients who did not receive any treatment and had negative lymph node status. Among these pre-selected candidates, only patients with extreme outcomes, either poor outcomes (recurrence or metastasis within five years) or good outcomes (neither recurrence nor metastasis within eight years) are selected. The final data is highly biased because there are 224 good and 18 poor out of a total 242 samples. The data have a feature size of 247,965, which corresponds to the gene expression results obtained from micro-array experiments. They were preprocessed to keep only the top 15 features ranked using t-test (see [140] for details). Figure 9.1 shows boxplots of these selected gene features. It can be observed in the figure below that effective gene features are different from each other in different population groups.



(c) GSE3494_U133A



(d) GSE3494_U133B

Figure 9.1: Boxplots of Breast Cancer Gene Expression Data. Each column corresponds to one feature vector and the last column indicates the outcome variable.

9.2 Motivation

Probabilistic models are important to many tasks in biomedical informatics. Early detection and accurate prediction help decision makers to interpret observations quickly and come up with more informed responses to potential adverse events. However, a major difficulty lies in the fact that traditional machine learning approaches require a reasonable number of labels from both positive and negative observations, which are often hard to obtain. Because the labeling of observations requires experts' knowledge, staff time, and expensive laboratory test. Oftentimes, the care providers label a small fraction of highly suspicious observations and leave the rest unlabeled.

My own motivation to investigate this problem comes from a clinical decision support problem. In one type of the diagnosis error, failure to follow-up on test results in a timely fashion can lead to significant delays in diagnosis and treatment, and may causes patient morbidity and mortality. In health-care settings with well-developed computerized information systems, a post-discharge test result follow-up process could be automated and tracked. However, many hospitals and clinics in the U.S. have not adopted electronic health records. While laboratory systems can often identify culture results that show the growth of an organism,

these results are usually associated with the ordering provider. By the time the results return, the provider responsible for the patient may have changed or the patient may have left the hospital. In these settings, the test results pending at the time of discharge from the hospital often need to be followed up manually. Typically, this responsibility is placed on the individual provider or team that cared for the patient in the hospital. By assigning staff to track these results, adverse events could be avoided. The cost of hospital staff required to follow-up post-discharge results could be reduced by allowing them to focus on high-risk results estimated by some effective predictive model. The difficulty is exactly what I have discussed earlier: only a small amount of positive observations are confirmed while most are left unlabeled.

For both observations, the dilemma for traditional supervised learning approaches is the tiny number of positive-labeled observations are not sufficient to train a model but the abundant number of unlabeled observations cannot be effectively utilized. To face these challenges, I suggest a novel approach, Structured Biased Support Vector Machine (BSVM), to exact additional information from different perspectives: 1) the sample unbalance factor between labeled and unlabeled observations; 2) the feature correlation from unlabeled observations. This information, which offers additional granularity to model the data faithfully, is not considered in earlier works.

9.3 Previous Work

Traditional approaches to dealing with class-imbalanced biomedical data use "under-sampling" or "over-sampling" of the majority class or the minority class, respectively [7, 40, 195, 198]. These methods are based on assumptions about cohort studies that testing data follow the same distribution as the observed data. However, their performance on individual tests, which depends on the randomness of sampling, are not consistent. Recently, more sophisticated methods [8, 88, 100, 171, 174] have been developed to tackle this challenge, but the problem remains to be an open question.

The problem that is the focus of this chapter is slightly different from the typical class-imbalanced challenge where training data contains both positive and negative labels. In this problem, the available data for training are an incomplete subset of positive observations and a set of unlabeled observations, which makes the problem more challenging and less studied. Among the few studies that exist, there are generally two categories of models that aim to address this problem.

The first category is is comprised of one-class models that do not consider the unlabeled data [73, 84, 91,

120, 121]. A representative model is the one-class Support Vector Machine (OC-SVM), which aims to find the smallest possible kernel ball that encompasses all the positive observations but does not overfit them. The trade-off factor here is the radius of the ball and the mistake that is made in excluding observations outside it. These models cannot always convince the users because the models completely ignore the abundant number of unlabeled observations in training.

The second category of models considers both positive and unlabeled observations. An early approach proposed by Yu et al. [200] built Support Vector Machines (SVMs) with smaller subsets of observations at each iteration with the expectation of converging to some "nature boundary" between labeled positive observations and the “hidden” negative observations. This heuristic approach could be very slow and often ends up with local optimal. Another study by Liu and his colleagues [118] applied unbalanced weights to positive and unlabeled observations under a maximum margin framework. This approach doubly penalized a SVM and demonstrated good empirical performance. A more recently paper [61] discussed an alternative Bayesian approach to learning a classifier from only positive and unlabeled observations.

Unfortunately, all these methods concentrated on the very limited positive labels and failed to exploit the structural information embedded in the abundant unlabeled data. I suggested a disciplined approach to learn transferable knowledge (feature correlations) from unlabeled observations to build a more comprehensive model.

9.4 Methodology

I first overview the Biased Support Vector Machine (BSVM), an approach closely related to ours, followed by detailed discussion of my method.

9.4.1 Biased Support Vector Machine

A natural way of representing a maximum margin classification optimization led to the following format:

$$\min \ \frac{1}{2}||W||_2^2 + C\sum_i L(sign(W^T X_i + b), Y_i) \tag{9.1}$$

Here $L(sign(W^T X_i + b), Y_i)$ is the $0/1$ loss while $||W||_2^2$ is the penalty term specifies the maximum margin between two classes of data. C tradeoffs between model bias and variance. Unfortunately, Equation 9.1 is

not convex. A relaxation of the problem leads to the using of a Hinge loss function to approximate the $0/1$ loss. This result is known as the famous norm-2 SVM. It is equivalent to fitting a model that:

$$\min \frac{1}{2}||W||^2 + C\sum_i \epsilon_i \tag{9.2}$$
$$s.t. Y_i(W^T X_i + b) \geq 1 - \epsilon_i, \epsilon_i \geq 0, \forall i$$

Reorganizing the above equation by plugging in the constraints to the objective function, we get the following,

$$\min \frac{||W||^2}{2} + C\sum_i \max(0, Y^i(W^T X_i + b)), \tag{9.3}$$

where $\max(0, 1 - X)$ is known as the hinge loss function and $||W||^2$ is called ridge penalty function. The function $f(X) = W^T X + b$ represents a linear decision boundary, which maximizes the margin between positive and negative cases. Figure 9.2 illustrates the separating hyperplane and the idea of maximum margin optimization.

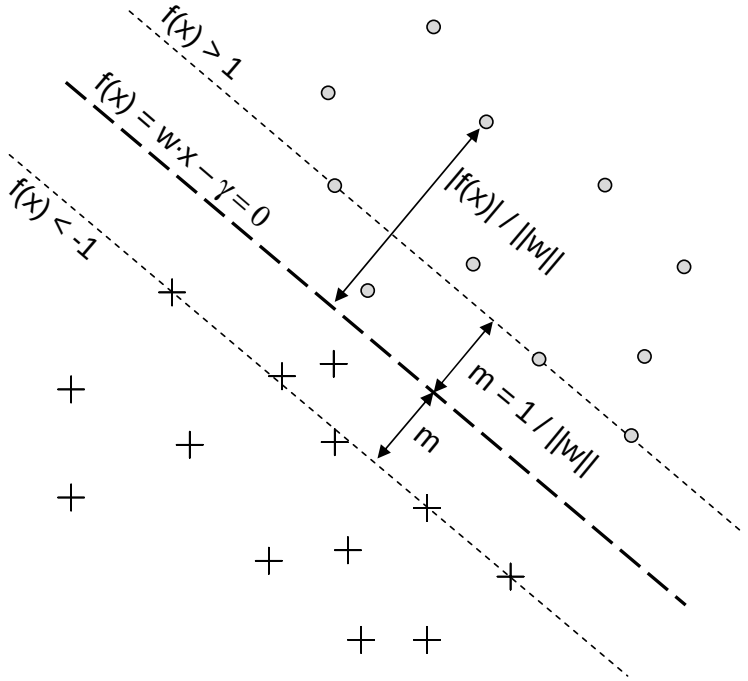


Figure 9.2: The separating hyperplane that maximizes the margin. ("o" is a positive data point, i.e. $f("o") > 0$, and "+" is a negative data point, i.e. $f(" + ") < 0$.

The ridge penalty term controls the smoothness or the complexity of the model while the loss function minimizes the bias. It is well known that the ridge penalty has the effect of controlling the variance of W [213]. In its dual format in Equation 9.4, the model shrinks the fitted coefficients α_i towards zero and uses only a small portion of the given data to construct margin hyper-planes.

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j c_i c_j X_i^T X_j, \tag{9.4}$$
$$s.t. \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i c_i = 0,$$

Computational learning theory and statistical function estimation have associated good performances of SVM in low data dimension p and large data size n to its marginal maximization property. However, many data have $p \gg n$ and the features embed sparsity, e.g., genetic data often have hundreds and thousands of genetic markers (features) investigated on a small number of patients. The authors of [118] suggested a weighted soft-margin Support Vector Machine model and reported superior performance than other heuristic models in identifying negative examples. The model looks like:

$$\min \frac{1}{2}||W||^2 + C_1 \sum_{i \in P} \epsilon_i + C_2 \sum_{i \in U} \epsilon_i \tag{9.5}$$
$$s.t.\ y_i(W^T X_i + b) \geq 1 - \epsilon_i,$$
$$\epsilon_i \geq 0, \forall i,$$

where each sample i is associated with a p-dimensional feature $X_i = [x_1, x_2, \ldots, x_p]^T$, a binary class label y_i and a hinge loss $\epsilon_i \geq 0$. P stands for the positive observations and U represents the unlabeled observations; C_1 and C_2 represent the penalty parameters for positive and unlabeled observations, respectively. W corresponds to the weights of the feature X_i and b is a constant factor.

> **Interpretation:** The doubly penalized Support Vector Machine applies different penalties to positive labeled observations and unlabeled observations so that important expert labeled data are treated more favorably.

To make losses on P be penalized more heavily than losses on U, C_1 is usually set much higher than C_2. The value of these parameters is usually determined by cross validation. This method, called Biased Support

Vector Machine (BSVM), demonstrates the state-of-the-art performance for learning from only positive and unlabeled documents. Thanks to the maximum margin criteria, it can be easily kernelized to introduce non-linearity. My model extends BSVM and incorporates additional granularity for improved performance.

9.4.2 Structured Biased Support Vector Machine

9.4.2.1 The Model

My approach extends the Bias Support Vector Machine framework of classification:

$$\text{argmin}_{W,b}\ C_1 \sum_{i \in P} L(y_i, W^T X_i + b)) + C_2 \sum_{i \in U} L(y_i, W^T X_i + b)) + W^T \Sigma_s^{-1} W, \tag{9.6}$$

where $L(x, y) = 1 - xy$ indicates a hinge loss function. Like BSVM, the confirmed positive observations are penalized much heavier than the unlabeled observations at misclassificaiton.

> **Interpretation:** I utilized the structural information implied in unlabeled data to enhance the quality of classification model based on limited data.

The difference between the Structured Biased Support Vector Machine (SB-SVM) and BSVM is in the last term of the above equation $W^T \Sigma_s^{-1} W$, where an additional matrix Σ_s is included. By introducing Σ_s, the feature correlations estimated from unlabeled observations, I intend to regularize the SB-SVM model with an imposed prior. Intuitively, correlated features should receive weights in a similar way. With the guidance of a prior, labeled observations contribute more informed to the decision boundary and thus less of them would be sufficient. These feature correlations, as useful information embedded in abundant unlabeled observations, are neglected in the previous research.

If I let Σ_s be an identity matrix $\mathbf{I}$, the term $W^T \mathbf{I}^{-1} W = W^T W$ corresponds to a L_2 norm regularization, which is default to the BSVM. In this case, Equation 9.6 is equivalent to Equation 9.5. From a Bayesian perspective, I can interpret $\Sigma_s^{-1} = \mathbf{I}^{-1}$ as imposing a Gaussian prior with zero means and an identity correlation matrix. In other word, the identity correlation matrix indicates not sufficient knowledge or unreluctant to explore the feature space information.

By some simple algebra, I can project Equation 9.6 into the standard BSVM formulation and solve the following equation accordingly.

$$\text{argmin}_{\tilde{W},b}\ C_1 \sum_{i \in P} L(y_i^l, \tilde{W}^T \tilde{X}_i + b) + \sum_{i \in U} L(y_i^l, \tilde{W}^T \tilde{X}_i + b) + \tilde{W}^T \tilde{W}, \tag{9.7}$$

where $W = \Sigma^{\frac{1}{2}} \tilde{W}$ and $X_i = \Sigma^{-\frac{1}{2}} \tilde{X}_i$.

> **Interpretation:** The variable substitution makes the structured biased Support Vector Machine compatible with existing biased Support Vector Machine solvers, and can be optimized without additional computational overhead.

I can thus transfer my estimated feature correlations computed offline to regulate the classification task in an efficient and scalable manner. Next, I introduce how to estimate the feature correlations from unlabeled observations.

9.4.2.2 Learning Structural Feature Correlations

Estimating semantic feature correlation as an alternative way to gain from unlabeled observations is proposed by Zhang et al. [210]. I extend this idea to handle training data that contains only positive and unlabeled Data.

Dimensions reduction methods such as principle components analysis [162] or latent topic models [22] extract a higher level summary (lower dimension "topics") from the raw feature through a mapping process as follows:

$$X = \mathbf{A}Z, \tag{9.8}$$

where $X = [x_1, x_2, \ldots, x_p]^T$ is the p dimensional feature vector of an unlabeled observation, $Z = [z_1, z_2, \ldots, z_k]$ is the k dimensional topics. $\mathbf{A}$ is a $p \times k$ matrix, representing the latent structure that projects "topics" onto the raw feature.

Equation 9.8 describes how a vector of raw feature X can be represented by a latent k-dimensional topic distribution and a distribution of p features in k latent topics. Different observations have different topic distributions Z, thus different feature distributions. But $\mathbf{A}$ is invariant across various observations. I can consider this $\mathbf{A}$ as the semantic implication that is transferable. Each column vector of $\mathbf{A}$ has an observation in the p dimensional feature space, corresponding to the semantic roles of the features in this topic. Given a large number of $|k|$ observations, I can build the following semantic covariance,

$$
\begin{aligned}
cov_s(X_i, X_j) &= \frac{1}{k}\sum_{t=1}^{k}(\mathbf{a}_{it} - \bar{\mathbf{a}}_{(i,)})(\mathbf{a}_{jt} - \bar{\mathbf{a}}_{(i,)}) \\
&= \frac{1}{k}\sum_{t=1}^{k}\mathbf{a}_{it}\mathbf{a}_{it} - \bar{\mathbf{a}}_{(i,)}\bar{\mathbf{a}}_{(j,)}, \qquad (9.9)
\end{aligned}
$$

where $\mathbf{a}_{it}$ represents an element of matrix$\mathbf{A}$, while $\bar{\mathbf{a}}_{(i,)}$ represents the average of the ith row in $\mathbf{A}$.

> **Interpretation:** Semantic covariances are consistent and are transferable across datasets.

I can calculate the semantic correlations as:

$$
corr_s(X_i, X_j) = \frac{cov_s(X_i, X_j)}{\sqrt{cov_s(X_{i,}X_i)cov_s(X_j, X_j)}}
$$

Consider a set of n unlabeled observations $\mathcal{D}_u = \{X_i \in \mathbf{X}, i = 1, \ldots, n\}$, I can thus learn the transferable knowledge (semantic feature correlations) from the unlabeled observations with Algorithm 5 to give SB-SVM a good prior.

Algorithm 5 Learning semantic feature correlations through bootstrap.

Input: Unlabeled observations $\mathcal{D}_u$, latent variable model PCA
Output: Semantic feature correlation matrix Σ_s
Parameters: α: number of samples at each iteration, k: dimension of hidden topics, N: an scale factor.

1: $\mathbf{V} \leftarrow \emptyset$
2: **repeat**
3: $\quad D_{samp} \leftarrow Sampling(\mathcal{D}_u, \alpha)$
4: $\quad \{(Z_1, \mathbf{a}_{(,1)}), (Z_2, \mathbf{a}_{(,2)}), \ldots, (Z_k, \mathbf{a}_{(,k)})\} \leftarrow PCA(k, D_{samp})$
5: $\quad \mathbf{V} \leftarrow \mathbf{V} \cup \{\mathbf{a}_{(,1)}, \mathbf{a}_{(,2)}, \ldots, \mathbf{a}_{(,k)}\}$
6: **until** $|\mathbf{V}| \geq kN$
7: $\Sigma_s : \Sigma_s(i, j) \leftarrow corr_s(V_i, V_j)$

In Algorithm 5, I estimated the semantic feature correlations using the principle component analysis (PCA) by extracting a large number (kN) of eigenvectors. However, obtaining a large set of diverse and meaningful eigenvectors is hard since the number of representative eigenvectors for the entire dataset are usually small. As a re-sampling technique, I used bootstrap to estimate reliable semantic feature correlations. The algorithm takes unlabeled observations $\mathcal{D}_u$ as inputs and outputs a correlation matrix Σ_s. I need to set three parameters α, K, N, correspond to the sampling rate, dimension of hidden topics and the number of iterations, respectively. The model repeats N times until large enough number of eigenvectors $\mathbf{V}$ is

collected. Finally, I calculated the correlation between these eigenvectors (V_i, V_j) to fill the semantic feature correlation matrix $\Sigma_s(i, j)$.

9.5 Experiments

I evaluate the performance of the my model on both synthetic data and real clinical data. Unfortunately, I cannot use widely accepted F score for the measurement because $F = \frac{2pr}{(p+r)}$ involves the precision (p) and the recall (r) but I do not have negative labels to compute p. I thus adopt a pseudo-F score proposed by Liu and his colleagues in [118], $F^* = \frac{r^2}{Pr[f(X)=1]}$, where $Pr[f(X) = 1]$ is the probability that an observation is classified as positive; r can be estimated using the positive examples in the validation set and $Pr[f(X) = 1]$ can be estimated from the whole validation set. The authors claim the pseudo-F score works in a similar behavior to the F score in the sense that it is larger when both p and r are larger and is small if either p or r is small. The Structured Biased Support Vector Machine (SB-SVM) model is compared to other BSVMs with various kernels: linear, poly2 and XOR. The linear kernel uses the raw feature X to compute an equal-dimension weight W; the poly2 kernel is built with the original feature set X plus and its squares: $X^* = [X, X^2]$. The XOR kernel consists of the original feature set X and all the pairwise XOR features $X^* = [X, Z]$, $Z = \{x_i \oplus x_j, \forall x_i, x_j \sim \mathcal{B}\}$, where $\mathcal{B}$ indicates the family of binary features.

9.5.1 Synthetic Data

I generate the synthetic data by sampling the positive observations and unlabeled observations from the following Gaussian Mixture Models (GMM):

$$\begin{aligned} X_p &\sim \alpha N(\mu_p, \sigma_1) + (1-\alpha)N(\mu_p, \sigma_2), \\ X_u &\sim \alpha N(\mu_u, \sigma_1) + (1-\alpha)N(\mu_u, \sigma_2), \end{aligned}$$

where $\alpha = 0.6$, $\mu_p = [0.5, 3.5]$, $\mu_u = [-0.5, 0.5]$, $\sigma_1 = \begin{bmatrix} 1.1 & 0 \\ 0 & 2.9 \end{bmatrix}$ and $\sigma_2 = \begin{bmatrix} 1.3 & 0 \\ 0 & 2.7 \end{bmatrix}$.

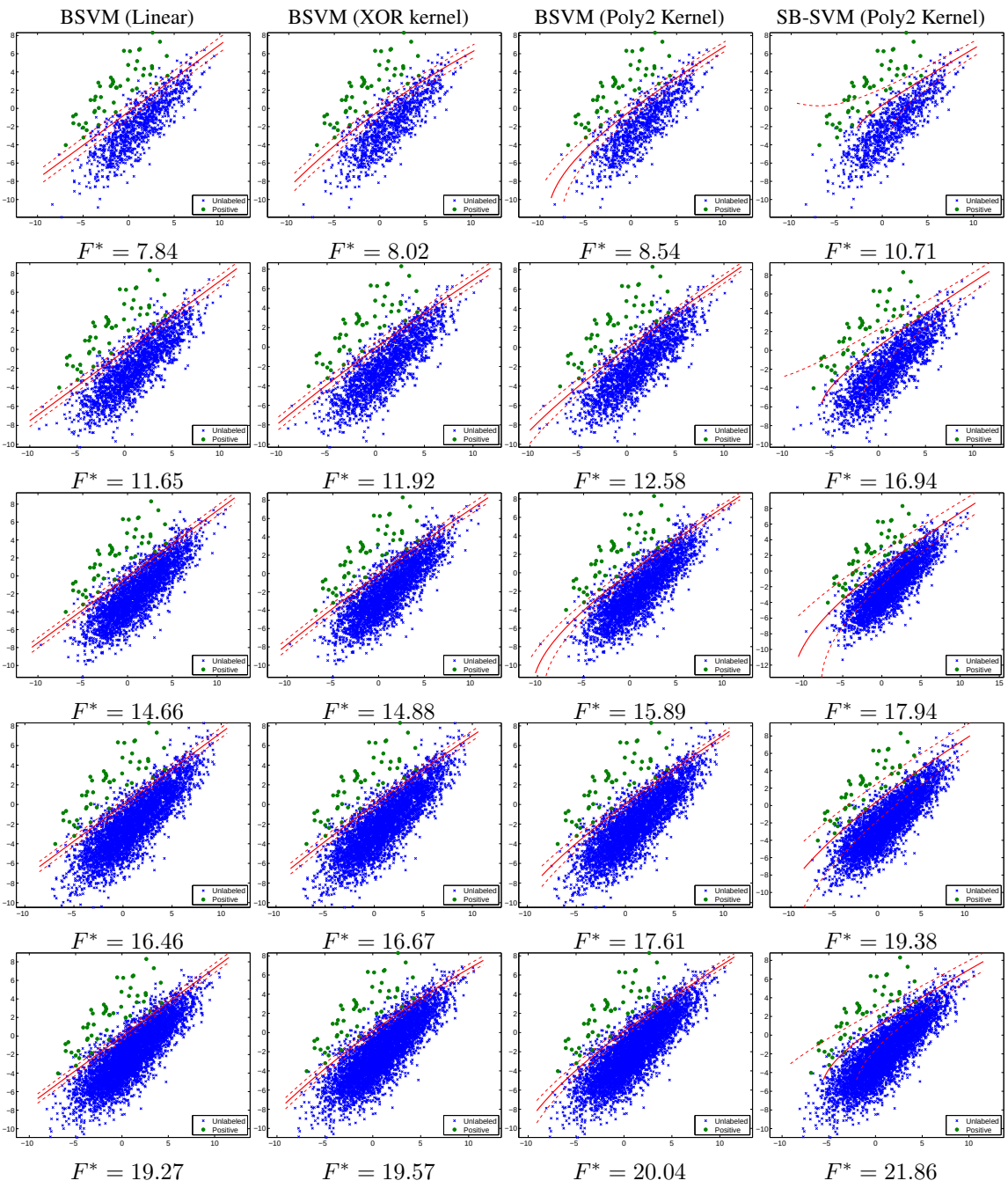


Figure 9.3: Visual comparison of different models for the "Biased Labeling" task. From left to right are: BSVM with linear kernel, BSVM with XOR kernel, BSVM with polynomial kernel and SB-SVM with polynomial kernel. Rows of these figures correspond to increasing amount of data samples, from 1050 to 6050. The number below each figure indicates the pseudo-F score.

I fix the size of $|X_p|$ to be 50 and increase the size of $|X_u|$ from $1,000$ to $6,000$ to create highly

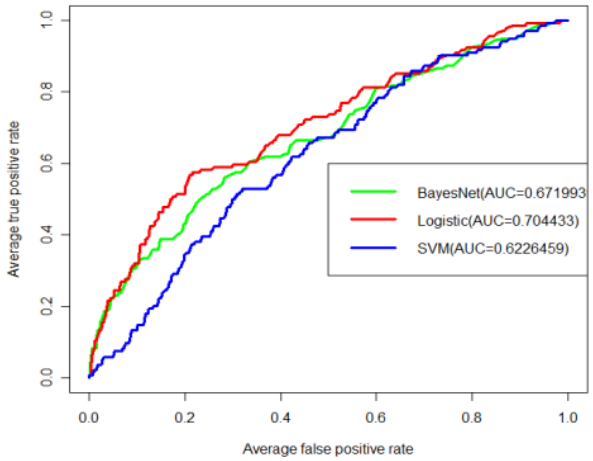
unbalanced observations, as indicated in Figure 9.3. I plot the positive observations in green and the rest of the unlabeled observations in blue. The region of positive observations overlaps with the unlabeled observations. Figure 9.3 illustrates the performance of four different models that handle training data with only positive and unlabeled observations. My model, BS-SVM, outperforms the others in terms of the pseudo-F scores, which can be interpreted as having a larger precision and recall.

9.5.2 Hospital Discharge Data-set

I evaluated models using a real world clinical data: Hospital_Discharge_Error. Please refer to Chapter 2 for details.

9.5.2.1 Predicting use traditional models

I first demonstrate how traditional models perform on this dataset. In specific, I use Naive Bayes, Logistic Regression and Support Vector Machine (SVM). The first two models are implemented in Weka [82] and the last one is available in LibSVM [35]. I use a linear kernel for SVM here so that all three models have a linear separation plane. Due to the lack of negative training observations, I have to use unlabeled observations as negative training samples in order to use these off-the-shelf supervised learning models. In the following experiment, all three models are evaluated on the entire data using a 10-cross-validation method.



	AUC	Recall
BayesNet	0.67	0.014
LR	0.70	0.019
SVM	0.62	0.038

Figure 9.4: Performance evaluations of off-the-shelf machine learning tools using the hospital discharge error data.

Figure 9.4 demonstrates the performance of these models. I have to note that the AUC value is merely

an approximation to the truth, which does not reflect the performance of the models. As I do not have "real" negative labeled observations, the averaged False Positive Rate (FPR^*) is calculated as $FPR^* = \frac{UP}{|Unlabeled|}$ to approximate $FPR = \frac{FP}{FP+TN}$, where UP indicates an unknown observation predicted as positive, FP corresponds to false positive and TN corresponds to true negatives. Even if the approximated AUC of these models are not bad, I will be very disappointed at a second look at the recall ($\frac{TP}{TP+FN}$), where FN indicates false negatives. All three models have very small recalls, which indicates they have missed almost every confirmed positive observation. Such performance makes these off-the-shelf models useless in my case because the positive observations are what really matters.

9.5.2.2 Predicting use SB-SVM

I have observed that existing supervised learning models are insufficient to handle only positive and unlabeled data. Next, I will evaluate how well my model and several BSVMs would perform. The following Table 9.3 summarizes the performance of various models . As in the synthetic data experiment, I include BSVM (linear), BSVM (Poly2) and BSVM (XOR) to be compared with their corresponding structured version. Obviously, the model SB-SVM does much better than BSVM in terms of both F^* score and recall. The SB-SVM (XOR) demonstrate the best performance in terms of both pseudo-F score and the recall because most raw features are binary or categorical.

Table 9.3: Performance comparison of SB-SVM and several BSVM models with different kernels

Model	F^* **score**	**Recall**
BSVM (Linear)	1.038	0.396
SB-SVM (Linear)	1.042	0.422
BSVM (Poly2)	1.185	0.444
SB-SVM (Poly2)	1.197	0.474
BSVM (XOR)	1.205	0.615
SB-SVM (XOR)	1.370	0.673

The metrics listed in Table 9.3 demonstrates the advantage of including semantic feature correlations to build models with only positive and unlabeled data.

9.5.3 Breast Cancer Gene Expression Data

The next experiment is conducted on real clinical gene expression data: GSE_3934 [128]. These data do not contain unlabeled observations but are highly unbalanced (224 good/18 poor). Unfortunately, traditional

machine learning algorithms do not work well even with the negative training labels available due to data unbalance. They tend to ignore the impact of the minor negative class for the sake of a higher overall *discrimination*. I divided the data into five folds and computed various models in terms of their average F score and recall using their best parameters. The F score is $\frac{2*Precision*Recall}{Precision+Recall}$, where $Recall = \frac{TP}{TP+FN}$ and $Precision = \frac{TP}{TP+FP}$. Note I calculate the true F score instead of a pseudo one as true labeling is available for this data. In this experiment, I include SVM, BSVM (linear) and BSVM (Poly2) to be compared with the SB-SVM that I developed in this chapter.

Table 9.4: Performance comparison of SB-SVM and several other models.

Model	F **score**	**Recall**
SVM	0.824	0.778
BSVM (Linear)	0.735	1.000
SB-SVM (Linear)	0.750	1.000
BSVM (Poly2)	1.000	1.000
SB-SVM (Poly2)	1.000	1.000

Table 9.4 again demonstrates the performance advantage of Structured Biased Models against previous approaches. Compared with SVM, SB-SVMs consider more important positive data and achieve higher recalls. On the other hand, SB-SVM outperforms BSVM in the F score by incorporating useful semantic feature correlations in the model construction. Figure 9.5 illustrates a Graphical User Interface I designed.

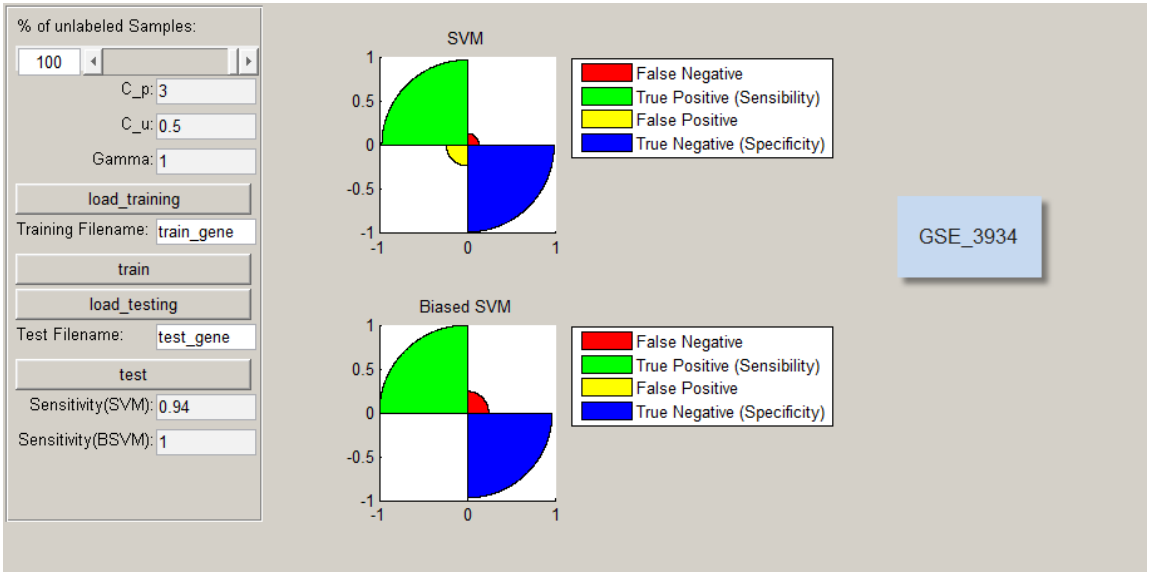


Figure 9.5: Graphical User Interface for interacting with Biased Support Vector Machine.

9.6 Discussion

Learning a predictive model for bias labeled data, e.g., only positive and unlabeled data, is a problem of great importance in biomedical informatics. In many cases, there are abundant unlabeled observations while it is only affordable to manually review and label the class membership of a few cases. Such limitation is largely due to the cost of additional follow-up, extra staff time, and expensive lab tests.

Traditional supervised learning algorithms usually both positive and negative labeled inputs to determine a decision boundary between their co-variate patterns. However, bias labeled data do not have negative labeled samples, and there traditional supervised learning approaches find it difficult to handle them. Actually, this dilemma is twofold: on one hand, there is a very limited amount of labeled data that supervised learning algorithms can use; on the other hand, the number of unlabeled samples are tremendous, but supervised learning algorithms cannot extract useful information from them.

To tackle this challenge, I decided to construct a model that can synthesize labeling information and structural knowledge from data. Specifically, I reformulated the Biased Support Vector Machine framework to incorporate semantic correlations, which are estimated from unlabeled data. This semantic correlation serves as a structural prior that regulates my model towards the narrowed feasible regions of optimization. Thus, the method jointly optimizes information from different perspectives: class membership and semantic feature correlations to minimize the ambiguity in either perspective, if considered separately.

9.7 Conclusion

Biomedical applications often involve constructing models with only positive and unlabeled samples. Traditional supervised learning algorithms require training samples of both class memberships, thus having difficulties in handling such a situation. In contrast, I developed a method to jointly consider the labeling information and the structural knowledge from bias-labeled data. I reformulated Biased Support Vector Machine to consider semantic correlations. The semantic correlation are estimated from unlabeled data, and provided a prior that structurally regulates my model towards narrower feasible regions of optimization. That is, the joint optimization of two complementary perspectives, class membership and semantic feature correlations, offers modeling advantages over methods that consider these perspectives independently. This data-driven approach has demonstrated performance advantage over previous methods like Support Vector Machine and Biased Support Vector Machine using both synthetic data and clinically related data.

Chapter 10

Applicability Across Different Data

Frequently, machine learning models developed for specific tasks involve context specific assumptions and their generalizability to a different setting or applications is limited. However, useful approaches should capture the transferable knowledge of problems they study while not overfitting them with problem specific assumptions. Successful machine learning models (e.g., Logistic Regression, Support Vector Machines, Decision Trees and et al.) follow the law of succinctness, that is, models should be kept as simple as possible, but no simpler, to tackle the problem. Along the same lines of thought, I developed learning models to assist biomedical decision making processes with minimum assumptions about particular problems of interest. The models developed in this thesis, including TM3N, SIO and AC-LR, are essentially data-driven approaches to learning and understanding observations from different sources. The models make only necessary constraints about the properties of the data to be applied but not the situation and the particular context to be applied.

For personalized clinical decision support systems, I require only the outcome variable to be a single dichotomous variable so that my classifiers can use these labels to estimate the "true probability" of events. There are no limitations on the co-variants, and they can be categorical, binary, and numerical. I also made no assumptions about the co-variant occurrence patterns, but the models can learn such information from data. The outputs of these models are calibrated probabilistic values for each co-variable pattern.

For large-scale disease outbreak prediction, I need measurements and states of multiple correlated outcome variables as models' inputs. Note that this implies that these inputs should be provided together rather than separately. The correlations of these variables can be of different types, including temporal dependence,

spatial correlation, and relational correlation. I relaxed the values of these outcome variables to multi-class labels instead of dichotomous values. Finally, the outputs of these models are simultaneously estimated class labels for correlated outcome variables.

To verify the generalizability, I applied models developed in previous chapters to a variety of data. The following table summarizes the model V.S. data used in this chapter.

Table 10.1: Summary of models applied to various data to demonstrate model generalizability.

Algorithm	Breast Cancer	Myocardial Infarction	Hospital Discharge	Bankruptcy	PIMATR	HeightWeight
LR	x	x	x	x	x	x
LR-PS	x	x	x	x	x	x
LR-IR	x	x	x	x	x	x
LR-SIR	x	x	x	x	x	x
AC-LR	x	x	x	x	x	x

	Synthetic LDS	BioWar I	BioWar II	Building Occupancy
HMM	x	x	x	x
CRFs	x	x	x	x
M3N	x	x	x	x
TM3N	x	x	x	x

Specifically, data included for evaluating model generalizability are: Breast Cancer Gene Expression data (GSE2034, GSE2990, GSE3494), Myocardial_Infarction data (Edinburgh, Sheffield), Hospital Discharge Error Prediction data (HOSPITAL), three UCI machine learning data (Bankruptcy, PIMATR and HeightWeight) [69], BioWar I,II simulation data (BioWar-I,II) and Building Occupancy Detection data (OCCUPANCY). I compared various approaches with above mentioned data using standard *calibration* (HL test) and *discrimination* (AUC, accuracy) metrics.

10.1 Data

I used various data to test my model's generalizability. For single target variable prediction models, I used data with binary output variables, i.e., BREAST_CANCER, PIMATR, HOSPITAL, MI, HEIGHT_WEIGHT, BREAST_CANCER and BANKRUPTCY. For multiple target co-estimation models, I used data that contain correlated sources of observation which is recorded over time, i.e., BioWar-I, II and OCCUPANCY data.

10.1.1 Breast Cancer Gene Expression data

These data were obtained from the NCBI Gene Expression Omnibus (GEO). Three individual data downloaded were previously studied by Wang et al. (GSE2034) [185], Sotiriou et al. (GSE2990) [166], and Miller et al. (GSE3494) [128], respectively.

To make my data compatible with previous studies, I followed the criteria in [140] to select patients, who did not receive any treatment and had negative lymph node status. Among these pre-selected candidates, only patients with extreme outcomes, either poor outcomes (recurrence or metastasis within five years) or very good outcomes (neither recurrence nor metastasis within eight years) are selected. The number of samples after filtering are: 209 for GSE2034 (114 good/95 poor), 90 for GSE2990 (60 good/30 poor), and 242 for GSE2034 (224 good/18 poor).

I also applied a split to divide GSE3494 into two groups, as suggested by [140], GSE3494-A and GSE3493-B, according to the sample's Affymetrix platform. Thus, the breast cancer dataset has four separate data. All these data have a feature size of 247,965, which corresponds to the gene expression results obtained from micro-array experiments.

10.1.2 Pimatr data

A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization (WHO) criteria. The data contains the 532 complete records after dropping the (mainly missing) data on serum insulin.

Table 10.2: Descriptional statistic for the PIMATR data set.

obs	npreg	glu	bp	
Min. : 1.00	Min. : 0.00	Min. : 56.0	Min. : 38.00	
1st Qu.: 50.75	1st Qu.: 1.00	1st Qu.:100.0	1st Qu.: 64.00	
Median :100.50	Median : 2.00	Median :120.5	Median : 70.00	
Mean :100.50	Mean : 3.57	Mean :124.0	Mean : 71.26	
3rd Qu.:150.25	3rd Qu.: 6.00	3rd Qu.:144.0	3rd Qu.: 78.00	
Max. :200.00	Max. :14.00	Max. :199.0	Max. :110.00	
skin	bmi	pedigree	age	type
Min. : 7.00	Min. :18.20	Min. :0.0900	Min. :21.00	0: 132
1st Qu.:20.75	1st Qu.:27.57	1st Qu.:0.2500	1st Qu.:23.00	1: 68
Median :29.00	Median :32.80	Median :0.3700	Median :28.00	
Mean :29.21	Mean :32.31	Mean :0.4613	Mean :32.11	
3rd Qu.:36.00	3rd Qu.:36.50	3rd Qu.:0.6200	3rd Qu.:39.25	
Max. :99.00	Max. :47.90	Max. :2.2900	Max. :63.00	

The data contain the following columns: 'npreg' number of pregnancies; 'glu' plasma glucose concentration in an oral glucose tolerance test; 'bp' diastolic blood pressure (mm Hg); 'skin' triceps skin fold thickness (mm); 'bmi' body mass index (weight in kg/(height in m)^2); 'ped' diabetes pedigree function; 'age' age in years; finally, 'type' 1 for 'Yes' or 0 for 'No', for diabetic according to WHO criteria.

10.1.3 Hospital Discharge Error data

The "hospital discharge error" data consist of microbiology cultures and other variables related to hospital discharge errors [59]. The following table defines various features and outcome variables for these data.

Table 10.3: Details of the co-variables and the outcome variable in the hospital discharge error data. Eight out of ten explanatory variables are categorical and two are numerical.

Name	**Details**
	Features
Specimen:	0=blood, 1=urine, 2=sputum, 3=csf
Spec_days:	Number of days between admission date and specimen collection date.
Collect_week:	0=specimen collected on weekday, 1=specimen collected on weekend
Final_week:	0=final result on weekday, 1=final result on weekend
Vistyp:	1=admission, 0=non-admission
Svc:	0=<blank> (patient not admitted), 1=ONC, 2=MED, 3=Medical Sub-specialties, 4=Surgery and Surgical Sub-specialties, 5=Other
Age:	Age in years
Female:	0=male, 1=female
Race:	0=white, 1=black, 2=Asian, 3=Hispanic, 4=other, 5=unknown/declined
Insurance:	0=medicare, 1=medicaid, 2=commercial, 3=other
	Target Variable
Pot_error:	0=not a potential follow-up error, 1=a potential follow-up error

I also summarized features and the outcome variable by their description statistic in the table below. The clinical meaning for each column was explained in Chapter 2.There are 369 clinically important but highly suspicious observations out of 4,819 returned post-discharge observations, which makes the data highly unbalanced and challenge to calibrate.

Table 10.4: Descriptional statistics for the hospital discharge error data set.

specimen	specimen days	collect week	final week	vistype	svc
0: 233	1 :1245	0:3755	0:3583	1:4818	1: 665
1:2564	0 :1030	1:1063	1:1235		2:1287
2:1467	2 : 682				3:1217
3: 554	3 : 391				4:1608
	4 : 327				5: 41
	5 : 227				
	(Other): 916				
age	female	race	insurance	pot error	
Min. : 0.00	0:2252	0:3360	0:1996	0:4449	
1st Qu.:43.28	1:2566	1: 577	1: 554	1: 369	
Median :57.76		2: 110	2:2152		
Mean :56.51		3: 405	3: 116		
3rd Qu.:71.24		4: 55			
Max. :99.71		5: 311			

10.1.4 Myocardial Infarction data

Figure 10.1: Descriptional statistics for the Edinburgh data set

Abbreviation				
age	min: 13.0	median:59	mean:57.6	max: 92
Smokes	0: 785	1: 468		
Exsmoker	0: 959	1: 294		
Fhistory	0: 967	1: 286		
Diebetes	0: 1165	1: 88		
BP	0: 1053	1: 200		
Lipids	0: 1215	1: 38		
CPmajorSymp	0: 62	1: 1191		
Restrostern	0: 331	1: 922		
Lchest	0: 907	1: 346		
Rchest	0: 1109	1: 144		
Back	0: 1122	1: 131		
Larm	0: 670	1: 583		
Rarm	0: 1042	1: 211		
breath	0: 1031	1: 222		
postural	0: 1017	1: 236		
Cwtender	0: 1201	1: 52		
Sharp	0: 1208	1: 45		
Tight	0: 572	1: 681		
Sweating	0: 739	1: 514		
SOB	0: 731	1: 522		
Nausea	0: 1124	1: 129		
Vomiting	0: 1124	1: 129		
Syncope	0: 1208	1: 45		
Episodic	0: 1161	1: 92		
Worsening	min: 0.0	median: 4.0	mean: 17.4	max: 168
Duration	min: 0.0	median: 3.0	mean: 8.84	max: 168
prev-ang	0: 699	1: 554		
prev-MI	0: 836	1: 361		
Worse	0: 892	1: 361		
Crackles	0: 1106	1: 147		
Added-HS	0: 1247	1: 6		
Hypoperfusion	0: 1203	1: 50		
Stelve	0: 1199	1: 54		
NewQ	0: 1240	1: 13		
STorT-abnorm	0: 1240	1: 13		
LBBBorRBBB	0: 1203	1: 50		
Old-MI	0: 1101	0: 152		
Old-isch	0: 1141	1: 112		
MI	0: 979	1: 274		

Figure 10.2: Descriptional statistics for the Sheffield data set

Abbreviation				
age	min: 17.0	median:61	mean:59.9	max: 91
Smokes	0: 318	1: 182		
Exsmoker	0: 388	1: 112		
Fhistory	0: 373	1: 127		
Diebetes	0: 451	1: 49		
BP	0: 403	1: 97		
Lipids	0: 482	1: 18		
CPmajorSymp	0: 37	1: 463		
Restrostern	0: 110	1: 390		
Lchest	0: 373	1: 127		
Rchest	0: 438	1: 62		
Back	0: 426	1: 74		
Larm	0: 237	1: 263		
Rarm	0: 418	1: 82		
breath	0: 422	1: 78		
postural	0: 455	1: 45		
Cwtender	0: 491	1: 9		
Sharp	0: 400	1: 100		
Tight	0: 246	1: 254		
Sweating	0: 235	1: 265		
SOB	0: 281	1: 219		
Nausea	0: 341	1: 159		
Vomiting	0: 449	1: 51		
Syncope	0: 467	1: 33		
Episodic	0: 417	1: 83		
Worsening	min: 0.0	median: 6.0	mean: 50.37	max: 1000
Duration	min: 0.0	median: 4.0	mean: 12.34	max: 1000
prev-ang	0: 281	1: 219		
prev-MI	0: 377	1: 123		
Worse	0: 338	1: 162		
Crackles	0: 373	1: 127		
Added-HS	0: 476	1: 24		
Hypoperfusion	0: 441	1: 59		
Stelve	0: 403	1: 97		
NewQ	0: 470	1: 30		
STorT-abnorm	0: 403	1: 97		
LBBBorRBBB	0: 474	1: 26		
Old-MI	0: 454	1: 46		
Old-isch	0: 473	1: 27		
MI	0: 346	1: 154		

The Myocardial Infarction (MI) data correspond to results of patient both with and without myocardial infarction which were observed at emergency department in UK [98]. The data contain patient records from

two medical centers in Britain; among these, 600 patients attending at the emergency room (ER) with chest pain that were observed in Sheffield, England, and 1,253 patients with the same symptoms were observed in Edinburgh, Scotland. More details about the MI data set were provided in Chapter 2.

10.1.5 Height and Weight data

These data are on height vs weight for two groups: men and women. The subjects are 213 students of an academic University. There are 73 females and 140 males. The data contains the following features: height, weight, GPA, left arm length, right arm length, left foot size and right foot size. The following table summarized descriptional statistic of this data.

Table 10.5: Descriptional statistic for the HEIGHT_WEIGHT data set.

Sex	Height	Weight	GPA
0: 73	Min. :55.00	Min. : 95.0	Min. :1.240
1: 140	1st Qu.:64.00	1st Qu.:125.0	1st Qu.:2.670
	Median :67.00	Median :140.0	Median :3.000
	Mean :67.31	Mean :145.5	Mean :3.004
	3rd Qu.:70.50	3rd Qu.:160.0	3rd Qu.:3.400
	Max. :79.00	Max. :280.0	Max. :3.910
LArm	RArm	LFoot	RFoot
Min. :20.50	Min. :20.50	Min. :19.50	Min. :20.00
1st Qu.:24.00	1st Qu.:24.00	1st Qu.:23.40	1st Qu.:23.00
Median :25.00	Median :25.00	Median :24.70	Median :25.00
Mean :25.17	Mean :25.31	Mean :25.16	Mean :25.20
3rd Qu.:26.50	3rd Qu.:27.00	3rd Qu.:27.00	3rd Qu.:27.00
Max. :31.00	Max. :31.00	Max. :32.00	Max. :32.00

10.1.6 Breast Cancer Gene Expression data

This data were obtained from the NCBI Gene Expression Omnibus (GEO). Three individual data were previously studied by Wang et al. (GSE2034) [185], Sotiriou et al. (GSE2990) [166], and Miller et al. (GSE3494) [128], respectively.

To make my data compatible with previous studies, I followed the criteria in [140] to select patients, who did not receive any treatment and had negative lymph node status. Among these pre-selected candidates, only patients with extreme outcomes, either poor outcomes (recurrence or metastasis within five years) or very good outcomes (neither recurrence nor metastasis within eight years) were selected. The number of samples

after filtering were: 209 for GSE2034 (114 good/95 poor), 90 for GSE2990 (60 good/30 poor), and 242 for GSE2034 (224 good/18 poor).

I also applied a split to divide GSE3494 into two groups, as suggested by [140], GSE3494-A and GSE3493-B, according to the sample's Affymetrix platform. Thus, the breast cancer data-set has four separate data. All of these data have a feature size of $247,965$, corresponds to the gene expression results obtained from micro-array experiments.

10.1.7 Bankruptcy data

The Bankruptcy data contain two features: Return and EBIT (earnings before interest and taxes). The outcome variable "Bankruptcy" is binary. There are 66 samples in this data, where 33 samples correspond to observed bankruptcy and the others do not. The following table summarizes their description statistic, i.e., min, 1st Qu., median, 3rd Qu. ,max.

Table 10.6: Descriptional statistics for the BANKRUPTCY data set.

Return	EBIT	Bankruptcy
Min. :-308.90	Min. :-280.000	0:33
1st Qu.: -39.05	1st Qu.: -17.675	1:33
Median : 7.85	Median : 4.100	
Mean : -13.63	Mean : -8.226	
3rd Qu.: 35.75	3rd Qu.: 14.400	
Max. : 68.60	Max. : 34.100	

10.1.8 BioWar-I data

The BioWar-I data contain one-year-period observations in the city of Pittsburgh, PA, from 9/1/2002 to 8/31/2003. The total number of simulated agents are 306,181. There is one outbreak of airborne diseases (avian influenza) during the simulation period. The data incorporate both relational and temporal information.

In this data, the simulated agents interact and transmit airborne diseases (avian influenza) over time. There are six time ticks everyday; thus $365 * 6$ time ticks are observed for each year. The following table summarizes features and outcome variables of all 306,181 agents that are simulated. More details and motivation about this data were introduced in Chapter 2.

Table 10.7: Descriptional statistics for BioWar-I.

tick	dayOfWeek	month	day	dead	is.er
Min. : 0.0	Fri:312	Aug : 186	Min. : 1.00	Min. : 0.000	Min. : 0
1st Qu.: 547.2	Mon:312	Dec : 186	1st Qu.: 8.00	1st Qu.: 0.000	1st Qu.: 0
Median :1094.5	Sat:312	Jan : 186	Median :16.00	Median : 0.000	Median : 7
Mean :1094.5	Sun:318	Jul : 186	Mean :15.72	Mean : 4.338	Mean : 696
3rd Qu.:1641.8	Thu:312	Mar : 186	3rd Qu.:23.00	3rd Qu.: 0.000	3rd Qu.: 13
Max. :2189.0	Tue:312	May : 186	Max. :31.00	Max. :97.000	Max. :19401
	Wed:312	(Other):1074			
kidsAtHome	adultsAtHome	at.work	weblookup	medcalls	num.exchanges
Min. :20959	Min. : 37447	Min. : 0	Min. : 0.0	Min. : 0	Min. : 0.0
1st Qu.:31566	1st Qu.: 86910	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0	1st Qu.: 0.0
Median :78695	Median :217960	Median : 0	Median : 8.0	Median : 0	Median : 0.0
Mean :59889	Mean :151609	Mean : 41487	Mean : 695.9	Mean : 18867	Mean : 101.9
3rd Qu.:78701	3rd Qu.:217985	3rd Qu.: 0	3rd Qu.: 15.0	3rd Qu.: 0	3rd Qu.: 0.0
Max. :79497	Max. :226684	Max. :187787	Max. :19043.0	Max. :141408	Max. :11438.0
in.hospital	is.home	is.work	is.school	is.pharmacy	is.doctor
Min. : 0	Min. : 58563	Min. :0	Min. : 0	Min. : 0.0	Min. : 0.0
1st Qu.: 0	1st Qu.:118408	1st Qu.:0	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0.0
Median : 7	Median :296655	Median :0	Median : 0	Median : 0.0	Median : 0.0
Mean : 696	Mean :211498	Mean :0	Mean : 9277	Mean : 557.6	Mean : 125.9
3rd Qu.: 13	3rd Qu.:296683	3rd Qu.:0	3rd Qu.: 0	3rd Qu.: 13.0	3rd Qu.: 3.0
Max. :19401	Max. :306181	Max. :0	Max. :58370	Max. :25178.0	Max. :10315.0
is.stadium	is.theater	is.store	is.restaurant	is.university	is.military
Min. : 0	Min. : 0	Min. :0	Min. : 0	Min. :0	Min. :0
1st Qu.: 0	1st Qu.: 0	1st Qu.:0	1st Qu.: 0	1st Qu.:0	1st Qu.:0
Median : 0	Median : 0	Median :0	Median : 0	Median :0	Median :0
Mean :1437	Mean : 3997	Mean :0	Mean : 31342	Mean :0	Mean :0
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.:0
Max. :7408	Max. :20286	Max. :0	Max. :157115	Max. :0	Max. :0

10.1.9 BioWar-II data

The BioWar-II data contain multiple five-year-period observations from 9/1/2002 to 8/30/2007. The number of simulated agents are set to vary from 153,090 to 1,224,726, at an approximately equal scale ($150k$); specifically, the number of simulated agents vary from 10% (153,090) to 100% (1,224,726). The city of simulation is Norfolk, VA. There ws one outbreak of airborne diseases for every year during the simulated period. The data incorporate both relational and temporal information.

In this data, the simulated agents interact and transmit airborne diseases (avian influenza) over time.

There are six time ticks everyday; thus $365 * 6$ time ticks are observed for each year. I used the BioWar simulation engine to generate ten five-year periods rather than a single 50-year period to avoid the impact of birth and death factors on the disease modeling. The following table summarizes features and outcome variables of all 1,224,736 agents that are simulated. More details and motivation about this data were introduced in Chapter 2.

Table 10.8: Descriptional statistic for BioWar-II.

tick	dayOfWeek	month	day	dead	is.er
Min. : 0	Fri:1560	Dec : 930	Min. : 1.00	Min. :0	Min. : 0.00
1st Qu.: 2737	Mon:1566	Jan : 930	1st Qu.: 8.00	1st Qu.:0	1st Qu.: 0.00
Median : 5474	Sat:1560	Jul : 930	Median :16.00	Median :0	Median : 36.00
Mean : 5474	Sun:1566	Mar : 930	Mean :15.72	Mean :0	Mean : 38.94
3rd Qu.: 8212	Thu:1566	May : 930	3rd Qu.:23.00	3rd Qu.:0	3rd Qu.: 49.00
Max. :10949	Tue:1566	Oct : 930	Max. :31.00	Max. :0	Max. :368.00
	Wed:1566	(Other):5370			
kidsAtHome	adultsAtHome	at.work	weblookup	medcalls	num.exchanges
Min. : 85069	Min. :154198	Min. : 0	Min. : 0.00	Min. : 0	Min. : 0.0000
1st Qu.:126708	1st Qu.:362493	1st Qu.: 0	1st Qu.: 0.00	1st Qu.: 0	1st Qu.: 0.0000
Median :316864	Median :907737	Median : 0	Median : 42.00	Median : 0	Median : 0.0000
Mean :240999	Mean :622832	Mean :175130	Mean : 45.76	Mean :102464	Mean : 0.8463
3rd Qu.:316867	3rd Qu.:907859	3rd Qu.: 0	3rd Qu.: 57.00	3rd Qu.: 0	3rd Qu.: 1.0000
Max. :316867	Max. :907859	Max. :753630	Max. :375.00	Max. :595796	Max. :46.0000
in.hospital	is.home	is.work	is.school	is.pharmacy	is.doctor
Min. : 0.00	Min. : 239387	Min. :0	Min. : 0	Min. : 0.00	Min. : 0.000
1st Qu.: 0.00	1st Qu.: 489130	1st Qu.:0	1st Qu.: 0	1st Qu.: 0.00	1st Qu.: 0.000
Median : 36.00	Median :1224600	Median :0	Median : 0	Median : 0.00	Median : 0.000
Mean : 38.94	Mean : 863832	Mean :0	Mean : 37529	Mean : 37.51	Mean : 9.815
3rd Qu.: 49.00	3rd Qu.:1224726	3rd Qu.:0	3rd Qu.: 0	3rd Qu.: 56.00	3rd Qu.: 15.000
Max. :368.00	Max. :1224726	Max. :0	Max. :231797	Max. :542.00	Max. :237.000
is.stadium	is.theater	is.store	is.restaurant	is.university	is.military
Min. : 0	Min. : 0	Min. :0	Min. : 0	Min. :0	Min. :0
1st Qu.: 0	1st Qu.: 0	1st Qu.:0	1st Qu.: 0	1st Qu.:0	1st Qu.:0
Median : 0	Median : 0	Median :0	Median : 0	Median :0	Median :0
Mean : 4988	Mean :15666	Mean :0	Mean :127495	Mean :0	Mean :0
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.: 0	3rd Qu.:0	3rd Qu.:0
Max. :25269	Max. :78581	Max. :0	Max. :634041	Max. :0	Max. :0

10.1.10 Occupancy data

The OCCUPANCY data were collected by the School of Architecture, Carnegie Mellon University, for the cost-efficient operation and better understanding of occupancy behavior in buildings. The sensor network is setup in an open plan office space with six rooms and one kitchen/printer room, as indicated in Figure 10.3.

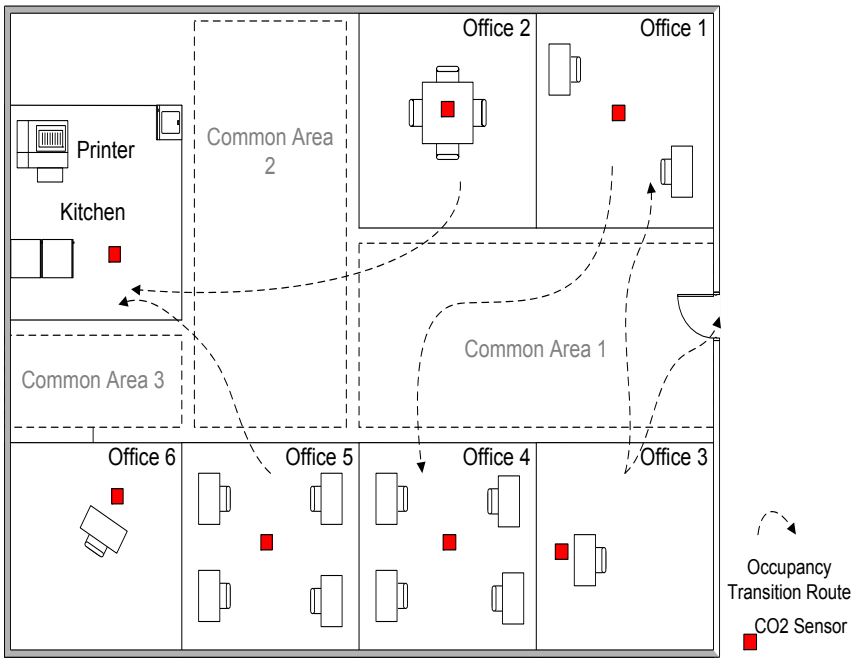


Figure 10.3: Geometric flat view of the office area testbed

The physical sensor network includes a wired CO_2 network and a data server. One CO_2 sensor is installed in the center of each office at the nose level $(1.1m)$ above the ground. Data collection for this experiment was for one continuous period, with a sampling rate of every one minutes, capturing CO_2 measurements and the number of occupants in four offices. The time period was three weeks from March 17th, 2008 to April 4th, 2008 excluding weekends. Occupancy data was recorded from 8:00am to 8:00pm from the four offices (2, 3, 4 and 5). Office 2 and 5 have four Ph.D. students; office 4 has two graduate students ; and office 3 has 1 faculty. I synchronized the measurements from all sensors; and aggregated measurements for every 10 minutes to predict the averaged occupant numbers in a ten-minute window. The truth occupancy information is collected use a network of commercial cameras. The following table summarizes the features and outcome variables for the OCCUPANCY data.

Table 10.9: Summary of outcome variables for the building occupancy data.

nodeID	linkQuality	batteryVoltage	tSensorAvg
office 2:10815	Min. :60.00	Min. :3.288	Min. :19.49
office 3:10747	1st Qu.:70.00	1st Qu.:3.293	1st Qu.:23.07
office 4:10763	Median :74.00	Median :3.294	Median :23.93
office 5:11163	Mean :73.99	Mean :3.307	Mean :24.08
	3rd Qu.:77.00	3rd Qu.:3.299	3rd Qu.:25.06
	Max. :92.00	Max. :3.343	Max. :28.36
tSensorOutlier	hSensorAvg	hSensorOutlier	iSensorAvg
Min. :19.59	Min. : 6.435	Min. : 0.00	Min. : -0.3644
1st Qu.:23.26	1st Qu.:19.636	1st Qu.:19.84	1st Qu.:116.3161
Median :24.13	Median :22.708	Median :22.90	Median :117.2441
Mean :24.26	Mean :25.107	Mean :25.29	Mean :107.9557
3rd Qu.:25.16	3rd Qu.:30.102	3rd Qu.:30.28	3rd Qu.:118.0678
Max. :28.55	Max. :44.709	Max. :44.99	Max. :119.9538
iSensorOutlier	pSensorAvg	pSensorOutlier	aSensorAvg
Min. : -0.4047	Min. : -5.80226	Min. :-5.081e+00	Min. : -0.9299
1st Qu.:116.4016	1st Qu.: -0.00099	1st Qu.:-3.094e-04	1st Qu.: 2.1183
Median :117.3187	Median : 0.16477	Median : 2.482e+00	Median : 3.6830
Mean :108.0394	Mean : 7.40846	Mean : 8.631e+00	Mean : 3.5792
3rd Qu.:118.1897	3rd Qu.: 0.76869	3rd Qu.: 2.494e+00	3rd Qu.: 4.6513
Max. :120.0339	Max. :375.12995	Max. : 9.778e+02	Max. :105.8745
aSensorOutlier	CO2	CO2outside	total
Min. : -1.164	Min. : -1.0	Min. : -1.0	No Obervation:12918
1st Qu.: 4.070	1st Qu.: 451.0	1st Qu.:355.0	0 occupant :11200
Median : 5.721	Median : 511.0	Median :377.0	1 occupant :10397
Mean : 8.349	Mean : 491.9	Mean :368.5	2 occupants: 6863
3rd Qu.: 8.579	3rd Qu.: 570.0	3rd Qu.:401.0	3 occupants: 1372
Max. :262.026	Max. :1127.0	Max. :534.0	4 occupants: 733
			5 occupants: 5

10.2 Single Target Variable Prediction Models

All data described in the previous section are randomly split into training and testing at a ratio factor that varies from 0.1 to 0.9. Figure 10.4 to 10.12 illustrate the performance of the models; each figure corresponds to one particular training/testing split. Each column of these figures corresponds to one specific data, from the top row to the bottom row, I plot the distribution of probabilistic outputs, Logistic Regression reliability diagram (LR), Platt Scaling reliability diagram (PS), Isotonic Regression reliability diagram (IR), Smooth

Isotonic Regression reliability diagram (SIR) and Adaptive Calibrated Logistic Regression reliability diagram (ACLR), respectively. The first three were existing models in the literature and the last two are models that I developed in this thesis.

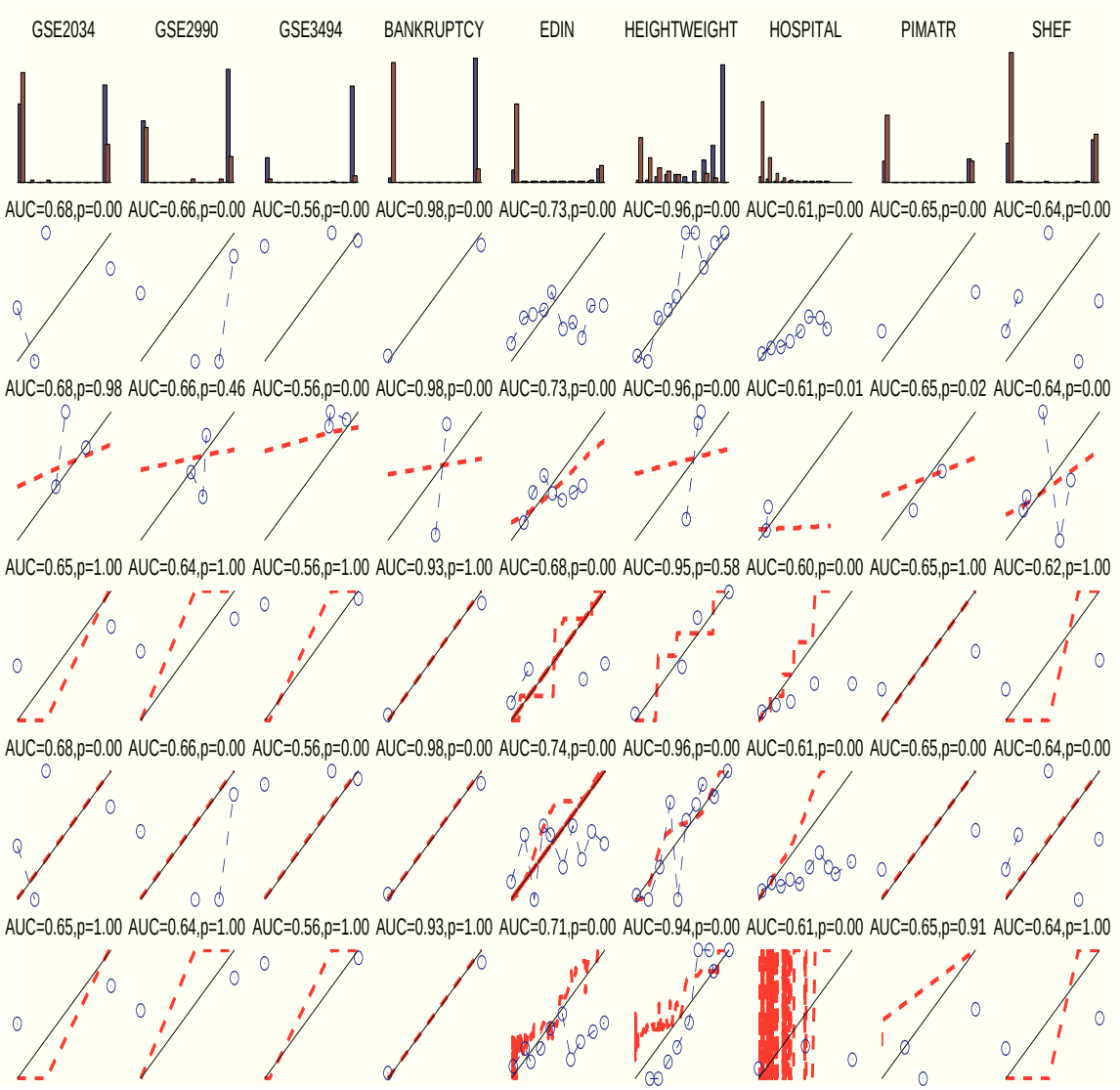


Figure 10.4: Model generalizability evaluation: training/testing ratio 1:9.

I evaluated all five models (LR, PS, IR, SIR and ACLR) using a small fraction of all nine datasets. I plotted histograms of the original predicted probabilities (blue bars for class "1", red bars for class "0"), the reliability diagrams for all five models (blue circles) and their corresponding *calibration* mapping function (red curves). LR failed to pass the HL-test at significance level 0.05 for all cases. Platt Scaling failed in

most cases showing its insufficiency to handle complex real world scenarios. Even though IR demonstrates better *calibration*, its *calibration* mapping functions are not smooth and are unrealistically sharp at the corners. The SIR methods had slighted worse *calibration* than IR, which indicates smoothness might not be a useful objective when observations are limited. Finally, ACLR demonstrated comparable performance to IR approach in terms of both AUC and HL-test performance .

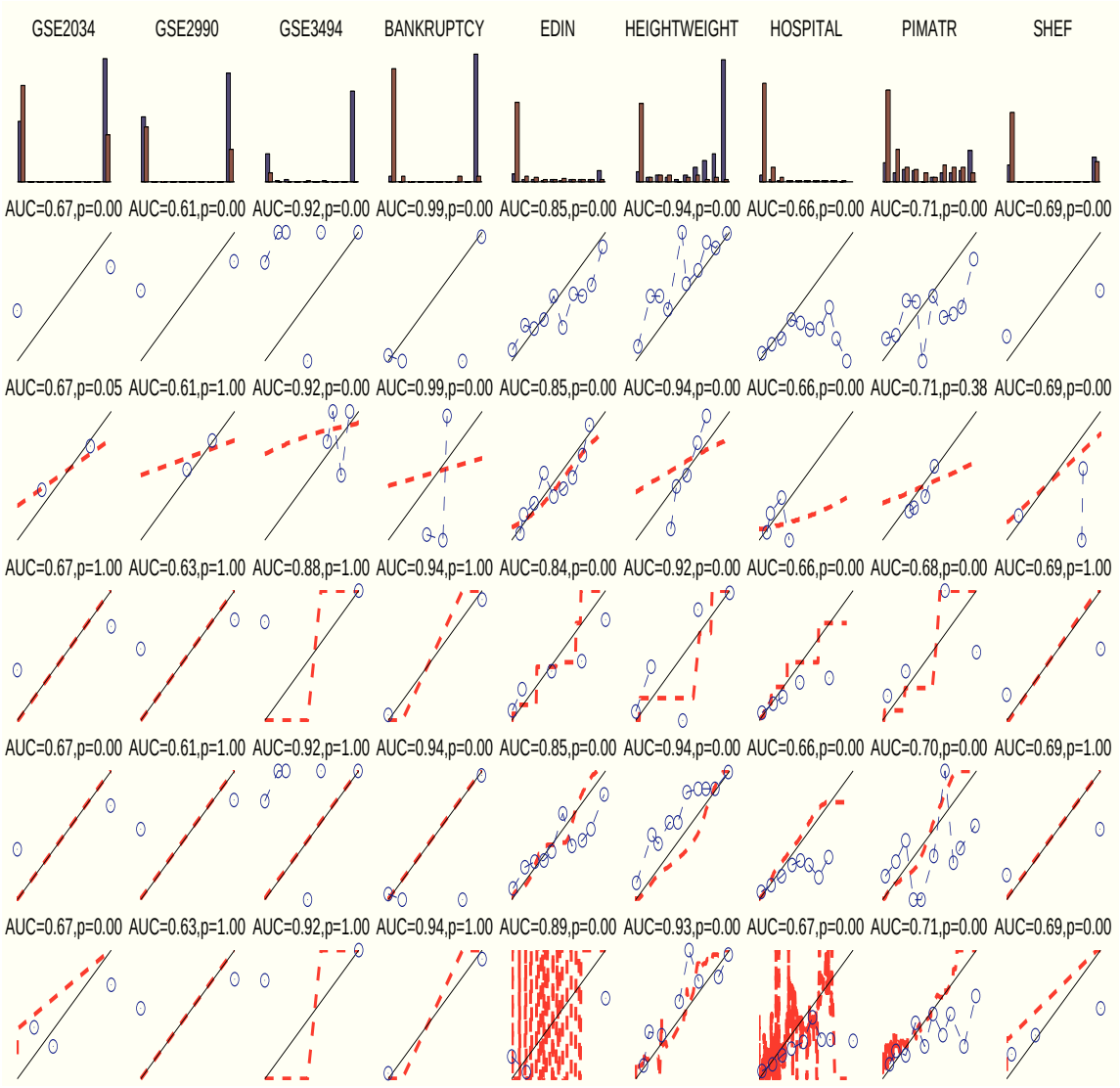


Figure 10.5: Model generalizability evaluation: training/testing ratio 2:8

Next, I increased the training/test ratio to 2:8 for all nine datasets. Similar to the previous experiment, LR failed in all HL-test at significance level at 0.05. Again, PS failed in 6 out of 9 datasets, thus showing

its limitation in calibrating. IR retained its *calibration* ability, and SIR demonstrated similar *calibration* performance but with a smoother mapping function. Finally, ACLR showed slightly worse *calibration* but better AUC in most cases. The results so far indicated that all these methods, including existing approaches and these developed in this thesis, generalized well to different data with consistent performance.

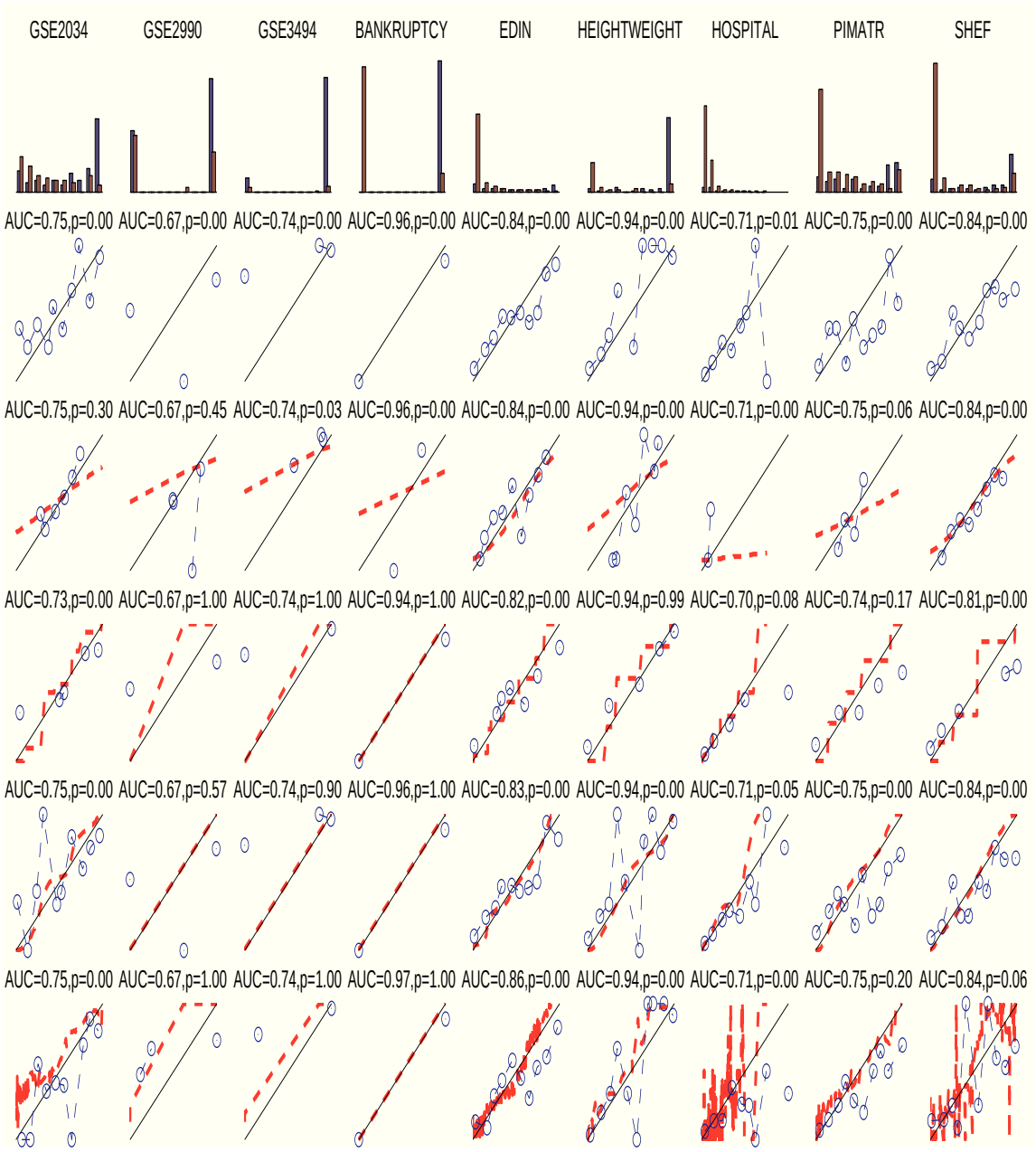


Figure 10.6: Model generalizability evaluation: training/testing ratio 3:7

As before, I increased the training/testing ratio to 3:7 so that more data are used to train models. LR still failed in the HL-tests and PS showed poor *calibration* ability, i.e. only 3 out of 9 cases are calibrated. IR, SIR and ACLR demonstrated similar performance and good generalizability across data from different sources.

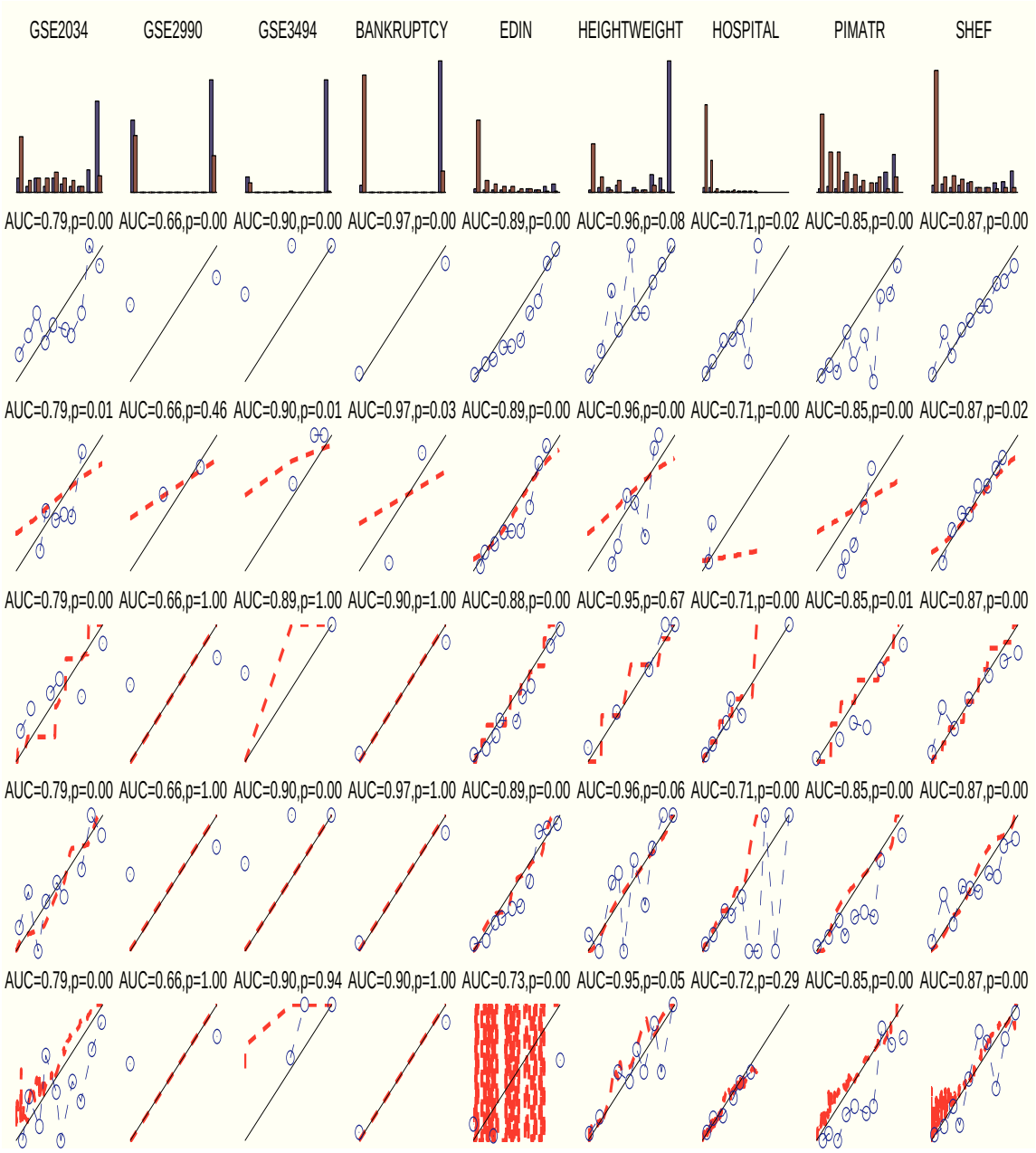


Figure 10.7: Model generalizability evaluation: training/testing ratio 4:6

I further increased the training/testing ratio to 4:6. At increased training samples, LR still failed in all HL-test while PS did not do well in *calibration*. IR and SIR showed similar performance in both AUC and HL-test. ACLR outperformed the rest methods in comparison, in which it calibrated 5 out of 9 cases and demonstrated superior AUC values.

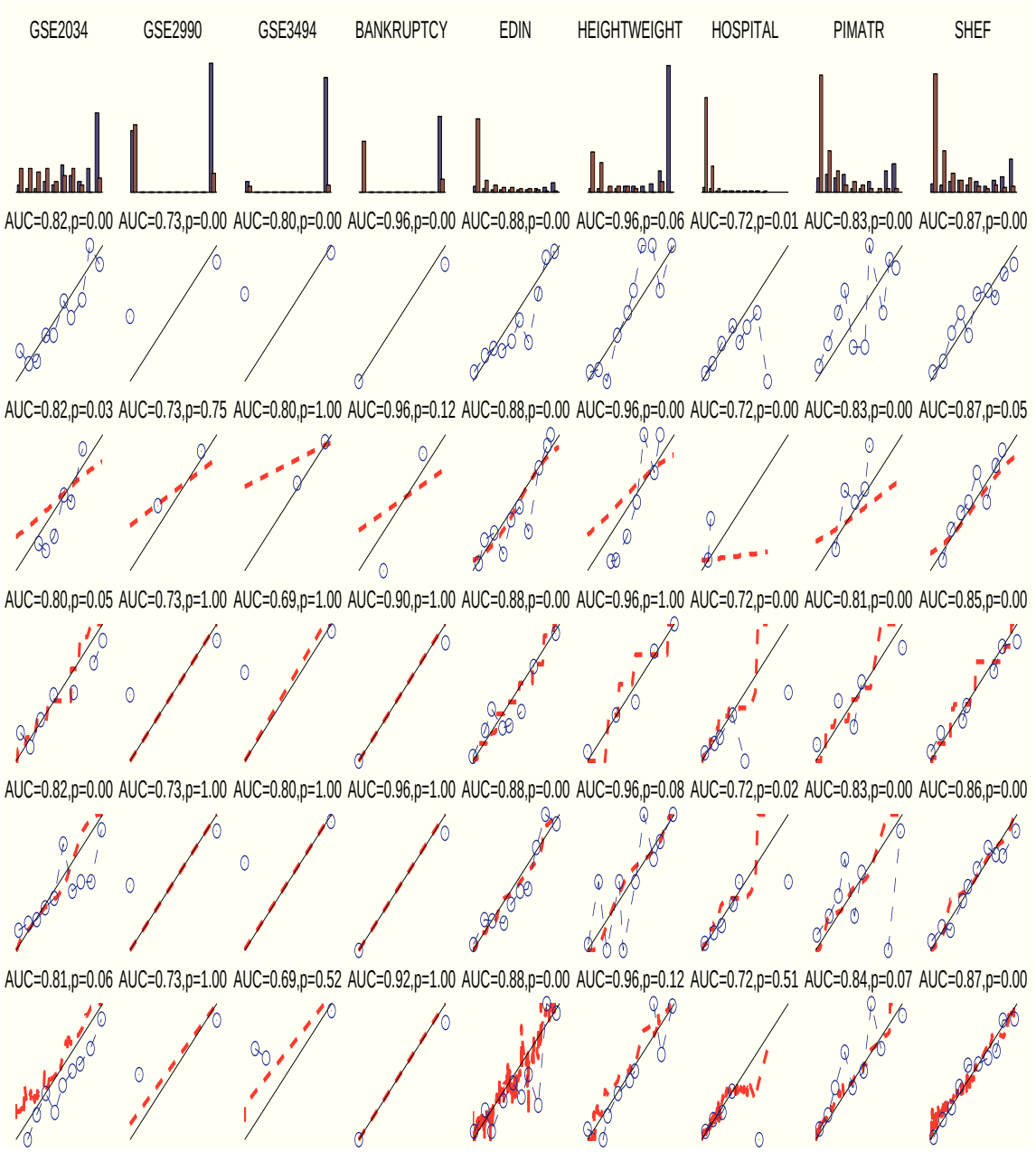


Figure 10.8: Model generalizability evaluation: training/testing ratio 5:5

At training/testing ratio of 5:5, I trained all five models and compared their performance. LR and PS continued to demonstrate poor performance, which indicates that both are not appropriate methods for *calibration*. IR and SIR retained their performance in *calibration*. ACLR again outperformed both approaches and led the performance with 7 out of 9 success in its *calibration*.

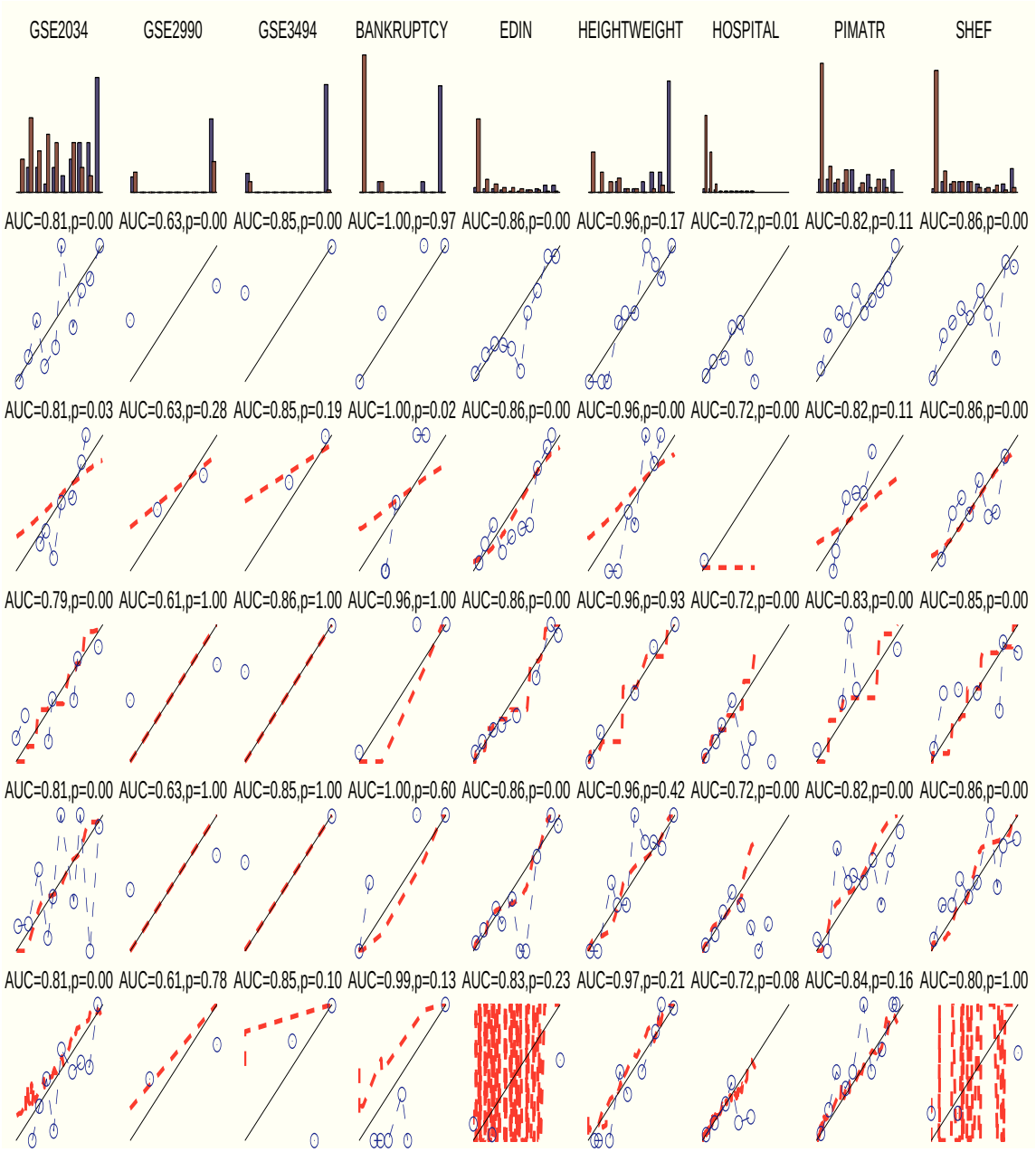


Figure 10.9: Model generalizability evaluation: training/testing ratio 6:4

With an increased training/testing ratio of 6:4, I compared five models' AUC and their *calibration* ability in terms of HL-test. While LR showed poor performance again, PS showed increased *calibration* ability with more training data. It passed three out of nine HL-tests and indicated that PS could benefit from having more training data. Similarly, ACLR showed increased *calibration* ability with more observations. In this experiment, ACLR passed seven out of nine HL-tests at significance level 0.05, which outperformed both IR and SIR. Specifically, SIR and IR passed two out of nine and three out of nine HL-tests, respectively.

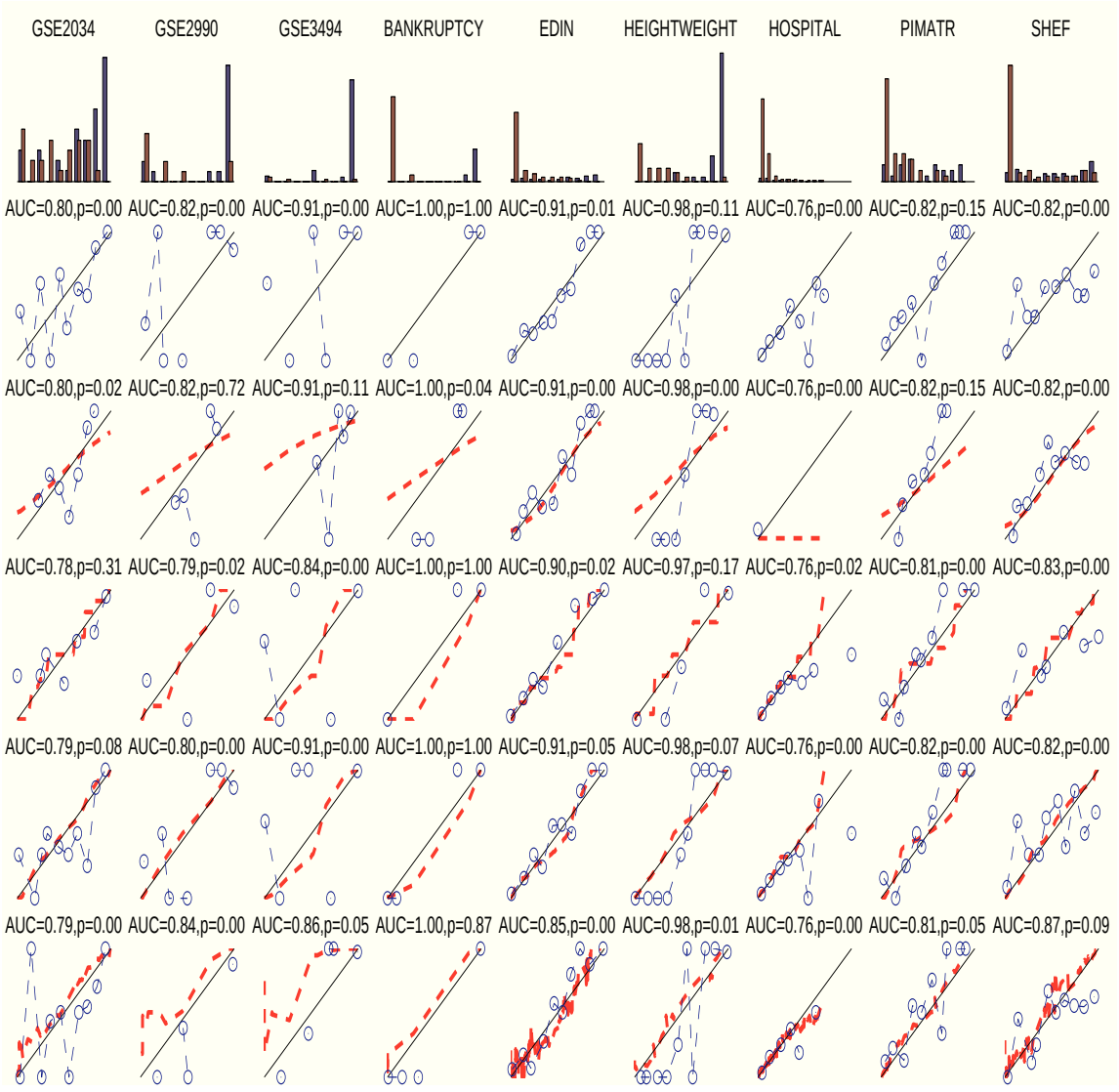


Figure 10.10: Model generalizability evaluation: training/testing ratio 7:3

Similar to the previous case, at a training/testing ratio of 7:3, LR failed in HL-tests. PS passed 3 out of

9 tests, which is a significant improvement compared with its previous performance with less training data. ACLR still led the competition of *calibration*, followed by IR and SIR. The latter two approaches tied in their performance.

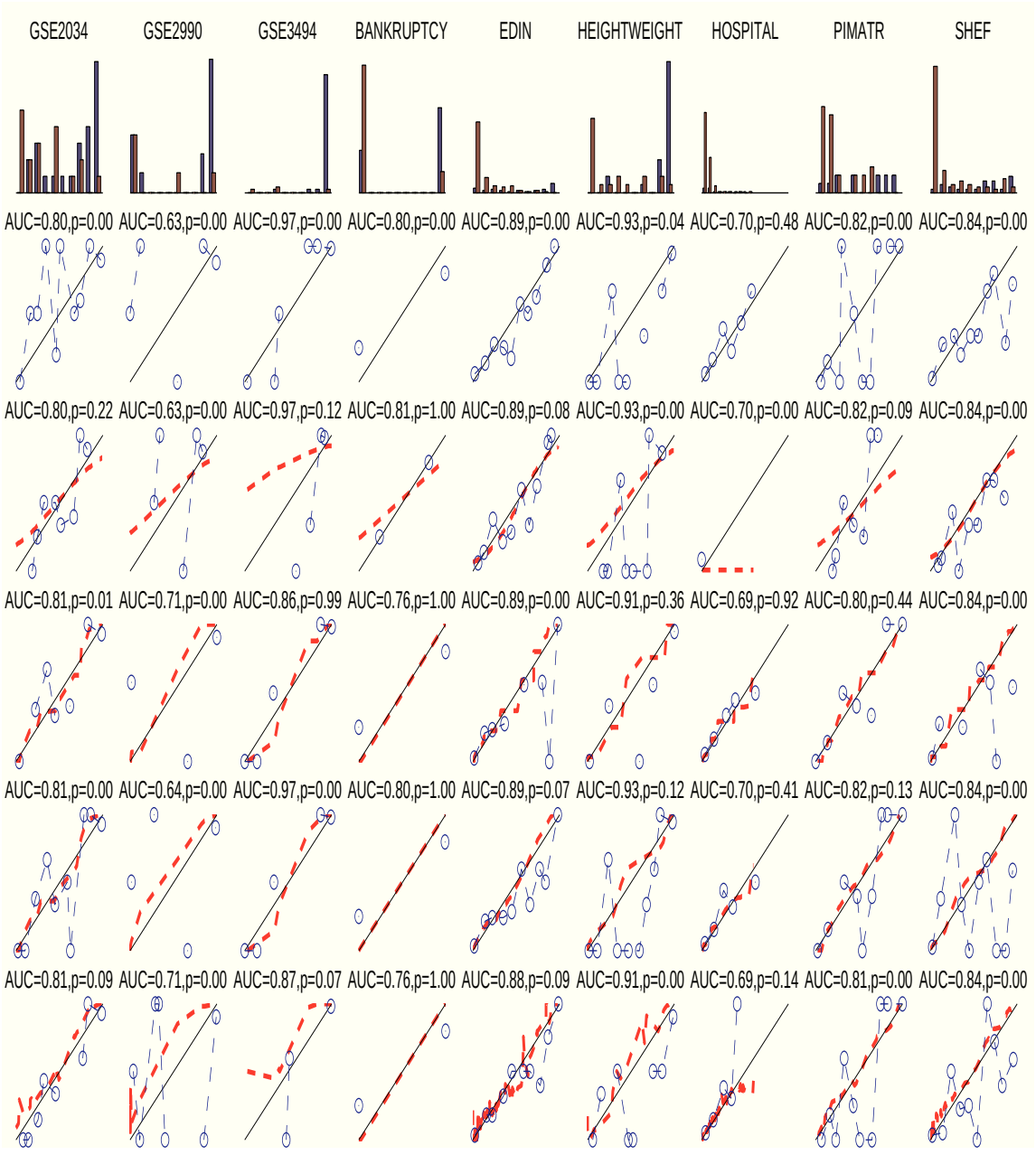


Figure 10.11: Model generalizability evaluation: training/testing ratio 8:2

With an increased the training/testing ratio of 8:2, I compared five models' AUC and their *calibration*

ability in terms of HL-test. As before, LR did not do well with its raw probabilistic outputs and needed to be rectified. Interestingly, PS, IR, SIR and ACLR tried in their *calibration* performance, which seems to indicate that all *calibration* models benefit from increasing the amount of training data, regardless of their model assumptions.

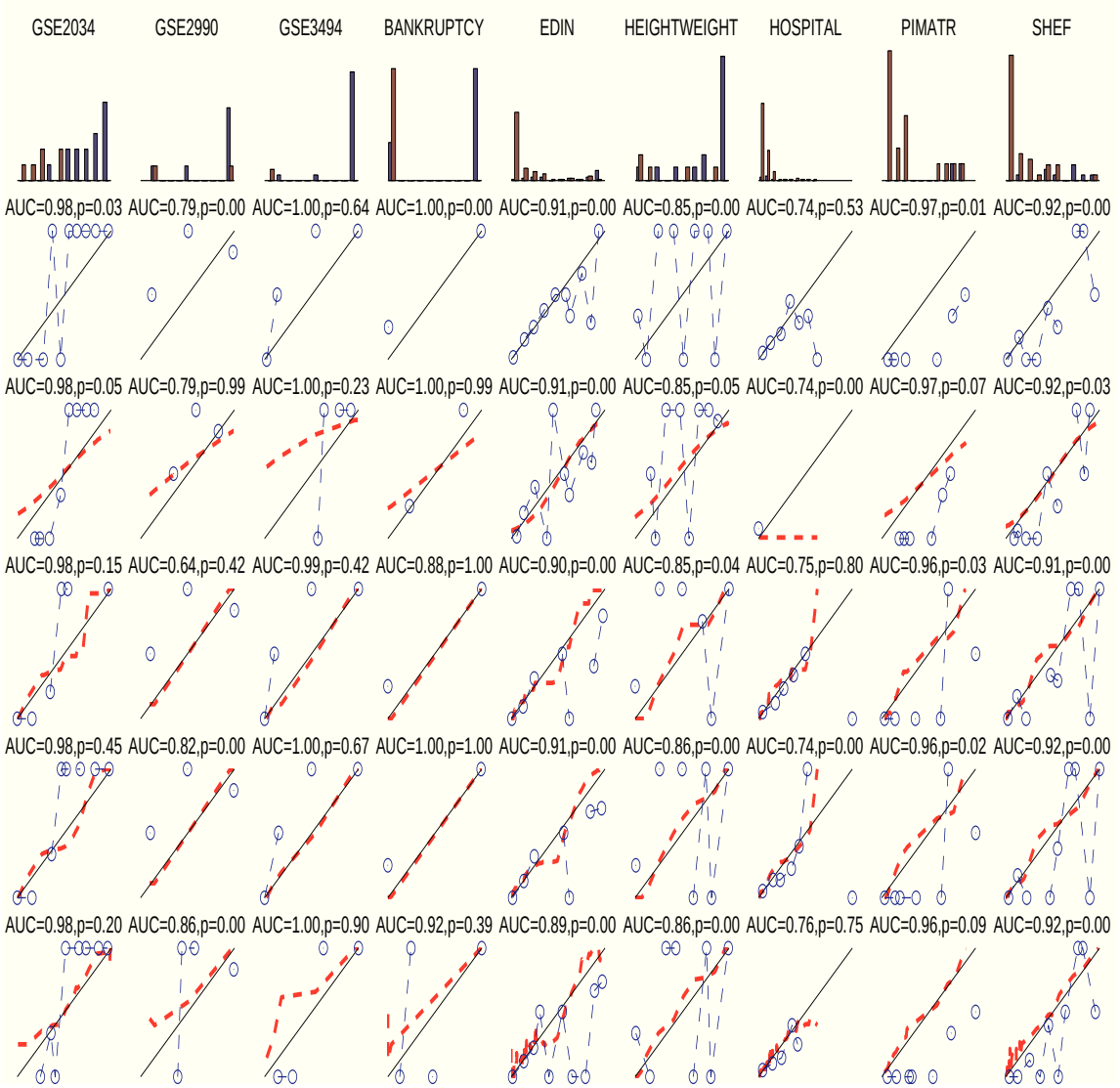


Figure 10.12: Model generalizability evaluation: training/testing ratio 9:1

The last experiment used a training/testing ratio of 9:1. That is, most sample data enters the training to build the models. LR did not pass any HL-test, which indicated raw outputs of LR need *calibration*. PS demonstrated increased performance with more training data. It passed 5 out of 9 tests, which is comparable

to that of IR, SIR and ACLR. The result indicated PS, IR, SIR and ACLR could all have good *calibration* performance with abundant training data.

All these examples demonstrated the generalizability of SIR and AC-LR approaches developed in this thesis. In most cases, these two are the leading methods in terms of *calibration* and their performance is consistent across different sources of data and various ratios of the training/test sample size.

10.3 Multiple Target Variables Co-Estimation Models

I apply my multi-variable prediction model (TM3N) to simulated LDS data, BioWa- I, II and OCCUPANCY to show its applicability across different data. Along with my model TM3N, I test CRFs, M3N and HMM models.

The synthetic Linear Dynamic System (LDS) data is generated using formula introduced in Chapter 4. The equations below simulate a temporal and spatial dependent linear system, which specifies the hidden state Y_t^i that depends temporally on the previous state Y_{t-1}^i and correlates spatially with the states of the neighboring sites Y_t^j, $j \in \mathcal{N}_i$.

$$Y_t^i = \alpha Y_{t-1}^i + (1-\alpha) \sum_{j \in \mathcal{N}_i} \beta^j Y_t^j + e_1, \tag{10.1}$$

$$X_t^i = A Y_t^i + e_2, \tag{10.2}$$

$$e_1 \sim N(0, \sigma_{e_1}^2), \tag{10.3}$$

$$e_2 \sim N(0, \sigma_{e_2}^2), \tag{10.4}$$

where $\mathcal{N}_i$ corresponds to the neighboring sites of i but excludes i; A is a projection vector that maps hidden states to the observations; X_t^i corresponds to the observations at site i, time tick t; e_1 and e_2 are the environmental Gaussian noises;α represents the temporal/relational trade-off parameter. If α is set to be zero, the system considers no time dependence. Otherwise, if α is set as one, the system ignores relational correlations. The following table lists the average accuracy of applying four methods to synthetic LDS data with various α value.

Table 10.10: Averaged accuracy of four different methods using synthetic LDS data with various α value. The number in each cell indicates the averaged accuracy.

	Value of α							
Models	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
HMM	0.01	0.11	0.21	0.25	0.33	0.38	0.45	0.51
CRFs	0.66	0.54	0.52	0.38	0.34	0.29	0.2	0.23
M3N	0.68	0.54	0.47	0.4	0.35	0.39	0.34	0.23
TM3N	0.68	0.64	0.59	0.58	0.59	0.6	0.58	0.63

Table 10.11: Comparison of the average accuracy for the OCCUPANCY data.

OCCUPANCY	
Algorithm	Accuracy
HMM	36.5%
CRFs	49.81%
M3N	49.05%
TM3N	$\mathbf{69.76}\%$

Table 10.12: Comparison of the average accuracy for BioWar-I data.

BioWar I	
Algorithm	Accuracy
HMM	65%
CRFs	57%
M3N	58%
TM3N	$\mathbf{69}\%$

Table 10.11 summarizes the comparison results of the four different methods applied to the OCCUPATION data. For this task, the HMM model gives an average accuracy of 36.5% when temporal correlations are considered. Clearly, the first order Markov model over time is not sufficient to capture system dynamics. On the other hand, relational models such as CRFs and M3N show similar results. Although slightly better than HMM, are still unsatisfactory. A significant improvement in average accuracy was observed when I combined both temporal and structural influence into a unified model, TM3N. The results indicated that the temporal and relational aspects complement each other, and joint optimization reduces the ambiguity when single aspect is considered separately.

Regarding the BioWar-I data, a similar pattern of improvements can be observed in Figure 10.12. However, temporal correlations dominate the system dynamics of this data. As a result, HMM showed a better performance than its relational counterparts M3N and CRFs. TM3N, which synthesized both information, achieved best average accuracy at 69%, outperforming the rest in comparison.

Both improvements in predictions using OCCUPANCY and BioWar-I data own a big part of their success

to the global modeling of both temporal and structural perspectives, which offers a more comprehensive description to complicated yet noisy observations of the unknown system dynamics.

Finally, all four methods are evaluated using BioWar II. Figure 10.13 illustrates the accuracies, where all four methods stabilize around the scale of 80%. TM3N's accuracy increases 10%, which outperforms HMM(6%), CRF(2%) and M3N(4%).

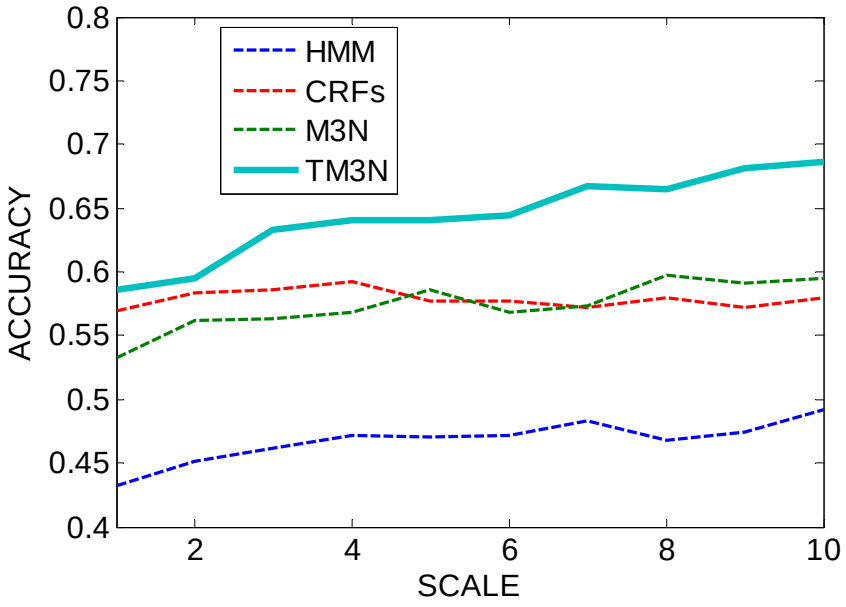


Figure 10.13: Model generalizability evaluation using BioWar II data. The X axis corresponds to the scale of agents at ten different levels (10%-100%); and Y axis represents the accuracy. The figure shows TM3N generalize to various scales of BioWar data with a better performance than conventional methods in comparison.

10.4 A Universal Model Access Platform

To make the research in *calibration* and *discrimination* more useful and applicable across various data and different models, I developed a free web-service, WEBCALIBSIS, to provide evaluation of performance for any probabilistic predictive models. It investigates *calibration* and *discrimination* using statistical and graphical standard measures and tools, and is capable of comparing various models. WEBCALIBSIS integrates a simple upload interface, flexible PHP and Java-scripts, and a powerful computational engine in a three-layer structure capable of processing multiple requests efficiently. This tool is available at

`http://128.2.219.203/webcalibsis`, and can be used under the terms of GNU general public license as published by the Free Software Foundation.

As opposed to other applications that focus on ROC analysis (Table 10.13), I provided a more comprehensive evaluation of a classifier. My tool also allows statistical comparison of classifiers. Furthermore, WEBCALIBSIS offers easy and simple access via an intuitive web interface that can evaluate different models across platform. Figure 10.14 illustrated a made-up example that WEBCALIBSIS evaluated four different models (Logistic Regression, Naive Bayes, Support Vector Machine and Adaptive Calibration for Logistic Regression) using the same simulation data.

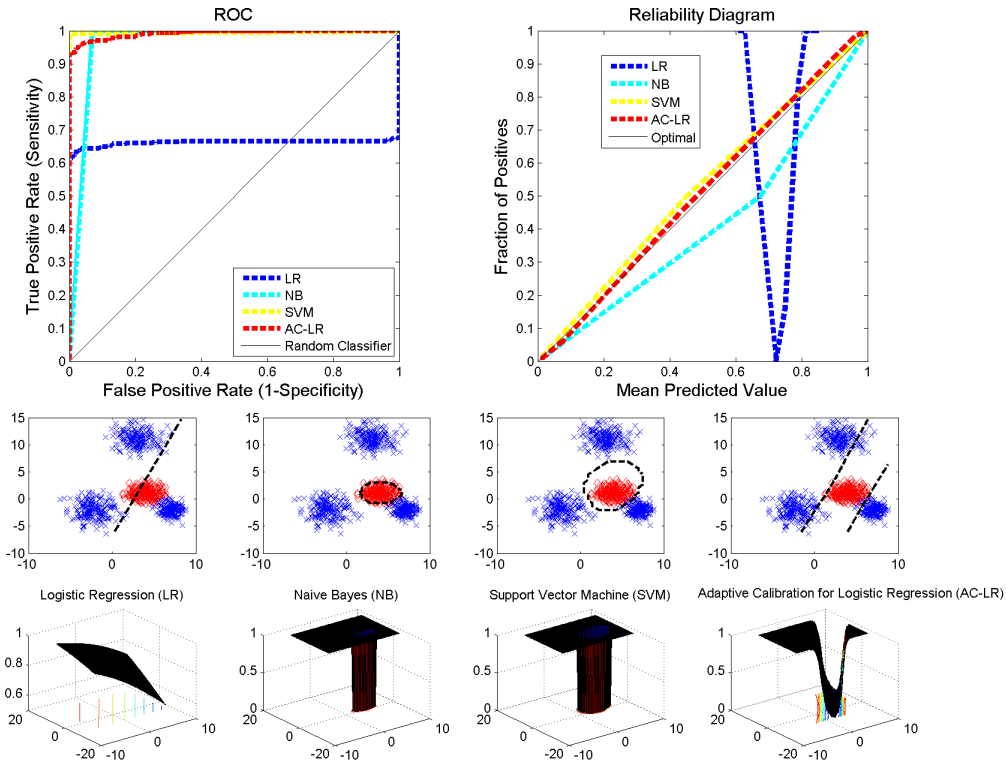


Figure 10.14: A made-up example that uses WEBCALIBSIS to evaluate various models.

To show the difference between WEBCALIBSIS and other existing applications, the following table lists various models (AccuROC [181], ROCR [164], Analyse-It [6], Web-based ROC [1], MedCalc [125], LABMRMC [106]), and their capabilities.

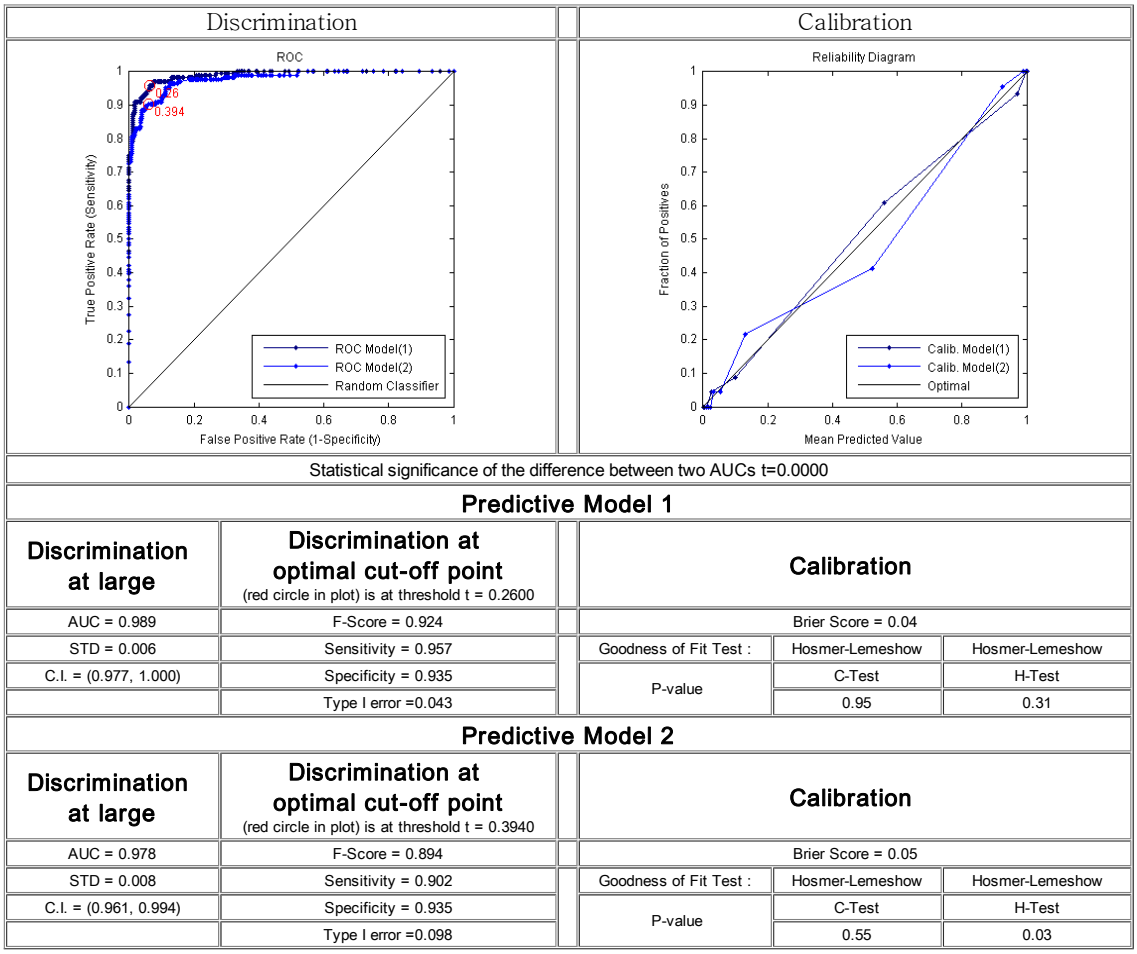
Table 10.13: Comparison of existing applications for the assessment of the quality of predictive models.

Application	ROC Analysis	Calibration Analysis	Visualization (AUC)	Model Comparison	Web Access	Freeware
AccuROC	x		x			
ROCR	x	x	x		x	x
Analyse-It	x		x	x		
Web-based ROC	x		x			x
MedCalc	x		x			
LABMRMC	x		x			x
Webcalibsis	x	x	x	x	x	x

A simple way to assess the *calibration* of a predictive model is to plot the average estimate for groups representing either (a) pre-defined ranges of the classifier estimates, or (b) percentiles of estimated risk against (c) the proportion of positive cases in that group. Such *calibration* plots constitute a useful subjective visualization tool, but lack quantitative evidence. A global assessment of *calibration* is represented by the Brier score [23], which measures the average squared error (where error corresponds to the difference between the estimate and the observed outcome. The well-known Hosmer-Lemeshow (HL) goodness-of-fit statistic [90] for logistic regression models provides a quantitative *calibration* index by measuring to what extent an estimate for a case approximates the relative frequency for a group of cases that have similar estimates. Here, too, groups of cases are represented either by pre-defined ranges of the estimates (HL-H) or percentiles of estimated risks (HL-C). In addition to *calibration* analysis, WEBCALIBSIS presents standard measures for assessing the *discrimination* of classifiers in terms of the ROC curve, the area under the ROC curve (AUC), the AUC's standard deviation and confidence interval, as well as the ROC cut-off point maximizing the Youden index [66].

WEBCALIBSIS is organized in a three-layer structure. The front-end for the user is a simple upload interface that accepts plain text files. The format of the input files is defined in the way that each row consists of probabilistic estimates for a case and the case's associated class label, e.g., the first $N-1$ columns represent outputs from one or more probabilistic models and the last column contains the binary coded presence ("1") or absence ("0") of an event. In the mid-layer, WEBCALIBSIS utilizes Java-script and PHP to link the inputs from the front-end to local databases, one folder for each requesting IP address. Thus, requests can

be processed in parallel without interfering with one another. The computational engine, the bottom layer of the three-layer structure, calculates and provides visualization of the statistics. Finally, WEBCALIBSIS generates a report summarizing the evaluation, as illustrated in Figure 10.15. I used Wisconsin Diagnostic Breast Cancer (WDBC) data [69], which contains 569 observations of cell nuclei (212 malignant and 357 benign), for my illustration. I display the comparison of two predictive models (Model 1: Logistic Regression; Model 2: Support Vector Machine), both constructed with 32 variables obtained from images of fine needle aspiration (FNA) of breast masses.



Statistical significance of the difference between two AUCs t=0.0000

Predictive Model 1

Discrimination at large	Discrimination at optimal cut-off point (red circle in plot) is at threshold t = 0.2600		Calibration		
AUC = 0.989	F-Score = 0.924		Brier Score = 0.04		
STD = 0.006	Sensitivity = 0.957		Goodness of Fit Test :	Hosmer-Lemeshow	Hosmer-Lemeshow
C.I. = (0.977, 1.000)	Specificity = 0.935		P-value	C-Test	H-Test
	Type I error =0.043			0.95	0.31

Predictive Model 2

Discrimination at large	Discrimination at optimal cut-off point (red circle in plot) is at threshold t = 0.3940		Calibration		
AUC = 0.978	F-Score = 0.894		Brier Score = 0.05		
STD = 0.008	Sensitivity = 0.902		Goodness of Fit Test :	Hosmer-Lemeshow	Hosmer-Lemeshow
C.I. = (0.961, 0.994)	Specificity = 0.935		P-value	C-Test	H-Test
	Type I error =0.098			0.55	0.03

Figure 10.15: Sample output of Webcalibsis. On the left column, I plot the ROC and show the discrimination at large as well as at the optimal cut-off point; on the right column, I illustrate the calibration of the model in terms of its reliability diagram, Brier score and HL-test p-values.

10.5 Discussion

The generalizability of a learning model is important because it allows it to be used in related studies without having to rebuilt it. It is also a necessary condition for the system to be versatile. This chapter investigated the cross-data and cross-platform performance of models developed in this thesis. For this purpose, I used data from various sources, e.g., clinically related ones and more general machine learning benchmarks for measuring the performance of different models, including existing approaches (Platt Scaling, Isotonic Regression) and Smooth Isotonic Regression (SIR), Adaptive Calibration for Logistic Regression (AC-LR) and Temporal Maximum Margin Markov Networks (TM3N) developed in this thesis.

The results demonstrated consistent and comparable results of these models developed in the thesis. Specifically, SIR and AC-LR showed improved *calibration* ability compared with the other methods while retaining their discriminative power. The SIR and AC-LR outperforms the others in terms of the *calibration* measurement using Hosmer and Lemeshow goodness-of-fit test. The *discrimination* abilities of these models are close. Actually, Platt Scaling, Isotonic Regression and Smooth Isotonic Regression were at the same level. AC-LR did slightly better in discriminating between positive and negative cases thanks to its adaptive nature. For multiple variables co-estimation problem, TM3N demonstrated superior performance across various experiments involving data generated by a Linear Dynamic System, simulated by the BioWar disease outbreak engine and collected for building occupancy detection. All these experiments indicated that models developed in the thesis are generalizable to other data besides motivational examples.

10.6 Conclusion

In this chapter, I accessed the generalizability of the models by evaluating them using a wide range of datasets. Comparison of the models included both state-of-the-art approaches (Platt Scaling, Isotonic Regression, Conditional Random Fields and Hidden Markov Model) and those developed in the thesis, e.g., Smooth Isotonic Regression (SIR), Adaptive Calibration for Logistic Regression (AC-LR) and Temporal Maximum Margin Markov Networks (TM3N). The variety of experiences provided empirical evidence of models' applicability across different data. I evaluated the Area Under Curve (AUC) and HL test p-value of approaches that predict the probability of dichotomous outcomes, and recorded average accuracy of methods that predict multiple hidden states. The results indicated that models developed in the thesis are generalizable to other data besides motivational examples. SIR and AC-LR demonstrated better *calibration* ability

compared to the other methods. AC-LR also showed better *discrimination* ability in linearly non-separable cases. In variable co-estimation tasks, TM3N constantly outperformed other models in various experiments.

In summary, TM3N framework offers enough flexibility in modeling complex systems; and it can be easily generalized to other dynamic systems involving temporal, spatial and relational dependencies. The framework can be applied to cases involving multiple dependent manifestations, which are time correlated. For example, ICU patients management and object tracking in video sequences are good candidates for applying TM3N framework. More specifically, inputs to TM3N can be any features (continuous or categorical) from a set of instances while the outputs are some or all of their corresponding states. Note these states must be discrete values as TM3N is a discriminative model that cares about the accuracy of predictions. On the other hand, SIO model has good generalization ability. It can be directly used in calibrating outputs of binary outcomes from "uncalibrated" models like logistic regression, support vector machine and decision tree. SIR provides a better fitted model than PS while less overfitted results comparing to IR in most cases. Thanks to the consideration of model smoothness, it is capable of producing more reliable risk probabilities than existing approaches. The last model, AC-LR model can be easily generalized to other cases where a probabilistic outputs is preferred over a decision rule. The model can be directly applied to calibrate probabilistic estimates of binary outcomes. The model demonstrated best overall performance across various data from different sources. It owns a large amount of its success to the adaptive bandwidth selection using confidence intervals of predictions, which was merely studied by previous investigators.

The chapter ended up by demonstrating a dedicated online evaluation system, WEBCALIBSIS, to access the quality of predictive models across platforms.

Chapter 11

Contributions and Limitations

11.1 Summary

Recent progress in biomedicine and explosion in personal information have significantly increased the need for advanced biomedical informatics research. The next frontier in bioanalytical and detection science is developing predictive tools to support the decision system for emergency response.

Among the most important problems, bioterrorism related outbreak prediction and *calibration* for personalized medicine are two problems that are highly valued to the public and are heavily discussed. Accurate prediction can help decision makers in both problems to respond quickly and effectively. However, previous research cannot provide satisfactory solutions to these problems.

The first problem, bioterrorism related outbreak prediction, focuses on city-level predictions of disease outbreaks, which assists decision makers in responding more effectively. The major difficulty here is the mutual coupling between the temporal and relational factors. The second problem is about personalized medicine, which concentrates on individualized patient medicine and services so that caregivers can provide more specific diagnoses and therapies for patients. The difficulties in this problem are compounded by the accuracy of ranking and the reliability of probabilistic outputs. After briefly reviewing the relevant backgrounds, I started investigating the above-mentioned problems and developed solutions.

To model bioterrorism related outbreaks more faithfully, I exploited the temporal correlation concurrently with the relational dependency in simulated BioWar data. To this end, I developed a structured prediction model, Temporal Maximum Margin Markov Network (TM3N), to co-estimate multiple correlated

variables over time as a global optimization problem. That is, instead of predicting individual factors like "death rate" in the next time tick, TM3N predicts a network of outcome variables considering their mutual dependency over time. The complementary nature of temporal and relational information helps TM3N to achieve better accuracy and reliability in predicting bioterrorism related outbreaks. Furthermore, TM3N handles more general cases of learning noisy time series, can automatically adapt to new conditions, and can be achieved with tractable computational complexity.

For the *calibration* for personalized medicine problem, I focused on calibrating biomedical decision systems to identify the parameters of unique genetic populations. Previous theories were based on patient-diagnosis populations, i.e., if a patient has a Myocardial infarction (MI), the predictive model is then constructed for various treatments of patients with the same diagnosis. We know now, however, that medication right for one member of this diagnosis population may not work the same for all members. I believe predictions should be tailored and targeted to benefit patients in specific genetic groups, based on more detailed patient information. To this end, I studied an important but less-studied quality measurement for the probabilistic predictive model "*calibration*", which stratifies how outcomes affect various genetic population groups within a patient-diagnosis population. I designed a unified framework that combined two families of model quality measurements (i.e. *discrimination* and *calibration*). I demonstrated that models developed under this multi-targeted framework can achieve better performance of both quality measurements compared with single-targeted models.

Although formulations look different, my approaches to address both problems were data driven and based on more detailed information of targeted factors. With more detailed global models, my approaches maximized the likelihood of observed measurements and provided more reliable estimates of latent factors of interest.

Next, I evaluated the generalizability of developed methods. In most cases, training with larger numbers of samples may improve learning an algorithm's performance in testing when training data are representatives of the population. But too much training data could also hurt the testing performance due to overfitting and significantly increase the computational cost. Knowing the data scalability impact to methods developed in this thesis is useful in determining which, and how many data are required to construct reliable decision support predictive models.

Another difficulty in learning with the biomedical data that I investigated was biased labeling, i.e., a tiny number of confirmed hospital discharge errors vs. a large number of unconfirmed cases. This issue

is quite common in biomedical studies because the labeling procedure requires the expert's knowledge, staff time and even expensive laboratory test. Traditional approaches developed under supervised learning theory cannot handle data with biased labeling well due to the missing information of negative labels and extreme data unbalance. My study revealed that major deficiencies of previous approaches lies in their incapability of synthesizing information from different perspectives. To this end, I developed a structured biased Support Vector Machine model using feature correlations from abundant unlabeled cases concurrently with the positively labeled cases as a global optimization algorithm. Experiments on both synthetic and real datasets show performance advantages over conventional methods.

Finally, I demonstrated models' applicability across different databases. Specifically, two *calibration* methods, Smooth Isotonic Regression (SIO) and Adaptive Calibration for Logistic Regression (ACLR) were applied to nine different datasets while the structured inference approach, TM3N was evaluated on three temporal and relational correlated datasets.

11.2 Major Results

Through these investigations, I learned some points that are important to data-driven approaches. First of all, synthesizing information from different perspectives helps to reduce the ambiguity in estimating multiple outcome variables simultaneously. I developed the TM3N model to combine the power of methods based on only temporal correlation (HMM) and methods based on only relational dependency (CRF).

To verify the efficacy of TM3N, I first used synthetic data generated from a linear dynamic system, where synthesized temporal and relational factors were controlled by a trade-off parameter α. TM3N demonstrated superior performance at various levels of the trade-off parameter from 0.1 - 0.8. The results confirmed that combining complementary information helps to reduce the ambiguity existing in the individual perspective. I thus applied TM3N models to predict multiple states of correlated outcome variables in the BioWar simulation data. TM3N led the accuracy (69%) followed by HMM (65%), M3N (58%) and CRF (57%). The performance advantage of TM3N and its applicability were confirmed by the results of another real world experiment for building occupancy detection. Again, TM3N outperformed HMM, M3N and CRF and their average accuracies are 70%, 37%, 49% and 50%. Finally, I compared TM3N with HMM, M3N and CRF using the BioWar II data, which contained multiple five-year simulation periods of various sized agents (10% - 100%). The results showed TM3N scales well at increasing amount of training data and outperformed the

other models.

Another important finding is that considering *calibration* concurrently with *discrimination* can improve conventional single-target probabilistic models. Under a unified framework developed in Chapter 5, I implemented Smooth Isotonic Regression (SIO) and Adaptively Calibration for Logistic Regression (AC-LR). The SIO method introduced a smooth projection function to alleviate the problem of overfitting in Isotonic Regression, which is a state of the art *calibration* model. The AC-LR approach pushed the boundaries even further with the concept of adaptive binning based on input-specific information.

To verify the usefulness of both approaches, I compared them with a popular probabilistic model, Logistic Regression (LR) and existing *calibration* methods like Platt Scaling for Logistic Regression (LR-PS) and Isotonic Regression for Logistic Regression (LR-IS). The experiments using synthetic data showed that SIO has a superior *calibration* ability without decreasing the *discrimination* power. The real data experiments for SIO used a set of eight different data including Breast Cancer Gene Expression, Hospital Discharge Error and Pima Indian Diabetes. In general, SIO model demonstrated better *calibration* performance compared with LR, LR-PS and LR-IR in the Hosmer-Lemeshow goodness-of-fit test.

I conducted verification experiments for AC-LR in a similar way. For the synthetic data experiments, I showed intuitively how AC-LR is superior to existing approaches. I visualized 1D and 2D non-linear separable cases, which can be handled by AC-LR but not by the others. I conducted real data experiments using Hospital Discharge Error, Myocardial Infarction and Breast Cancer Gene Expression data. In the hospital discharge experiment, AC-LR passed HL-test at 0.05 significance level with a p-value of 0.349 while all the other methods failed. In addition, AC-LR even improved AUC from 0.704 (the best of previous approaches) to 0.717 showing joint optimization of *calibration* and *discrimination* improved single-target models in both perspectives. For the Myocardial Infarction dataset, AC-LR demonstrated its performance advantage over conventional methods again. AC-LR passed HL-test at 0.05 significance level with p-values of 0.645 and 0.246 for Sheffield data and Edinburgh data while LR, LR-PS and LR-IS failed. Improvements for *discrimination* were also prominent; AC-LR achieved an AUC of 0.880 and 0.863 comparing to 0.876 and 0.845 of LR, LR-PS and LR-IS for Sheffield data and Edinburgh data, respectively. Similarly, the performance of AC-LR led the competition of *discrimination* and *calibration* in the Breast Cancer Gene Expression data.

In terms of scalability, I tested LR-SIR and AC-LR along with LR, LR-PS and LR-IR at varying scales of training data, which included Breast Cancer Gene Expression Data (GSE2034, GSE2990, GSE3494),

Myocardial Infarction (Edinburgh, Sheffield), Hospital discharge Errors and three additional UCI machine learning repository data (Bankruptcy, PIMATR and HeightWeight) [69]. The variety of datasets provided empirical evidence of model's applicability across different data. I evaluated AUC, HL test p-value and time costs of all five approaches at ten different scales of training size. In general, LR-SIR and AC-LR demonstrated better *calibration* ability compared with the other methods without sacrificing their *discrimination* ability. In terms of computational cost, all methods are scalable, but LR-SIR is most expensive because it inserted an additional smoothing procedure to the LR-IR approach. The second and third expensive methods are LR-IR and LR-PS, respectively. Interestingly, the most balanced method, AC-LR turned out to be the least expensive.

11.3 Contributions

The goal of my thesis was to establish a coherent connection from machine learning techniques to life-saving biomedical applications, which often involve analyzing complex, expensive, and intensive health and clinical data in a timely manner. To this end, I developed new frameworks and implemented data-driven approaches to provide quick interpretation of observations, predict the consequence with high reliability and support the decision makers to respond effectively.

Specifically, I investigated two representative problems, the prediction of large-scale disease outbreaks (BioWar) and personalized clinical decision support (*calibration*), which cannot be well handled by conventional machine learning theory, which overlooks their biomedical characteristics. Although formulations look differently for these problems, my approaches to address both problems are data-driven based on more detailed information of targeted factors. With more detailed global models, my approaches maximize the likelihood of observed measurements and provide more reliable estimates of latent factors of interest.

I revealed that considering information from different sources simultaneously could improve the prediction accuracy of conventional time series models. I developed a new framework TM3N to optimize temporal coherence concurrently with relational dependence with tractable computation. As opposed to traditional approaches that predict states of outcome variables independently, my framework describes semantic correlations of heterogeneous variables and observations of individual variables in a global manner. The joint optimization reduces the ambiguity in estimating multiple outcome variables independently.

Furthermore, my framework offers more flexibility in modeling complex systems like disease outbreaks

and building occupancy; and it can be easily generalized to other dynamic systems involving temporal, spatial and relational dependencies. Synthetic experiments and real-world applications demonstrated TM3N's superior performance and wide adaptability.

Another contribution of my thesis is the development of a systematic framework for *discrimination* and *calibration*. In a principled way, I integrated two important model quality measurements: *discrimination* and *calibration*, which are traditionally considered separately. Through my investigation, I found that these two seemingly unrelated metrics are connected, and a well designed joint maximization algorithm can offer the best of both if they are optimized independently. This joint optimization using a combined objective function guards against learning degenerated model that performs well in one aspect but poorly in the other aspect.

Additionally, I found that the joint optimization can even improve *discrimination* performance without decreasing its *calibration* ability. This is because including a *calibration* term to conventional *discrimination*-based model brings an additional set of informative constraints that are not available to standard *discrimination*-based model. In many cases, such constraints effectively reduce the feasibility space and leave the solution with much fewer possibilities, i.e., $\min(|y_a| + |1 - y_b| + |y_c|)$ would significantly restrict the set of possible values for y_b, which satisfies $y_a < y_b < y_c$.

I developed an approach, Adaptive Calibration for Logistics Regression (AC-LR), to close the gap between traditional population learning theory and personalized medicine. That is, medication that is right for one member of a diagnosis population may not work the same for all members. In contrast, my AC-LR approach is tailored and targeted to benefit patients in specific genetic groups, based on more relevant patient information. As opposed to conventional methods constructed on the entire population of patients, my model used confidence intervals for individual predictions to construct a dynamic neighborhood for each patient. Thus, the predictions are based on more relevant information about the patient. Experiments on multiple clinical data demonstrated improved *calibration* and *discrimination* ability of this new model. Yet another advantage of AC-LR is that its computational cost is much lower compared to other methods that consider using a dynamic neighborhood.

I also investigated the "biased labeling" common to biomedical data, i.e., a tiny number of positively labeled cases vs. a large number of unlabeled cases, which gives difficulty to traditional supervised learning algorithms. My research showed that major deficiencies of previous approaches lies in their incapability of synthesizing information from different perspectives. I developed a structured biased Support Vector

Machine model using feature correlations from abundant unlabeled cases concurrently with the positively labeled cases as a global optimization algorithm. The method alleviates the underfitting problem with limited labels to a large extent by using alternative feature correlation constraints to regulate my model. Experiments on both synthetic and real datasets show performance advantages over conventional methods.

Through these investigations, I learned that model must be constructed from relevant information with consideration of different aspects of the observations simultaneously. First of all, global model must be targeted to fit the needs of prediction; second, multifaceted observations should be considered concurrently rather than independently.

11.4 Limitations

There are a few topics that I did not expect to make a major contribution of this thesis. The following are limitations of my works. Most of these limitations were indeed choices made to ensure the integrity of the thesis and interpretability of the results.

My experiments often involve large dataset, and carefully designed parallel computing could potentially improve the efficiency of my algorithm. For example, TM3N involves iterative steps of estimating most violated constraints and subgradient descent optimization that could be factorized into a series of parallel tasks. However, it is not trivial to design reliable paralleling, and the engineering overhead is very high. In addition, the parallel design may decrease the reproducibility of TM3N because it is already a very complex model. Considering these trade-offs, I decided not to include parallelization as a part of the thesis.

I developed discriminative models instead of generative ones for this thesis. I intended to design discriminative models because both disease outbreak and the *calibration* problems are related to decision making rather than the generation process for the observation. That is, these models tend to discriminate better with limited data but they cannot generate synthetic data or null hypothesis. I believe this compromise was necessary to ensure the focus of the thesis and its major contribution.

I limited my quantitative measurements to only a few metrics like accuracy, Area Under ROC Curve and Hosmer-Lemeshow goodness-of-the-fit test. They are measurements to access the model's prediction performance but there are more of these (e.g., R^2, F-test, Refinement and Resolution) in the statistical literature. The reason that I used only the most well known metrics was not to distract readers from understanding the fundamental aspects of main ideas by making aimless comparisons. Both AUC and the Hosmer-Lemeshow

test provided a simple one number summary of the probabilistic model under assessment.

The *calibration* for personalized medicine essentially involves constructing a model for each patient using the most relevant information, which had the potential to single out patterns of individual or small groups for analysis and evaluation. Although I used deidentified data with IRB approval and public records, I still think there could be important ethical and policy questions related to individualized prediction models. However, I decided to exclude the privacy question to maintain the integrity of the thesis.

Despite these limitations, the contribution of this thesis still holds, which is to provide useful learning tools to support decision systems for emergency response. I demonstrated various successful applications of the models developed in this thesis. An important insight gained from these studies is that models must be constructed from relevant information with consideration of different aspects of the observations simultaneously.

The main message is twofold: first, a global model must be tailored to fit the needs of prediction; second, it is risky to make an assumption independent of multifaceted observations. Thus, information of different perspectives should be synthesized rather than treated independently.

11.5 Future Works

The challenges in biomedical problems addressed in this thesis, present interesting extensions for future research.

Structured learning that synthesizing multifaceted information has received limited attention in many biomedical applications, e.g., in real time ICU risk estimation and long-term care for elderly. This thesis demonstrated the successful application of co-estimation of multiple outcome variables of interest in bioterrorism related diseases outbreaks. However, the potential of co-estimation for generally addressing temporal and relational dependent problems in biomedical research has yet to be realized. A reality gap between the techniques and the applications is belief in the ability of capturing long-range correlations over time and across regions. It is often infeasible for computer scientists to exhaustively search for the best solution within the entire feasible space while clinicians have no idea how to guide the algorithm towards meaningful optimization. This problem is largely due to the lack of consideration of domain-specific knowledge and infeasibility of including it in today's biomedical learning algorithm. The combination of co-estimation techniques with domain-specific priors is a promising direction for future research.

Regarding predictions in personalized medicine, a major dilemma is what to include and what not to include. Including too little information would cause model underfitting while including more but irrelevant information leads to overfitted models. The ideal solution would be testing every possible combinations of the observations in model construction, which is unfortunately not computationally feasible. This is an open problem and different applications might need different heuristics for determining the "relevancy." I demonstrated the success of an adaptive learning model, AC-LR, that tailors predictions towards more individual levels in clinical applications like hospital discharge error prediction. There are more to explore along this direction. One possible extension of my approach would be measuring distances between pairs of samples using meaningful kernels, and outputs predictions along with confidence intervals. Such non-parametric methods, providing reliability in addition to probability of individuals, have the potential to outperform their parametric counterparts and better address the prediction problems in personalized medicine. Another possible direction of research is to explore the sparsity pattern of individual patients based on more comprehensive information including genotype-phenotype correlation, social network profiles, and family disease history. A major challenge of modeling these apparently different areas of information is the ability to synthesize them meaningfully while not overfitting the data.

Bibliography

[1] Web-based calculator for roc curves, http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html. [Accessed Aug. 8, 2010]. 10.4

[2] D. H. Abdel-Qader, L. Harper, J. A. Cantrill, and M. P. Tully. Pharmacists' interventions in prescribing errors at hospital discharge: an observational study in the context of an electronic prescribing system in a UK teaching hospital. *Drug Saf*, 33:1027–1044, Nov 2010. 1, 1.1

[3] G. A. Ackerman. It is hard to predict the future: the evolving nature of threats and vulnerabilities. *Rev. - Off. Int. Epizoot.*, 25:353–360, Apr 2006. 4.1

[4] N. S. Altman. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 46(3):175, August 1992. 7.3.4

[5] E. Amir, O. C. Freedman, B. Seruga, and D. G. Evans. Assessing women at high risk of breast cancer: a review of risk assessment models. *J Natl Cancer Inst*, 102(10):680–691, 2010. 5

[6] Analyze-it. http://www.analyse-it.com/products/method_evaluation/. [Accessed Aug. 8, 2010]. 10.4

[7] A. Anand, G. Pugalenthi, G. B. Fogel, and P. N. Suganthan. An approach for classification of highly imbalanced data using weighting and undersampling. *Amino Acids*, 39:1385–1391, Nov 2010. 9.3

[8] R. Anand, K. G. Mehrotra, C. K. Mohan, and S. Ranka. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Trans Neural Netw*, 4:962–969, 1993. 9.3

[9] M. Ancukiewicz, D. M. Finkelstein, and D. A. Schoenfeld. Modelling the relationship between continuous covariates and clinical events using isotonic regression. *Stat Med*, 22:3151–3159, Oct 2003. 5, 6.2

[10] Ioannis N Athanasiadis and Pericles A Mitkas. Supporting the decision-making process in environmental monitoring systems with knowledge discovery techniques. In *Knowledge Discovery for Environmental Management*, volume Workshop I of *Knowledge-based Services for the Public Sector Symposium*, pages 1–12. KDnet, 2004. 5

[11] M Ayati. A unified perspective on decision making and decision support systems. *Information Processing & Management*, 23(6):616–628, 1987. 5

[12] T. Ayer, O. Alagoz, J. Chhatwal, J. W. Shavlik, C. E. Kahn, and E. S. Burnside. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. *Cancer*, 116:3310–3321, Jul 2010. 3.2, 3.2.1, 5, 6

[13] S. G. Baker and D. J. Sargent. Designing a Randomized Clinical Trial to Evaluate Personalized Medicine: A New Approach Based on Risk Prediction. *J Natl Cancer Inst*, Nov 2010. 1.1, 1.3

[14] N. Balluerka, A. Gorostiaga, J. Gomez-Benito, and M. D. Hidalgo. Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema*, 22:1018–1025, Nov 2010. 7

[15] I. Barman, C. R. Kong, N. C. Dingari, R. R. Dasari, and M. S. Feld. Development of Robust Calibration Models Using Support Vector Machines for Spectroscopic Monitoring of Blood Glucose. *Anal Chem*, Nov 2010. 1.1

[16] L. Bartova, A. Berger, and L. Pezawas. Is there a personalized medicine for mood disorders? *Eur Arch Psychiatry Clin Neurosci*, 260 Suppl 2:S121–126, Nov 2010. 1.3

[17] G Baudat and F Anouar. Kernel-based methods and function approximation. In *Neural Networks 2001 Proceedings IJCNN 01 International Joint Conference on*, volume 2, pages 1244—-1249 vol.2, 2001. 4

[18] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41, 1970. 3.1.1.2

[19] P. Berche. Scientific progress and new biological weapons. *Med Sci (Paris)*, 22:206–211, Feb 2006. 1.1, 4.1

[20] E. V. Bernstam, J. W. Smith, and T. R. Johnson. What is biomedical informatics? *J Biomed Inform*, 43:104–110, Feb 2010. 1

[21] R. Blagus and L. Lusa. Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 11:523, Oct 2010. 1.1, 9

[22] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 9.4.2.2

[23] Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950. 10.4

[24] J. C. Brillman, T. Burr, D. Forslund, E. Joyce, R. Picard, and E. Umland. Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance. *BMC Med Inform Decis Mak*, 5:4, 2005. 1.1, 4.1

[25] A Buhot and Mirta B Gordon. Learning properties of Support Vector Machines. page 4, 1998. 4

[26] Oleg Burdakov, Anders Grimvall, and Oleg Sysoev. Data preordering in generalized PAV algorithm for monotonic regression. (1):1–21. 5.2, 7.2

[27] Shaun Burke. Regression and Calibration, 2001. 5

[28] J. Callen, J. McIntosh, and J. Li. Accuracy of medication documentation in hospital discharge summaries: A retrospective analysis of medication transcription errors in manual and electronic discharge summaries. *Int J Med Inform*, 79:58–64, Jan 2010. 1, 1.1

[29] Y. G. Cao, J. J. Cimino, J. Ely, and H. Yu. Automatically extracting information needs from complex clinical questions. *J Biomed Inform*, Jul 2010. 1, 1.1

[30] Kathleen Carley, Neal Altman, and Boris Kaminsky. BioWar : A City-Scale Multi-Agent Network Model of Weaponized Biological Attacks. *CASOS Technical Report*, (Cdc), 2004. 2.1, 4

[31] Kathleen Carley, D.B. Fridsma, E. Casman, a. Yahja, N. Altman, B. Kaminsky, and D. Nave. BioWar: scalable agent-based model of bioattacks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 36(2):252–265, March 2006. 2.1, 4

[32] Kathleen Carley, Douglas Fridsma, Elizabeth Casman, Jack Chang, Boris Kaminsky, Demian Nave, and Alex Yahja. BioWar : Scalable Multi-Agent Social and Epidemiological Simulation of Bioterrorism Events. *NAACSOS Conference, IEEE SMCA03*, (412), 2003. 2.1, 4

[33] O. Carugo. Detailed estimation of bioinformatics prediction reliability through the Fragmented Prediction Performance Plots. *BMC Bioinformatics*, 8:380, 2007. 2.5, 3.2.2

[34] C. Castillo-Salgado. Trends and directions of global public health surveillance. *Epidemiol Rev*, 32:93–109, Apr 2010. 4

[35] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 9.5.2.1

[36] B. Chen, M. Chang, and C. Lin. Load forecasting using support vector machines: A study on eunite competition 2001. *IEEE Transactions on Power Systems*, 19(4):1821 1830, 2006. 2.1

[37] L. Chen, A. M. Tonkin, L. Moon, P. Mitchell, A. Dobson, G. Giles, M. Hobbs, P. J. Phillips, J. E. Shaw, D. Simmons, L. A. Simons, A. P. Fitzgerald, G. De Backer, and D. De Bacquer. Recalibration and validation of the SCORE risk chart in the Australian population: the AusSCORE chart. *Eur J Cardiovasc Prev Rehabil*, 16:562–570, Oct 2009. 3.2, 5

[38] Li-chiou Chen, Kathleen Carley, Boris Kaminsky, Tiffany Tummino, Elizabeth Casman, Douglas Fridsma, and

Alex Yahja. Aligning Simulation Models of Biological Attacks Contact : Aligning Simulation Models of Biological Attacks. *NAACSOS*, 2004. 2.1, 4

[39] Li-chiou Chen, Boris Kaminsky, Tiffany Tummino, Kathleen M, Elizabeth Casman, Douglas Fridsma, and Alex Yahja. Aligning Simulation Models of Smallpox Outbreaks. *Proceedings of the Second Symposium on Intelligence and Security Informatics*, 2004. 4, 4.8

[40] S. Chen, H. He, and E. A. Garcia. RAMOBoost: Ranked Minority Oversampling in Boosting. *IEEE Trans Neural Netw*, 21:1624–1642, Oct 2010. 9.3

[41] Yueh-Yun Chi and Xiao-Hua Zhou. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by Pencina et al., . *Statistics in medicine*, 27(2):182–4, January 2008. 3.2.1, 5, 5.2, 5.6

[42] Jainn-Shiun Chiu, Fu-Chiu Yu, and Yu-Chuan Li. Discrimination and calibration are concurrently required for model comparison., 2006. 5

[43] Ariel Cintrón-Arias, Carlos Castillo-Chávez, Luís M A Bettencourt, Alun L Lloyd, and H T Banks. The estimation of the effective reproductive number from disease outbreak data. *Mathematical biosciences and engineering MBE*, 6(2):261–282, 2009. 4

[44] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler. Learning from imbalanced data in surveillance of nosocomial infection. *Artif Intell Med*, 37:7–18, May 2006. 9

[45] Ira Cohen and Moises Goldszmidt. Properties and benefits of calibrated classifiers. In *8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 125–136, 2004. 5, 5.2, 7.2

[46] N. J. Cox, S. E. Tamblyn, and T. Tam. Influenza pandemic planning. *Vaccine*, 21:1801–1803, May 2003. 1.1, 4, 4.1

[47] J. S. Crane, J. D. McCluskey, G. T. Johnson, and R. D. Harbison. Assessment of community healthcare providers ability and willingness to respond to emergencies resulting from bioterrorist attacks. *J Emerg Trauma Shock*, 3:13–20, Jan 2010. 1.1

[48] Robert L. Cross, Andrew Parker, and Rob Cross. *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations*. Harvard Business School Press, June 2004. 4

[49] R. Cruz-Cano, D. S. Chew, C. Kwok-Pui, and L. Ming-Ying. Least-Squares Support Vector Machine Approach to Viral Replication Origin Prediction. *INFORMS J Comput*, 22:457–470, Jun 2010. 1.1

[50] Sr. D'Agostino, R. B., S. Grundy, L. M. Sullivan, and P. Wilson. Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA*, 286(2):180–7, 2001. 5

[51] Thomas H Davenport. Make better decisions. *Harvard Business Review*, 87(11):117–118, 120–123, 134, 2009. 5

[52] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *International Conference on Computer Vision (ICCV)*, pages 229 – 236, 2009. 3.1.2, 4.3

[53] G. Di Minno and M. Mancini. Measuring plasma fibrinogen to predict stroke and myocardial infarction. *Arteriosclerosis*, 10:1–7, 1990. 1.1

[54] G A Diamond. What price perfection? Calibration and discrimination of clinical prediction models. *Journal of Clinical Epidemiology*, 45(1):85–89, 1992. 5

[55] Joachim Diederich, Jörg Kindermann, Ella Leopold, and Gerhard Paass. Authorship Attribution with Support Vector Machines. *Applied Intelligence*, 19(1):109–123, 2003. 4

[56] Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Physical Review Letters*, 82(14):4, 1998. 4

[57] V. Digalakis, J. R. Rohlicek, and M. Ostendorf. Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(4), 1993. 3.1.1.1

[58] Pedro Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple Bayesian classi er. *Machine Learning*, 29(2/3):103–130, 1997. 7.2

[59] Robert El-Kareh and Xiaoqian Jiang. Hospital Discharge Error Prediction. *DBMI-Tech-Report 2010-002, UCSD*, 2010. 2.4, 6.1.2, 7.1.1, 8.1.1, 9.1.1, 10.1.3

[60] Charles Elkan. Logistic Regression and Stochastic Gradient Training. *UCSD Technical Report*, pages 1–10, 2007. 5.2

[61] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. *Proceeding of international conference on Knowledge discovery and data mining (KDD)*, pages 213–220, 2008. 9.3

[62] Tom Fawcett and Alexandru Niculescu-Mizil. PAV and the ROC convex hull. *Machine Learning*, 68(1):97–106, 2007. 5.2, 7

[63] J Feng and P Williams. The generalization error of the symmetric and scaled support vector machines. *IEEE Transactions on Neural Networks*, 12(5):1255–1260, 2001. 4

[64] A. G. Fiks and M. E. Jimenez. The promise of shared decision-making in paediatrics. *Acta Paediatr*, 99:1464–1466, Oct 2010. 1.1

[65] Robert Fildes. Forecasting structural time series models and the kalman filter. *International Journal of Forecast-*

ing, 8(4):635–635, December 1992. 4.3

[66] Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the Youden Index and its associated cutoff point. *Biometrical journal. Biometrische Zeitschrift*, 47(4):458–72, August 2005. 10.4

[67] D. Foster. Asymptotic calibration. *Biometrika*, 85(2):379–390, June 1998. 5.2

[68] John Fox, David Glasspool, and Jonathan Bury. Quantitative and qualitative approaches to reasoning under uncertainty in medical decision making. In *Artificial Intelligence in Medicine*, volume 2101 of *Lecture Notes in Computer Science*, pages 272–282. Springer Berlin / Heidelberg, 2001. 4.4

[69] A. Frank and A. Asuncion. Uci machine learning repository, 2010. 6.1, 6.1.3, 6.1.5, 6.1.6, 8.2, 10, 10.4, 11.2

[70] F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980. 6.3

[71] H. Fujikawa. Development of a new logistic model for microbial growth in foods. *Biocontrol Sci*, 15:75–80, Sep 2010. 7

[72] R Fuller. Linear dynamic systems. *Wiley Online Library*, Jan 2008. 3.1.1

[73] Andrew B Gardner, Hayden Hall, and Abba M Krieger. One-Class Novelty Detection for Seizure Analysis from Intracranial EEG. *Journal of Machine Learning Research*, 7:1025 – 1044, 2006. 9.3

[74] Y. Garten, A. Coulet, and R. B. Altman. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics*, 11:1467–1489, Oct 2010. 1, 1.1

[75] Z. Ghahramani and G. Hinton. Parameter estimation for linear dynamical systems. *Tech. Report, Department of Computer Science, University of Toronto*, 1996. 3.1.1, 3.1.1.1

[76] Laura M Glass and Robert J Glass. Social contact networks for the spread of pandemic influenza in children and teenagers. *BMC Public Health*, 8(61):61, 2008. 4

[77] Isabel Gordo, M Gabriela M Gomes, Daniel G Reis, and Paulo R A Campos. Genetic Diversity in the SIR Model of Pathogen Evolution. *PLoS ONE*, 4(3):8, 2009. 4

[78] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *International Conference on Computer Vision (ICCV)*, pages 1 – 8, 2009. 3.1.2, 4.3

[79] R A Greenes. *Clinical decision support: the road ahead*. Academic Press, 2006. 5

[80] Kamalakar Gulukota. Immunoinformatics in personalized medicine. *Novartis Foundation Symposium*, 254:43–50; discussion 50–56, 98–101, 250–252, 2003. 3.2, 6

[81] J D Habbema, J Hilden, and B Bjerregaard. The measurement of performance in probabilistic diagnosis. V.

General recommendations. *Methods of Information in Medicine*, 20(2):97–100, 1981. 5

[82] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009. 9.5.2.1

[83] J. A. Hanley and B. J. Mcneil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, April 1982. 3.2, 3.2.1, 6.4, 7.4

[84] K.A. Heller, K.M. Svore, A.D. Keromytis, and S.J. Stolfo. One class support vector machines for detecting anomalous windows registry accesses. In *Proc. of the workshop on Data Mining for Computer Security*, 2003. 9.3

[85] A. Herment, M. Lefort, N. Kachenoura, A. De Cesare, V. Taviani, M. J. Graves, C. Pellot-Barakat, F. Frouin, and E. Mousseaux. Automated estimation of aortic strain from steady-state free-precession and phase contrast MR images. *Magn Reson Med*, Nov 2010. 1, 1.1

[86] C. Hernandez Prats, A. Mira Carrio, E. Arroyo Domingo, M. Diaz Castellano, L. Andreu Gimenez, and M. I. Sanchez Casado. Conciliation discrepancies at hospital discharge. *Aten Primaria*, 40:597–601, Dec 2008. 1, 1.1

[87] K. W. Hong and B. Oh. Overview of personalized medicine in the disease genomic era. *BMB Rep*, 43:643–648, Oct 2010. 1.3, 6

[88] X. Hong, S. Chen, and C. J. Harris. A kernel-based two-class classifier for imbalanced data sets. *IEEE Trans Neural Netw*, 18:28–41, Jan 2007. 9.3

[89] T Hosaka, T Kobayashi, and N Otsu. Image Segmentation Using MAP-MRF Estimation and Support Vector Machine. *Interdisciplinary Information Sciences*, 13(1):33–42, 2007. 4

[90] D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980, 1997. 3.2.2, 5.3.2, 6.4, 7.4, 10.4

[91] Bo Jiang, Michael Q Zhang, and Xuegong Zhang. OSCAR: one-class SVM for accurate recognition of cis-elements. *Bioinformatics (Oxford, England)*, 23(21):2823–8, November 2007. 9.3

[92] Xiaoqian Jiang, Bing Dong, and Latanya Sweeney. Temporal maximum margin markov network. In *ECML/PKDD (1)*, volume 6321 of *Lecture Notes in Computer Science*, pages 587–600. Springer, 2010. 1

[93] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Adaptive calibration for logistic regression. *Nature Biotechnology*, 2010 (under review). 1

[94] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Smooth isotonic regression. *AMIA Summit on Clinical Research Informatics*, 2010 (under review). 1

[95] C. Jonquet, M. A. Musen, and N. H. Shah. Building a biomedical ontology recommender web service. *J Biomed Semantics*, 1 Suppl 1:S1, 2010. 1

[96] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960. 3.1, 3.1.1, 3.1.1.1, 3.1.1.1, 4.3

[97] K. Kaplan. Burdens of biodefence. *Nature*, 465:386–387, May 2010. 4

[98] R L Kennedy, a M Burton, H S Fraser, L N McStay, and R F Harrison. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. *European heart journal*, 17(8):1181–91, August 1996. 2.3, 7.1.2, 7.4.2.2, 8.1.2, 10.1.4

[99] A. S. Khan, A. Fleischauer, J. Casani, and S. L. Groseclose. The next public health revolution: public health information fusion and social networks. *Am J Public Health*, 100:1237–1242, Jul 2010. 4

[100] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Supervised neural network modeling: an empirical investigation into learning from imbalanced data with labeling errors. *IEEE Trans Neural Netw*, 21:813–830, May 2010. 9.3

[101] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid, 1996. 6.1.4, 8.1.4

[102] S. Konovalov, M. Scotch, L. Post, and C. Brandt. Biomedical informatics techniques for processing and analyzing web blogs of military service members. *J. Med. Internet Res.*, 12:e45, 2010. 1, 1.1

[103] A. A. Kramer and J. E. Zimmerman. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.*, 35:2052–2056, Sep 2007. 3.2.2, 5.3.2

[104] C. A. Kulikowski and A. Geissbuhler. Health informatics: building capacity worldwide. Editorial. *Yearb Med Inform*, pages 6–7, 2010. 1.1

[105] S. Laberge, M. Albert, and B. D. Hodges. Perspectives of clinician and biomedical scientists on interdisciplinary health research. *CMAJ*, 181:797–803, Nov 2009. 1

[106] LABMRMC. `http://www-radiology.uchicago.edu/krl/KRL_ROC/software_index.htm`. [Accessed Aug. 8, 2010]. 10.4

[107] John Lafferty. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. *Computer*, 1999. 3.1, 3.1.2, 4, 4, 4.3

[108] Thomas a Lasko, Jui G Bhagwat, Kelly H Zou, and Lucila Ohno-Machado. The use of receiver operating characteristic curves in biomedical informatics. *Journal of biomedical informatics*, 38(5):404–15, October 2005. 3.2.1

[109] Chung Hong Lee and Hsin Chang Yang. A text mining approach for measuring semantic relatedness using support vector machines. In *The 8th World Multi Conference on Systemics Cybernetics and Informatics*, volume 4, pages 320–323, 2004. 4

[110] S. Lemeshow and D. W. Hosmer. Logistic regression analysis: applications to ophthalmic research. *Am. J. Ophthalmol.*, 147:766–767, May 2009. 7

[111] W. D. Leslie, L. M. Lix, H. Johansson, A. Oden, E. McCloskey, and J. A. Kanis. Independent clinical validation of a Canadian FRAX tool: fracture prediction and model calibration. *J. Bone Miner. Res.*, 25:2350–2358, Nov 2010. 3.2, 3.2

[112] B. E. Lewis. Narrative Medicine and Healthcare Reform. *J Med Humanit*, Oct 2010. 1.3

[113] Rui Li, Tai-Peng Tian, and Stan Sclaroff. Simultaneous Learning of Nonlinear Manifold and Dynamical Models for High-dimensional Time Series. *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, October 2007. 4.3

[114] D. V. Lim, J. M. Simpson, E. A. Kearns, and M. F. Kramer. Current and developing technologies for monitoring agents of bioterrorism and biowarfare. *Clin. Microbiol. Rev.*, 18:583–607, Oct 2005. 1.1, 4

[115] H. C. Lin, H. C. Wu, C. H. Chang, T. C. Li, W. M. Liang, and J. Y. Wang. A real time online assessment system with modelized architecture on clinical infometrics for patient reported outcomes of prostate cancer. *Comput Methods Programs Biomed*, Nov 2010. 1, 1.1

[116] Jessica Lin, Eamonn Keogh, Stefano Lonardi, Jeffrey P. Lankford, and Donna M. Nystrom. Visually mining and monitoring massive time series. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, page 460, 2004. 4.4

[117] J. Lipson, J. Bernhardt, U. Block, W. R. Freeman, R. Hofmeister, M. Hristakeva, T. Lenosky, R. McNamara, D. Petrasek, D. Veltkamp, and S. Waydo. Requirements for calibration in noninvasive glucose monitoring by Raman spectroscopy. *J Diabetes Sci Technol*, 3:233–241, Mar 2009. 3.2

[118] Bing Liu, Y. Dai, Xiaoli Li, W.S. Lee, and P.S. Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining, (ICDM)*, pages 179–186, 2003. 9.3, 9.4.1, 9.5

[119] K. Liu, W. R. Hogan, and R. S. Crowley. Natural Language Processing methods and systems for biomedical ontology learning. *J Biomed Inform*, Jul 2010. 1

[120] Yang Liu and Michael G Madden. One-Class Support Vector Machine Calibration Using Particle Swarm Optimisation. *Machine Learning*, (August):91–100, 2007. 9.3

[121] Larry M. Manevitz and Malik Yousef. One-Class SVMs for Document Classification. *Journal of Machine*

Learning Research, 2(2):139–154, May 2002. 9.3

[122] S. Mani. Note on Friedman's 'fundamental theorem of biomedical informatics'. *J Am Med Inform Assoc*, 17:614, 2010. 1, 1.1

[123] C. Marrocco, M. Molinara, and F. Tortorella. On Linear Combinations of Dichotomizers for Maximizing the Area Under the ROC Curve. *IEEE Trans Syst Man Cybern B Cybern*, Aug 2010. 3.2

[124] M. E. Matheny, L. Ohno-Machado, and F. S. Resnic. Discrimination and calibration of mortality risk prediction models in interventional cardiology. *J Biomed Inform*, 38:367–375, Oct 2005. 5

[125] MedCalc. `http://www.medcalc.be`. [Accessed Aug. 8, 2010]. 10.4

[126] G. B. Melton. Biomedical and health informatics for surgery. *Adv Surg*, 44:117–130, 2010. 1, 1.1

[127] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol*, 10:70, 2010. 1

[128] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. U.S.A.*, 102:13550–13555, Sep 2005. 2.2, 7.1.3, 8.1.5, 9.1.2, 9.5.3, 10.1.1, 10.1.6

[129] T M Mitchell. Generative and Disciminative classifiers: Naive Bayes and Logistic Regression learning classifiers based on Bayes Rule. *Machine Learning*, pages 1–17, 2010. 3.2

[130] T. Morton-Jones, P. Diggle, L. Parker, H. O. Dickinson, and K. Binks. Additive isotonic regression models in epidemiology. *Stat Med*, 19:849–859, Mar 2000. 6.2

[131] F. A. Murphy. Emerging zoonoses: the challenge for public health and biodefense. *Prev. Vet. Med.*, 86:216–223, Sep 2008. 1, 1.1

[132] Kevin P. Murphy, Yair Weiss, and Michael I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of Uncertainty in AI*, pages 467–475, 1999. 4.5.4.1, 2

[133] N. J. Nagelkerke, S. Moses, F. A. Plummer, R. C. Brunham, and D. Fish. Logistic regression in case-control studies: the effect of using independent as dependent variables. *Stat Med*, 14:769–775, Apr 1995. 1.1, 7

[134] A. Nalca and D. K. Nichols. Rabbitpox: A model of airborne transmission of smallpox. *J Gen Virol*, Oct 2010. 4

[135] B. Neelon and D. B. Dunson. Bayesian isotonic regression and trend analysis. *Biometrics*, 60:398–406, Jun 2004. 6.2

[136] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. *Proceed-*

ings of the 22nd international conference on Machine learning (ICML), (1999):625–632, 2005. 5.2, 6.4.2, 7

[137] R. W. Niska. Hospital collaboration with public safety organizations on bioterrorism response. *Prehosp Emerg Care*, 12:12–17, 2008. 1.1

[138] S. Oh, M. S. Lee, and B. T. Zhang. Ensemble Learning with Active Example Selection for Imbalanced Biomedical Data Classification. *IEEE/ACM Trans Comput Biol Bioinform*, Sep 2010. 1.1

[139] Sang Min Oh, Ananth Ranganathan, James M Rehg, and Frank Dellaert. A Variational inference method for Switching Linear Dynamic Systems Need for Variational methods Variational method. *GVU Technical Report;GIT-GVU-05-16*, 2005. 4.3

[140] Melanie Osl, Stephan Dreiseitl, Jihoon Kim, Kiltesh Patel, and Lucila Ohno-Machado. Effect of Data Combination on Predictive Modeling : A Study Using Gene Expression Data Division of BioMedical Informatics. *AMIA Annual Symposium proceedings*, 2010. 2.2, 3.2, 5.1, 5.5, 6.1.1, 7.1.3, 7.1.3, 7.4.2.3, 8.1.5, 9.1.2, 10.1.1, 10.1.6

[141] Melanie Osl, Lucila Ohno-Machado, Christian Baumgartner, Bernhard Tilg, and Stephan Dreiseitl. Improving calibration of logistic regression models by local estimates. *AMIA Annual Symposium proceedings*, pages 535–539, January 2008. 7.2

[142] O. Ovaskainen, J. Hottola, and J. Siitonen. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, 91:2514–2521, Sep 2010. 7

[143] S. V. Pakhomov, P. L. Hanson, S. S. Bjornsen, and S. A. Smith. Automatic classification of foot examination findings using clinical notes and machine learning. *J Am Med Inform Assoc*, 15:198–202, 2008. 1, 1.1

[144] P. R. Payne, P. J. Embi, and J. Niland. Foundational biomedical informatics research in the clinical and translational science era: a call to action. *J Am Med Inform Assoc*, 17:615–616, Nov 2010. 1, 1.1

[145] A. S. Pena. Personalized medicine: inevitable. *Rev Esp Enferm Dig*, 102:573–576, Oct 2010. 1.3, 5

[146] M. S. Pepe. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. *American Journal of Epidemiology*, 159(9):882–890, May 2004. 3.2.1

[147] J Platt. Probabilistic Output for Support Vector Machines and Comarisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999. 3.2, 5, 5.2, 5.4.2.2, 6, 6.2, 7.2

[148] R. M. Plenge and S. L. Bridges. Personalized medicine in rheumatoid arthritis: Miles to go before we sleep. *Arthritis Rheum*, Nov 2010. 1.3

[149] Philip M Polgreen, Troy Leo Tassier, Sriram Venkata Pemmaraju, and Alberto Maria Segre. Prioritizing healthcare worker vaccinations on the basis of social network analysis. *Infection control and hospital epidemiology the official journal of the Society of Hospital Epidemiologists of America*, 31(9):893–900, 2010. 4

[150] Sarah D Pressman, Sheldon Cohen, Gregory E Miller, Anita Barkin, Bruce S Rabin, and John J Treanor. Loneliness, social network size, and immune response to influenza vaccination in college freshmen. *Health psychology official journal of the Division of Health Psychology American Psychological Association*, 24(3):297–306, 2005. 4

[151] Xian Qian, Xiaoqian Jiang, Qi Zhang, Xuanjing Huang, and Lide Wu. Sparse higher order conditional random fields for improved sequence labeling. In *International Conference on Machine Learning (ICML*. 3.1.2, 4.3

[152] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989. 3.1, 3.1.1, 4, 4, 4.3

[153] P. Radivojac, N. V. Chawla, A. K. Dunker, and Z. Obradovic. Classification and knowledge discovery in protein databases. *J Biomed Inform*, 37:224–239, Aug 2004. 9

[154] L. D. Rotz and J. M. Hughes. Advances in detecting and responding to threats from bioterrorism and emerging infectious disease. *Nat. Med.*, 10:S130–136, Dec 2004. 1, 1.1, 4

[155] D. H. Roukos. Next-generation sequencing and epigenome technologies: potential medical applications. *Expert Rev Med Devices*, 7:723–726, Nov 2010. 1.3

[156] Sunita Sarawagi and Rahul Gupta. Accurate max-margin training for structured output spaces. In *Proceedings of the 25th international conference on Machine learning (ICML)*, pages 888–895, New York, NY, USA, 2008. ACM. 4.3

[157] A. J. Scheen and H. Kulbertus. Interheart: nine risk factors predict nine out of ten myocardial infarctions. *Rev Med Liege*, 59:676–679, Nov 2004. 1.1

[158] C. Schlotelburg, T. Becks, and T. Stieglitz. [Biomedical engineering today : An overview from the viewpoint of the German Biomedical Engineering Society]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, 53:759–767, Aug 2010. 1

[159] Bernhard Scholkopf. The Kernel Trick for Distances. In *Neural Information Processing Systems*, volume 13, pages 301–307. Microsoft Research, The MIT Press, 2000. 4

[160] S. J. Schonfeld, D. Pee, R. T. Greenlee, P. Hartge, Jr. J. V. Lacey, Y. Park, A. Schatzkin, K. Visvanathan, and R. M. Pfeiffer. Effect of changing breast cancer incidence rates on the calibration of the gail model. *J Clin Oncol*, 28(14):2411–2417, 2010. 5

[161] Pengliang Shi and Michael Small. Modelling of SARS for Hong Kong. *The Lancet*, (2):6, 2003. 4

[162] Jonathon Shlens. A tutorial on principal component analysis, December 2005. 9.4.2.2

[163] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the em algorithm.

Journal of Time Series Analysis, 3(4), 1982. 3.1.1

[164] Tobias Sing, Oliver Sander, and Niko Beerenwinkel. BIOINFORMATICS ROCR : Visualizing classifier performance in R. *Bioinformatics*, pages 2003–2004, 2005. 10.4

[165] S. K. Sinha, N. M. Laird, and G. M. Fitzmaurice. Multivariate logistic regression with incomplete covariate and auxiliary information. *J Multivar Anal*, 101:2389–2397, Nov 2010. 7

[166] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst*, 98:262–272, 2006. 2.2, 5.1, 5.5, 6.1.1, 7.1.3, 8.1.5, 10.1.1, 10.1.6

[167] W. W. Stead, J. R. Searle, H. E. Fessler, J. W. Smith, and E. H. Shortliffe. Biomedical Informatics: Changing What Physicians Need to Know and How They Learn. *Acad Med*, Aug 2010. 1, 1.1

[168] T. Steuber, P. Nurmikko, A. Haese, K. Pettersson, M. Graefen, P. Hammerer, H. Huland, and H. Lilja. Discrimination of benign from malignant prostatic disease by selective measurements of single chain, intact free prostate specific antigen. *J. Urol.*, 168:1917–1922, Nov 2002. 3.2

[169] B. Tadmor and B. Tidor. Interdisciplinary research and education at the biology-engineering-computer science interface: a perspective. *Drug Discov. Today*, 10:1183–1189, Sep 2005. 1

[170] Azzam F G Taktak, Antonio Eleuteri, Stephen P Lake, and Anthony C Fisher. A web-based tool for the assessment of discrimination and calibration properties of prognostic models. *Computers in Biology and Medicine*, 38(7):785–791, 2008. 3.2

[171] Y. Tang, Y. Q. Zhang, N. V. Chawla, and S. Krasser. SVMs modeling for highly imbalanced classification. *IEEE Trans Syst Man Cybern B Cybern*, 39:281–288, Feb 2009. 9.3

[172] Ben Taskar. Learning structured prediction models: A large margin approach. Stanford University, Ph.D. thesis, 2004. 3.1, 3.1.2, 4.3

[173] A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. Magnor, and D. Keim. Automated Analytical Methods to Support Visual Exploration of High-Dimensional Data. *IEEE Trans Vis Comput Graph*, Oct 2010. 1, 1.1

[174] R. Teramoto. Balanced gradient boosting from imbalanced data for clinical outcome prediction. *Stat Appl Genet Mol Biol*, 8:Article20, 2009. 9.3

[175] Tânia Tomé and Robert M Ziff. On the critical behavior of the Susceptible-Infected-Recovered (SIR) model on a square lattice. *Arxiv preprint arXiv10062129*, page 9, 2010. 4

[176] G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali. Statistical methods for the assessment of prognostic biomarkers(part II): calibration and re-classification. *Nephrol. Dial. Transplant.*, 25:1402–1405, May 2010. 3.2, 3.2.1, 5

[177] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005. 4.3

[178] Duong T.V., Phung D.Q., Bui H.H., and Venkatesh S. Human. Behavior recognition with generic exponential family duration modeling in the hidden semi-markov model. volume 3 of *Proceedings of the 18th international Conference on Pattern Recognition*, 2006. 4.3

[179] B. Van Calster and S. Van Huffel. Integrated discrimination improvement and probability-sensitive AUC variants. *Stat Med*, 29:318–319, Jan 2010. 3.2

[180] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. 4, 5.4.2

[181] S Vida. A computer program for non-parametric receiver operating characteristic analysis. *Computer methods and programs in biomedicine*, 40(2):95–101, June 1993. 3.2.1, 10.4

[182] Erik Volz and Lauren Ancel Meyers. Susceptible-infected-recovered epidemics in dynamic contact networks. *Proceedings of the Royal Society B Biological Sciences*, 274(1628):2925–2933, 2007. 4

[183] C. von Schacky. Cardiovascular disease prevention and treatment. *Prostaglandins Leukot. Essent. Fatty Acids*, 81:193–198, 2009. 1.1

[184] E. A. Wagar, M. J. Mitchell, K. C. Carroll, K. G. Beavis, C. A. Petti, R. Schlaberg, and B. Yasin. A review of sentinel laboratory performance: identification and notification of bioterrorism agents. *Arch. Pathol. Lab. Med.*, 134:1490–1503, Oct 2010. 4

[185] J. Wang, K. A. Do, S. Wen, S. Tsavachidis, T. J. McDonnell, C. J. Logothetis, and K. R. Coombes. Merging microarray data, robust feature selection, and predicting prognosis in prostate cancer. *Cancer Inform*, 2:87–97, 2006. 6.1.1, 8.1.5, 10.1.1, 10.1.6

[186] X. Wang and F. Li. Isotonic Smoothing Spline Regression. *Journal of Computational and Graphical Statistics*, 17(1):21–37, 2008. 6.2, 6.3, 6.3, 7.2

[187] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679, 2005. 2.2, 5.1, 5.5, 7.1.3

[188] K. Wei. Assessment of myocardial viability using myocardial contrast echocardiography. *Echocardiography*,

22:85–94, Jan 2005. 1.1

[189] Lloyd R. Welch. Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4), December 2003. 4.3

[190] E. D. Williamson, M. G. Duchars, and R. Kohberger. Predictive models and correlates of protection for testing biodefence vaccines. *Expert Rev Vaccines*, 9:527–537, May 2010. 1, 4

[191] T. E. Workman, M. Fiszman, J. F. Hurdle, and T. C. Rindflesch. Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information. *J Med Libr Assoc*, 98:273–281, Oct 2010. 1

[192] F. Xiao, C. C. Liao, K. C. Huang, I. J. Chiang, and J. M. Wong. Automated assessment of midline shift in head injury patients. *Clin Neurol Neurosurg*, 112:785–790, Nov 2010. 1.1

[193] F. Yaghouby, A. Ayatollahi, R. Bahramali, M. Yaghouby, and A. H. Alavi. Towards automatic detection of atrial fibrillation: A hybrid computational approach. *Comput Biol Med*, Nov 2010. 1, 1.1

[194] Q. Yan. Translational bioinformatics and systems biology approaches for personalized medicine. *Methods Mol. Biol.*, 662:167–178, 2010. 1.1, 1.3

[195] P. Yang, L. Xu, B. B. Zhou, Z. Zhang, and A. Y. Zomaya. A particle swarm based hybrid system for imbalanced medical data sampling. *BMC Genomics*, 10 Suppl 3:S34, 2009. 9, 9.3

[196] X. Ye, G. Beddoe, and G. Slabaugh. Automatic Graph Cut Segmentation of Lesions in CT Using Mean Shift Superpixels. *Int J Biomed Imaging*, 2010:983963, 2010. 1, 1.1

[197] L. Yeganova and W. Wilbur. Isotonic regression under lipschitz constraint. *Journal of Optimization Theory and Applications*, 141:429–443, 2009. 10.1007/s10957-008-9477-0. 6.2

[198] H. L. Yin and T. Y. Leong. A model driven approach to imbalanced data sampling in medical decision making. *Stud Health Technol Inform*, 160:856–860, 2010. 9.3

[199] A. Yoneyama. Survey of costs of laboratory tests aiming at revision of medical treatment fees and improvement of work. *Rinsho Byori*, 58:920–924, Sep 2010. 1

[200] Hwanjo Yu. Single-Class Classification with Mapping Convergence. *Machine Learning*, 61(1-3):49–69, June 2005. 9.3

[201] W. Yu, M. Clyne, S. M. Dolan, A. Yesupriya, A. Wulf, T. Liu, M. J. Khoury, and M. Gwinn. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*, 9:205, 2008. 1

[202] Zhiwen Yu, Hau-San Wong, and Guihua Wen. A modified support vector machine and its application to image segmentation. *Image and Vision Computing*, 29(1):29–40, 2010. 4

[203] Z. Yuan and R. Chappell. Isotonic designs for phase I cancer clinical trials with multiple risk groups. *Clin Trials*, 1:499–508, 2004. 6.2

[204] A L Zacharakis and D A Shepherd. The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, 16(4):311–332, 2001. 5

[205] B Zadrozny and C Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *International Conference on Machine Learning (ICML)*, pages 609–616. Citeseer, 2001. 3.2, 5, 7.2

[206] B Zadrozny and C Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *The Eighth International Conference on Knowledge Discovery and Data Mining (KDD)*, volume 02, pages //www–cseucsdedu/zadrozny/kdd2002–Transfpdf, 2002. 3.2, 5, 6.2, 7.2

[207] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. *Twenty-first international conference on Machine learning (ICML)*, page 114, 2004. 5.2, 6, 7.2

[208] Gul Zaman, Yong Han Kang, and Il Hyo Jung. Stability analysis and optimal vaccination of an SIR epidemic model. *Bio Systems*, 93(3):240–249, 2008. 4

[209] M. Zhang, J. Li, Y. M. Cai, H. Ma, J. M. Xiao, J. Liu, L. Zhao, T. Guo, and M. H. Han. A risk-predictive score for cardiogenic shock after acute myocardial infarction in Chinese patients. *Clin Cardiol*, 30:171–176, Apr 2007. 1.1

[210] Y. Zhang, Jeff Schneider, and A. Dubrawski. Learning the semantic correlation: An alternative way to gain from unlabeled text. In *Proceedings of the 22nd Conference on Neural Information Processing Systems (NIPS)*, pages 1945–1952. Citeseer, 2008. 4, 9.4.2.2

[211] X. M. Zhao, X. Li, L. Chen, and K. Aihara. Protein classification with imbalanced data. *Proteins*, 70:1125–1132, Mar 2008. 9

[212] Xing-Ming Zhao, Yong Wang, Luonan Chen, and Kazuyuki Aihara. Gene function prediction using labeled and unlabeled data. *BMC bioinformatics*, 9:57, January 2008. 9

[213] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. In *Neural Information Processing Systems (NIPS)*, 2003. 9.4.1