Detecting Anomalous Groups in Categorical Datasets

Kaustav Das Jeff Schneider Daniel B. Neill

April 2009
CMU-ML-09-104

Carnegie Mellon

Detecting Anomalous Groups in Categorical Datasets

Kaustav Das **Jeff Schneider** **Daniel B. Neill**

April 2009
CMU-ML-09-104

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We propose a new method for detecting groups of anomalies in categorical datasets. Our approach is a generalization of the spatial scan statistic, a commonly used method for detecting clusters of increased counts in spatial data. We extend this framework to non-spatial datasets with discrete valued attributes, where the degree of anomalousness of each record depends on its attribute values and we wish to find self-similar groups of anomalous records. We model the relationship between the attributes using a probabilistic model (e.g. Bayesian network), define a likelihood ratio statistic in terms of the pseudo-likelihoods for the null and alternative hypotheses, and maximize this statistic over all subsets of records. Since an exhaustive search over all such groups is computationally infeasible, we propose an efficient (but approximate) search heuristic. We show that this algorithm is able to accurately detect anomalous groups in real-world hospital, container shipping and network connections data.

This publication was supported in part by Grant Number 8-R01-HK000020-02 from CDC and by NSF under award IIS-0325581.

Keywords: Pattern Detection, Anomaly Detection, Machine Learning

1 Introduction

Anomalies can be defined as any observations or patterns that are different from the *normal* behavior of the data. We assume that under normal circumstances the records are generated by a particular process corresponding to the normal state of the system. In many applications, the motivating goal of anomaly detection is to detect the presence of some alternate process that might be affecting the normal behavior of the system. Such unexpected behavior might either be unwanted (e.g. in network intrusion detection), requiring user intervention, or it might be interesting (e.g. in astronomy), leading to a better understanding of the system or discovery of new phenomena. For example, in biosurveillance, we want to detect causes such as epidemics or bioterrorist attacks which give rise to unusual patterns of Emergency Department records. In customs monitoring, we are interested in detecting possible illegal activity, e.g. attempts to import unwanted illegal or dangerous material into the country. In general, these activities give rise to multiple records in the dataset which are anomalous, but are similar to each other. Our goal here is to detect such collections (or *groups*) of records which are generated by a process different from the normal behavior. Instead of trying to detect such records individually, we intend to use the presence of many such similar records (generated by a common process) to improve our detection performance.

To formalize our problem, assume we have a sufficiently large *training dataset* which defines the normal behavior of the system. We typically have unlabeled training data, in which we assume that no anomalies are present, but our methods can tolerate the presence of a small percentage of anomalies in the training set. Our goal is to detect the presence of groups of anomalies in an unlabeled *test dataset*. There might be single or multiple anomalous groups present, possibly generated by several distinct causes. We want to detect the anomalous groups of records, while minimizing the false positive rate.

2 Related work

Our proposed method can be thought of as generalizing two lines of previous research: the use of Bayesian networks and other probabilistic models to detect individually anomalous records in data, and the use of spatial scan statistics to detect clusters in spatial data. We extend the former method by integrating information from *groups* of anomalous records, and generalize the latter method from a simple univariate model (Poisson-distributed and spatially labeled counts) to multivariate datasets.

A Bayesian network is a popular representation of a probability model over the attributes for categorical data because of its parsimonious use of parameters, and efficient learning and inference techniques. Bayes Nets have been used for detecting anomalies in network intrusion detection [1, 12], detecting malicious emails [4] and disease outbreak detection [11]. A typical anomaly detection approach is to learn the structure and parameters of a Bayes Net using the training data, compute the likelihood of each record in the test dataset given the Bayes Net model, and report test records with unusually low likelihoods as potential anomalies. This method evaluates each test record in isolation and does not consider groupings of them or the relationship between test records. We test this individual record anomaly detector as a baseline algorithm in our empirical

studies. We also compare the performance of our proposed anomalous group detection method to another individual record anomaly detector, the Conditional Method, described in [2].

One of the most important statistical tools for cluster detection is the *spatial scan statistic* [8, 7, 10]. This method searches over a given set of spatial regions, finding those regions which maximize a likelihood ratio statistic and thus are most likely to be generated under the alternative hypothesis of clustering rather than the null hypothesis of no clustering. Kulldorff's framework assumes that the count of data points in a region S is Poisson distributed with some unknown rate of incidence q. Then the goal of the scan statistic is to find regions where the incidence rate is significantly higher inside the region than outside. The statistic used for this is the likelihood ratio $F(S) = \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$, where the null hypothesis H_0 assumes no clusters, and the alternative hypothesis $H_1(S)$ assumes a cluster in region S. Under H_0, we assume a uniform incidence rate q_{all}, while under $H_1(S)$ we assume that the incidence rate is higher inside region S than outside (i.e. $q_{in} > q_{out}$).

For the spatial scan, each data point consists of a set of real-valued location attributes, which can be mapped to a point in a Euclidean space, as well as a real-valued count. The search regions are defined in terms of the location attributes, while the likelihood ratio statistic is a function of the aggregate counts inside and outside a region. The spatial scan searches over subsets of the data which are geographically contiguous. For computational efficiency, further size and shape restrictions may be imposed on the set of search regions [7].

Rule-based algorithms have been proposed to detect groups of records. They find anomalous patterns by searching over *rules* of the form "$A_1 = v_1$ and $A_2 = v_2$" (e.g. Gender = Male and Symptom = Cough), where each rule defines a subset of records with the given attribute values. Anomaly Pattern Detection (APD) [3] begins with an individual anomaly detector and then uses the rule learning method to find groups of records that have an abnormally high proportion of individual anomalies. What's Strange About Recent Events (WSARE) [11] compares the actual and expected numbers of records fitting a rule using Fisher's Exact Test, and finds rules (subsets of records) with a higher or lower number of records than expected. We compare to both of these methods in our empirical studies.

The patterns detected by APD and WSARE are constrained to match a particular rule, and therefore are not flexible enough to include arbitrary subsets of the records. Another limitation of APD is that it can detect anomalous patterns only when the individual records forming the pattern are anomalous enough to be detected by the individual anomaly detector. We propose a method that can overcome the above limitations, finding arbitrary subsets of records that may not be individually anomalous but are anomalous when considered together.

3 Anomalous group detection

We would like to generalize the methodology of spatial scan statistics to find anomalous groups in arbitrary, non-spatial datasets with discrete valued attributes. This problem differs from spatial cluster detection in several respects. First, we do not have a defined set of location attributes, and thus we can no longer predefine a set of search regions based on geographical attributes such as size, shape, or contiguity. While we could conceivably define a distance metric between records

with categorical attributes, we do not have a direct embedding of the data points in Euclidean space or a notion of adjacency between different attribute values. Nevertheless, we want to formulate a measure of how well the data points fit as a *group* based on the similarity between them. We must then search over subsets of the data in order to find the most anomalous groups.

The second key difference is in the way we define the anomalousness of a data point or a group of points. Scan statistics are usually applied to detect over-densities of records in a given space. They assign the same level of interest or importance to each record, and aggregate individual records to counts to determine the anomalousness of a cluster. In our case, each record has many discrete-valued attributes rather than a single real-valued count, and can have an inherent degree of anomalousness depending on its features. Most records are generated from the "normal" (or usual) distribution of data and hence are not interesting for our purpose. We assume that the normal behavior of the data is defined by a model learned from a training dataset. Here we are no longer trying to detect simple over-densities of records in a certain feature space, but to detect groups of records that are both anomalous and also self-similar in some respect.

Instead of treating these two issues independently, we propose an approach that handles them simultaneously. As in the spatial scan statistic, our goal is to find a set of records that maximizes the likelihood ratio statistic $F(S) = \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$, where H_0 is the null hypothesis that there are no anomalies present, and $H_1(S)$ is the alternative hypothesis specifying that the set S is an anomalous group. We assume suitable probability distribution models for both the null and alternative hypothesis, and compute the data likelihoods given these models. More precisely, we learn a probability distribution model from the training dataset, which is assumed to contain no anomalies. Under the null hypothesis H_0, all data records are assumed to be drawn independently from this model. Under the alternative hypothesis $H_1(S)$, the records contained in subset S are assumed to have been drawn from a different probability model, while the rest of the data records are generated from the null model. We assume that data points are conditionally independent given the model, and thus records not contained in subset S have identical likelihoods given $H_1(S)$ and H_0. Thus the likelihood ratio statistic simplifies to:

$$F(S) = \frac{P(Data_S \mid H_1(S))}{P(Data_S \mid H_0)} = \frac{\prod_{i \in S} P(R_i \mid H_1(S))}{\prod_{i \in S} P(R_i \mid H_0)} \tag{1}$$

where $Data_S$ represents the subset of the data S and R_i is the ith record in $Data_S$. We note that the probability model parameters, but not the structure, for the alternative hypothesis $H_1(S)$ are learned directly from the records in $Data_S$. Since the number of records in group S may be small and we are using this data to fit a (potentially) large number of model parameters, data sparsity is a serious problem. In particular, learning the model parameters from the data $Data_S$ and evaluating the likelihood $P(Data_S \mid H_1(S))$, results in overfitting of the model. Using this as a part of the scoring function leads to the inclusion of a large number of irrelevant records in the best scoring group, as discussed in §3.2.

We use a two part approach to dealing with the problem of overfitting for the alternative hypothesis $H_1(S)$. First, we use Laplacian smoothing in the parameter estimation. Second, we use a "leave-one-out" method to compute the likelihood, which results in the following pseudo-

1. Learn the probability model for the null hypothesis H_0 from the training data.
2. For all subsets of the data S:
   (a) For each $R_i \in S$:
      i. Fit the alternate hypothesis probability model parameters using $Data_{(S-R_i)}$
      ii. Compute the leave-one-out likelihood $P(R_i \mid H_1(S - \{R_i\}))$.
   (b) Compute the group score, $F(S) = \frac{\prod_{i \in S} P(R_i \mid H_1(S-\{R_i\}))}{\prod_{i \in S} P(R_i \mid H_0)}$.
3. Output the groups with highest score.
4. Perform randomization testing to evaluate the statistical significance of the detected groups.

Figure 1: **Anomalous Group Detection Algorithm**

likelihood:

$$P_{pseudo}(Data_S \mid H_1(S)) = \prod_{i \in S} P(R_i \mid H_1(S - \{R_i\})) \tag{2}$$

This means that while computing the likelihood of the record R_i under the alternate hypothesis, we use a probability model with parameters learned from all the records in S minus R_i. Since we do not use the same record to estimate the parameters and to evaluate the likelihood, we expect to reduce the risk of over-fitting. We now define the group score as:

$$F(S) = \frac{P_{pseudo}(Data_S \mid H_1(S))}{P(Data_S \mid H_0)} \tag{3}$$

This scoring metric gives a higher score to anomalous records, as well as setting a constraint of similarity between the records in a group. If the records in S are similar to each other, then the alternate hypothesis will be able to model them tightly. This will result in a high value of the likelihood $P_{pseudo}(Data_S \mid H_1(S))$, thus increasing the score $F(S)$. Also, records that are poorly modeled by the null hypothesis will have a low value of the likelihood $P(Data_S \mid H_0)$, again increasing the group score $F(S)$. Hence maximizing this score leads to grouping of similar records and at the same time it prefers records that are anomalous (i.e. records with low likelihood under the null hypothesis).

3.1 The AGD Algorithm

We will now describe our method for anomalous group detection (AGD). An overview of the algorithm is given in Figure 1, and we now explain each step in detail. Although any probability distribution model can be used, we choose Bayesian Networks to model the probability distribution, and will specifically refer to them in the following description.

Step 1 of our algorithm is to learn the Bayes Net corresponding to the null hypothesis. We perform structure learning on the training dataset using the Optimal Reinsertion algorithm [9]. We assume this same Bayes Net structure for both H_0 and $H_1(S)$. We then learn the conditional probability table parameters of H_0 from the training dataset using smoothed maximum likelihood estimation.

Let us consider a node corresponding to the variable X_m in the Bayes Net. Let X_{Π_m} denote the set of variables corresponding to the parent nodes of X_m. The conditional probability table of X_m has parameters corresponding to the conditional probability values $\theta_{mjk} = P(X_m = j \,|\, X_{\Pi_m} = k)$. Here we need to estimate θ_{mjk} for each value of m, j and k. To deal with sparsity of the training data, we apply Laplace smoothing to adjust our estimate of each model parameter. We add $\frac{1}{J}$ to each N_{mjk} (the number of instances in the training dataset with $X_m = j$ and $X_{\Pi_m} = k$), where J is the arity of X_m. This makes the total weight of the prior add up to one for each variable X_m and each set of parent values k. The smoothed maximum likelihood estimates of the parameters are given by $\hat{\theta}_{mjk} = \frac{N_{mjk}+1/J}{\sum_{j'}(N_{mj'k}+1/J)}$

In **Steps 2-3**, we wish to find groups of records that maximize the likelihood ratio score $F(S)$. To do so, we search over all possible subsets of the test data. We note that an exhaustive search over all such subsets would require exponential time, but we will describe an efficient heuristic to make this search computationally feasible. For each subset of the data S, the alternative hypothesis assumes that the records in subset S form an anomalous group.

Step 2(a) of our algorithm computes the pseudo-likelihood of each record under the alternate hypothesis. To compute the pseudo-likelihood, in **Step 2(a)i** we first fit the parameters of the Bayesian Network for the alternative hypothesis $H_1(S - \{R_i\})$. These parameters are estimated from the counts in the subset of the test dataset represented by $S - \{R_i\}$. We follow an approach of smoothed maximum likelihood estimation similar to Step 1 above. In **Step 2(a)ii** we perform inference on the learned alternate hypothesis Bayesian Network model.

Step 2(b) of our algorithm computes the group likelihood ratio score $F(S)$, assuming conditional independence of the records given the models.

Note that Step 2(a) involves $|S|$ iterations of fitting the model parameters and performing inference. In the case of a Bayesian Network model using previously cached counts and a smoothed maximum likelihood estimation of parameters, this step can be done in time independent of the size of the group S. Using the notation from the description of Step 1,

$$P(R_i \mid H_1(S - \{R_i\})) = \prod_m \left[\frac{N_{mjk} + 1/J - 1}{\sum_{j'}(N_{mj'k} + 1/J) - 1} \right]_{\{j=X_m; k=X_{\Pi_m}\}} \tag{4}$$

$$P_{pseudo}(Data_S \mid H_1(S)) = \prod_m \prod_k \prod_{j=1}^{J} \left[\frac{N_{mjk} + 1/J - 1}{\sum_{j'}(N_{mj'k} + 1/J) - 1} \right]^{N_{mjk}} \tag{5}$$

Here N_{mjk} denotes the corresponding counts in subset of data $Data_S$. Notice that due to the exponentiation term N_{mjk}, this computation can be performed in time proportional to C, the number

of non-zero values of N_{mjk} in $Data_S$.

Step 3 of our algorithm outputs the highest scoring groups found in step 2. We use these scores to score the dataset with a measure of anomalousness. We assign the score of the most anomalous group detected as the score of the dataset: $F^*(Data) = \max_{S \in Groups} F(S)$. This is useful for distinguishing between datasets which contain anomalous groups and those without anomalous groups, e.g. distinguishing disease outbreaks from non-outbreak days.

Additionally, to identify individual records which are anomalies, we compute an anomalousness score for each individual record R in the test data, by finding the highest scoring group $S^*(R)$ that contains R. We can then compute the score of record R as $Score(R) = F(S^*(R))$. This gives a high score to any record that is contained in a highly anomalous group, regardless of whether the record is itself anomalous or just similar to other anomalous records.

In **Step 4**, we perform randomization testing to evaluate the statistical significance of the detected groups. To do so, we generate a large number N_{rand} of replica datasets under the null hypothesis that no anomalous groups are present. For each replica, we sample the training data uniformly at random to form a test dataset D_{rand} having the same number of records as the original test dataset, repeat steps 2 and 3 to find the highest scoring groups in the replica dataset, and record the maximum group score $F^*(D_{rand})$. To compute the p-value of a given subset of records S, we can compare the score $F(S)$ (from the original test dataset) to the distribution of maximum group scores from the replica datasets. The p-value is defined as $\frac{N_{beat}+1}{N_{rand}+1}$, where N_{beat} is the number of replica datasets with maximum group scores greater than $F(S)$. Since we are performing the same search procedure (maximization over subsets) for the original dataset and each replica dataset, the randomization testing approach correctly adjusts for the multiple hypothesis tests resulting from maximizing the score over many possible subsets.

We also note that, for a given dataset, the highest scoring subset will have the lowest p-value, and hence the ranking of regions is unchanged by randomization testing. When using the AGD method in practice, we can either choose a p-value threshold, and report all regions with p-values below the threshold, or choose a score threshold, and report all regions S with scores $F(S)$ above the threshold. In our evaluations discussed below, we have plotted the performance of AGD (and four other algorithms) over the entire range of such thresholds, and compared the area under these curves. For this type of evaluation, statistical significance testing by randomization is not necessary.

3.2 Search Heuristic

As noted previously, our method calls for searching over all possible subsets of the data. However, an exhaustive search requires exponential time and is thus likely to be computationally infeasible. Instead, we perform an efficient (but approximate) heuristic search in order to speed up the computation. More precisely, we adopt a greedy approach of growing the groups. We grow linearly many groups, starting from each record as an initial seed, and grow the group until no further additions can improve the likelihood ratio score. The algorithm is as follows:

1. Initialize $Groups \leftarrow \{\phi\}$

2. For each record $R_i \in Data_{test}$:

   (a) Initialize $S \leftarrow \{R_i\}$.

   (b) While S has changed over the previous iteration and $size(S) < MaxGroupSize$:

      i. Iterating over each record $R_j \in Data_{test} - Data_S$, find the record that maximizes the score $F(S \cup \{R_j\})$. Let the maximizing record be R_{max}.

      ii. If $F(S \cup \{R_{max}\}) > F(S)$ then set $S = S \cup \{R_{max}\}$; else $Groups = Groups \cup S$.

The anomalousness score of a record R in the test set is then defined as

$$Score(R) = \max_{S:\, S \in Groups,\, R \in S} F(S)$$

The impact of using the pseudo-likelihood score can be clearly seen during this greedy search procedure. When we use the full-likelihood scoring function as given by eqn. 1, overfitting results in an increase of the group score even when a dissimilar record is added to the group. This causes iteration 2(b) to keep adding records to the group until it reaches a size of $MaxGroupSize$. In most cases this results in the addition of many dissimilar records to the group before the iteration stops. The pseudo-likelihood scoring function (eqn. 2) helps us avoid this problem. In this case, the group score is increased only by the addition of records that are similar to the existing records within the group.

To evaluate the computational complexity of this search, let us consider a test set of size n. We treat each record as the initial seed and greedily grow the groups to some maximum size G. In our experiments below, we have used $G = 400$. Hence Step 2(b) is repeated at most nG times. We iterate over each record to find the one that best fits the group. Each such comparison can be done in time C, the number of non-zero values of N_{mjk} in $Data_S$. Hence, the overall complexity of the algorithm is $O(n^2GC)$. To make the algorithm efficient, we use a bounding strategy to prune the set of records for which we compute the score $F(S \cup \{R_j\})$ in step 2(b)i. Based on the current best candidate for inclusion in the group, it is possible to compute an upper bound on the null hypothesis likelihood score for any other candidate. Only records that have a null hypothesis likelihood score less than this bound needs to be considered. As we search through the records, we can dynamically update this upper bound based on the current best candidate. In certain cases, it allows us to significantly speedup the computation to determine the best record to add to a group.

3.3 Comparison to spatial scan

As noted above, our AGD algorithm can be thought of as a generalization of the spatial scan statistic [7] to arbitrary multivariate datasets without predefined location or count attributes. Here we summarize how the original spatial scan differs from our algorithm described in Figure 1:

1. The spatial scan searches over a set of contiguous spatial regions that are predefined based on the location attributes of the data, while we perform a heuristic search over arbitrary subsets of the data.

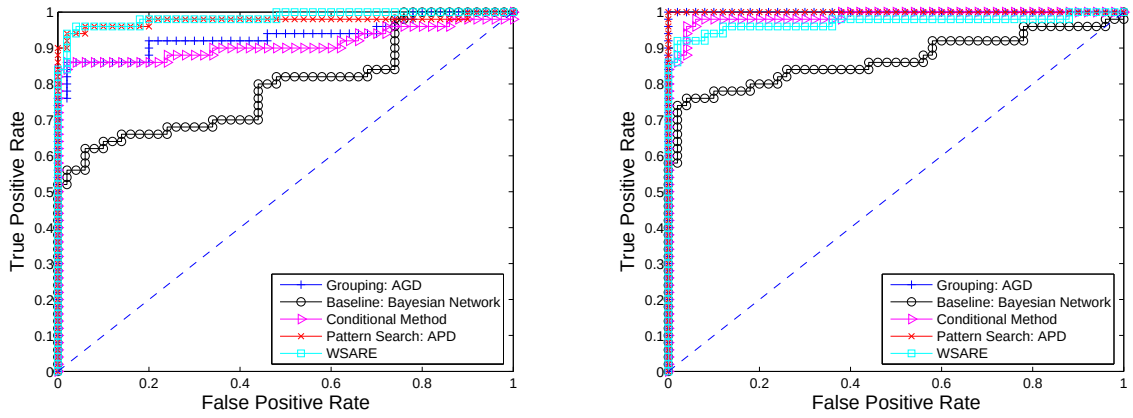
2. In **Step 1**, the spatial scan learns only a single parameter (the uniform incidence rate q_{all}) for the null hypothesis, rather than a probability model relating all variables in the multivariate dataset. Similarly, in **Step 2(a)i**, the spatial scan learns only two parameters (q_{in} and q_{out}) for $H_1(S)$. In **Step 2(a)ii**, it computes the likelihoods under the null and alternative hypotheses using a simple Poisson count model, rather than performing inference on a probability model.

4 Datasets

1. PIERS Dataset: Our first dataset consists of records describing containers imported into the country. Each record consists of 10 attributes. 7 of them are categorical: the container's country of origin, departing and arriving ports, shipping line, shipper name, vessel name and the commodity being shipped. There are three real valued attributes, the size, weight and value of the container. We have categorized these to five discrete levels. Since there were no labels in the original data, we create synthetic anomalies by randomly flipping attribute values. We first create a random partition of the dataset into training (100,000 records) and test (1000 records) sets. We modify a random 10% of the test set records to be an anomalous group. To create a group of anomalies G, we first make $Size_G$ identical copies of a randomly chosen record. Each record in the group is then modified by changing the value of a randomly chosen attribute. The new value is drawn from the marginal distribution of that attribute in the training dataset. The records within the group are similar to each other since each pair of records in G differs by at most two attributes. Each record in the group is anomalous because randomly changing an attribute value breaks the relationship of that attribute with the rest of the attributes. One possible real world scenario where such an anomalous group might occur is when a smuggler smuggles goods using similar methods which have proved successful in the past.

2. Emergency Department Dataset: This real-world dataset contains records of patients visiting Emergency Departments (ED) from hospitals around Allegheny County in the year 2004. Each record consists of six categorical attributes: the hospital id, prodrome, age decile, home zip code and the chief complaint class. The dataset is injected with simulated ED cases resembling an anthrax release. The simulated cases of anthrax were produced by a state-of-the-art simulator [5] that implements a realistic simulation model of the effects of an airborne anthrax release on the number and spatial distribution of respiratory ED cases. We treat the first two days when the attack symptoms begin to appear as the test data, thus evaluating our ability to detect anthrax attacks within two days of the appearance of symptoms. We train our model on the previous 90 days' data. Note that while we have a model for anthrax release, AGD is not given any information from it. Thus this dataset tests our ability to recognize a realistic, but previously unknown, disease outbreak.

3. KDD Cup 1999 Network Intrusion Detection Dataset: We have also evaluated AGD on the KDD Cup 1999 data [6], which contained a wide variety of intrusions simulated in a military network environment. Each record is a vector of extracted feature values from a connection record obtained from the raw network data. In total there are 41 features, most of them taking real values. Using all the features in the detection task causes most of the intrusion records to individually



(a) Emergency Department dataset

(b) PIERS dataset

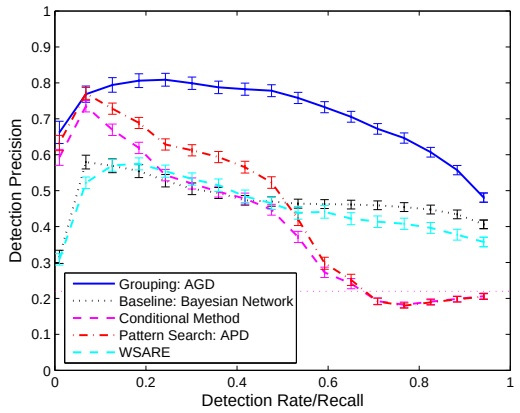
Figure 2: Algorithm performances for detection of datasets with anomalies

stand out from the normal ones as seen in [2, 3]. Hence, we chose a subset of 22 features that includes the basic features of individual TCP connections and the content features suggested by domain knowledge. This evaluation setup creates groups of self-similar anomalous records that are individually anomalous to a lesser degree. The real valued features were discretized to 5 levels. The goal of the KDD dataset was to produce a good training set for learning methods that use labeled data. Hence, in this case we have labeled anomalies (network attacks) and the proportion of attack instances to normal ones is very large. To create more realistic data, we have reduced the number of attack records to 10% of the test dataset. We have run our algorithms on the 7 most common types of attacks - apache2, guess password, mailbomb, neptune, smurf, snmpguess and warezmaster. Correspondingly, we created seven different test sets containing 10% records of the particular attack type, and 90% normal records. We use the rest of the normal records for training our model.

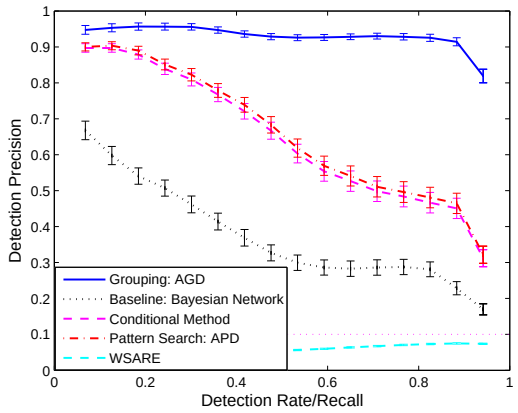
5 Evaluation

We compare the performance of our AGD method to the baseline method, which detects individual records with low likelihoods given the null hypothesis Bayes Net model. In our implementation of the baseline method, we use Optimal Reinsertion [9] to learn the structure, and perform smoothed maximum likelihood estimation of the network parameters. We also compare the performance to three other related methods discussed in Section 2: the Conditional Method [2], WSARE [11], and APD [3]. We note that the better-performing of the two individual anomaly detectors was used to detect individually anomalous records for APD on each dataset (i.e. we used the Conditional Method for the ED and PIERS datasets, and the Bayes Net method for the KDD Cup datasets).

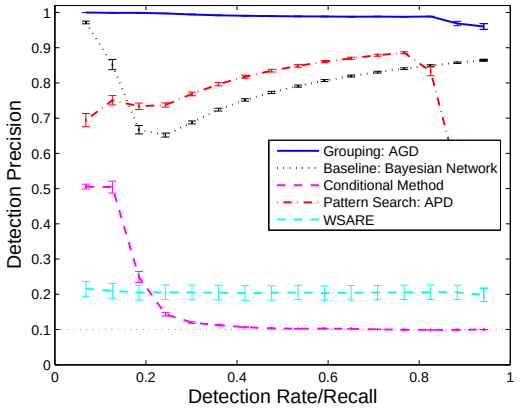
The procedure for randomly generating the test data and injecting anomalous groups in them was repeated 50 times for each of the nine experiments (ED, PIERS, and seven different KDD



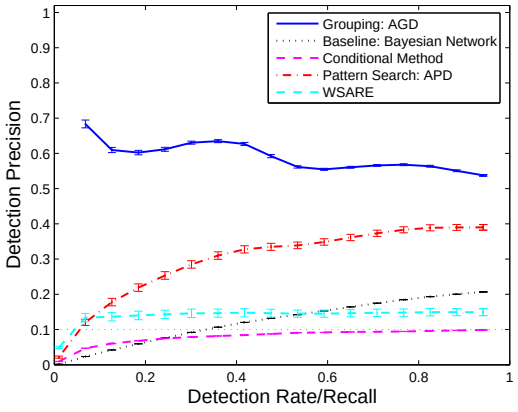
(a) Emergency Department dataset



(b) PIERS dataset



(c) KDD Cup 99: guess password



(d) KDD Cup 99: mailbomb

Figure 3: Comparison of detection precision vs. recall for AGD and baseline methods, with standard errors. The dashed line at precision = 0.1 (0.22 for the Emergency Department data) is the average performance of the "chance" algorithm that chooses records at random.

Table 1: Normalized area under the true positive rate vs. false positive rate curves for AGD and related methods, with standard errors

Dataset	AGD	Bayesian Network	Conditional Method	APD	WSARE
ED	0.932 ± 0.026	0.793 ± 0.041	0.910 ± 0.034	0.976 ± 0.018	$\mathbf{0.984 \pm 0.01}$
PIERS	$\mathbf{1.0 \pm 0.0}$	0.867 ± 0.038	0.987 ± 0.007	$\mathbf{1.0 \pm 0.0}$	0.970 ± 0.019
apache2	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	0.727 ± 0.051
guess passwd	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	0.957 ± 0.016	$\mathbf{1.0 \pm 0.0}$	0.610 ± 0.045
mailbomb	0.788 ± 0.02	0.82 ± 0.023	0.276 ± 0.036	$\mathbf{0.936 \pm 0.03}$	0.54 ± 0.048
neptune	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	0.695 ± 0.055
smurf	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	0.286 ± 0.031	$\mathbf{1.0 \pm 0.0}$	0.781 ± 0.048
snmpguess	$\mathbf{1.0 \pm 0.0}$	0.962 ± 0.023	0.294 ± 0.034	0.935 ± 0.02	0.679 ± 0.052
warezmaster	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	0.789 ± 0.042

Cup attack types). For each experiment, we also produced 50 additional sets of test data (of the same size) with no anomalies injected. These runs are helpful in determining the ability of the algorithms to differentiate between entire datasets containing anomalous groups and those without anomalous groups.

We evaluate the performance of the algorithms in two different ways. First, we examine the ability of the algorithms to identify and distinguish between entire test datasets which have anomalous groups against ones which are normal (i.e. do not have any anomalies). In the Emergency Department data, for example, this corresponds to distinguishing between an anthrax attack occurring and no attack occurring. As noted above, the algorithms are run over 100 test datasets, where half of these datasets contain injected anomalies.

For the three methods that explicitly search over sets of records, the dataset score is set as the score of the most anomalous group (AGD), pattern (APD), or rule (WSARE) detected. For two methods that score records individually (the baseline Bayesian Network method and the Conditional Method), the dataset score is calculated as the sum of the individual scores of all the records. Note that since these methods do not model groups of anomalies, summing up the individual record scores (as opposed to considering the single most anomalous record score) gives significantly better detection performance. We then examine each method's tradeoff between its false positive rate (proportion of datasets without anomalies that were falsely detected as being anomalous) and its true positive rate (proportion of datasets with anomalies that were correctly detected as being anomalous). This is the standard ROC curve: a higher curve denotes better detection performance, since it corresponds to a higher true positive rate for a given false positive rate. The area under the ROC curve (AUC) can be used as a summary measure, where higher AUC corresponds to better average performance.

For the 50 datasets that contain anomalies, we also evaluate the ability of each algorithm to identify which individual records were anomalous. For example, in the Emergency Department dataset, this corresponds to identifying which patients have been affected by the anthrax attack

Table 2: Area under the detection precision vs. recall curves for AGD and related methods, with standard errors

Dataset	AGD	Bayesian Network	Conditional Method	APD	WSARE
ED	**0.729 ± 0.032**	0.479 ± 0.027	0.375 ± 0.026	0.420 ± 0.027	0.465 ± 0.033
PIERS	**0.935 ± 0.018**	0.356 ± 0.043	0.641 ± 0.047	0.655 ± 0.046	0.053 ± 0.003
apache2	**1.0 ± 0.0**	0.973 ± 0.003	0.951 ± 0.004	0.882 ± 0.021	0.215 ± 0.042
guess passwd	**0.991 ± 0.002**	0.773 ± 0.008	0.124 ± 0.005	0.804 ± 0.013	0.205 ± 0.041
mailbomb	**0.587 ± 0.007**	0.136 ± 0.001	0.086 ± 0.001	0.329 ± 0.019	0.146 ± 0.022
neptune	0.993 ± 0.002	0.984 ± 0.003	**1.0 ± 0.0**	0.986 ± 0.003	0.217 ± 0.030
smurf	**0.974 ± 0.003**	0.640 ± 0.006	0.089 ± 0.001	0.889 ± 0.015	0.237 ± 0.032
snmpguess	**0.987 ± 0.002**	0.288 ± 0.002	0.087 ± 0.001	0.521 ± 0.030	0.266 ± 0.034
warezmaster	**0.892 ± 0.014**	0.852 ± 0.009	0.430 ± 0.014	0.677 ± 0.034	0.141 ± 0.021

and which patients are in the Emergency Department due to other causes. We plot the detection precision, i.e. the ratio of number of true positives to the total number of predicted positives, against the detection rate, i.e. the proportion of total true anomalies that are detected. The plots are generated by varying the threshold used to flag anomalies. The standard error estimates are also shown in the plots. Here, a higher curve denotes better performance, since it corresponds to a higher detection precision for a given detection rate.

The Bayesian Network method and the Conditional Method assign a anomalousness score to each individual record which can be directly used to perform this evaluation. For the rest of the methods, the score of a record is assigned as the score of the most anomalous group (AGD), pattern (APD) or rule (WSARE) that it belongs to.

6 Results

We first examine the performances of the algorithms in differentiating between test datasets that contain anomalous groups and datasets without injected anomalies. Figures 2(a) and 2(b) show the ROC curves for the ED and PIERS datasets respectively, and Table 1 shows the area under the ROC curve (AUC) for all nine experiments (ED, PIERS, and the seven attack types for KDD Cup).

We can see that, for both the ED and PIERS datasets, AGD performs better than the baseline Bayesian Network method, having a greater true positive rate for a given false positive rate. The AGD method has significantly larger area under the curve than the baseline method (using a paired t-test, $\alpha = 0.05$) in both cases. For the KDD Cup network intrusion dataset, AGD is able to perfectly differentiate the datasets (i.e., has a true positive rate = 1 for all false positive rates) for all attack types except mailbomb. AGD also performs well across all nine experiments as compared to APD, WSARE, and the Conditional Method. However, we see that for the Emergency Department data, WSARE gives us the best performance. This is not very surprising, since WSARE was originally developed to detect outbreaks among patients admitted to Emergency Departments, and

WSARE performs relatively poorly for the other experiments. APD performs similarly to AGD for this evaluation, but as we demonstrate below, AGD performs substantially better on identifying anomalous records.

Next, we look at the performances of the algorithms in identifying anomalous records in the datasets that contain anomalies. Figure 3 shows the relative performance of the five methods on the ED and PIERS datasets, as well as two of the seven KDD Cup experiments (guess password and mailbomb). In all of the plots, the baseline performance of randomly choosing which records are anomalous is shown by a dashed line. Table 2 gives the normalized area under the curve for the detection rate interval [0.1, 0.9], with standard errors, for each method on all nine experiments.

We see that AGD performed significantly better than the baseline Bayes Net method for all nine experiments, demonstrating that using the group information substantially improves our ability to detect which records are anomalous. On eight of the nine experiments, AGD also performed significantly better than the three related methods (APD, WSARE, and the Conditional Method). The one exception was the KDD Cup neptune attack, where the Conditional Method achieved perfect performance (AUC = 1) while AGD achieved near-perfect performance (AUC = 0.993). All differences in AUCs between the best method (performance shown in bold font) and second-best method were found to be significant at $\alpha = 0.05$.

7 Discussion

We note that, instead of using the maximum likelihood estimates of the parameters in §3, we have also explored the use of a Bayesian approach. In this approach, we consider a Dirichlet prior distribution over the parameters, and compute the marginal likelihood of the data as the score function $F(S)$. From a theoretical standpoint, it would seem that this approach might lessen the effect of overfitting while computing the likelihood under the alternate hypothesis. However, our preliminary empirical results indicate that using the marginal likelihood scoring function is not very effective at addressing overfitting, and the resulting groups still grow without bound (as was the case for the original maximum likelihood approach, motivating our use of the pseudo-likelihood). When we consider the marginal likelihood approach, the pseudo-likelihood is no longer well defined, since the likelihood of the data can no longer be expressed as a product of the individual record likelihoods when integrated over the multinomial parameters. Hence we chose to use the maximum likelihood, rather than marginal likelihood, estimates of the multinomial parameters in our pseudo-likelihood score function.

8 Conclusions and Future Work

In this work we describe a method of generalizing likelihood based anomaly detection (using Bayesian Networks) by integrating the information about *groups* of anomalous records. We evaluate the methods on three real-world datasets, injected with simulated and real anomalies. The performance is evaluated for the tasks of detecting individual anomalous records and distinguishing between datasets having or not having anomalous groups. The Anomalous Group Detec-

tion method gives significantly better detection performance over the baseline method for both of these tasks. Additionally, AGD was shown to outperform three previously proposed methods (WSARE, APD, and the Conditional Method), substantially improving the identification of anomalous records for all three datasets.

As a follow-up to the present study, we are currently evaluating how the size and self-similarity of the anomalous groups impact the relative performance of detection methods. Our preliminary results suggest that AGD performs best when the anomalous groups are larger and more self-similar, while APD outperforms AGD for detecting smaller groups where each individual record in the group is anomalous. In the extreme case of a group consisting of a single, highly anomalous record, we would expect individual anomaly detection methods such as the Conditional Method to outperform both APD and AGD.

Although we exclusively deal with categorical valued datasets in this work, we can easily generalize it to handle datasets containing real valued attributes as well, using Bayesian Network models containing both categorical and real valued nodes. Finally, while we believe that the chosen BARD outbreak simulation is a highly realistic model of anthrax release, we also plan to evaluate our methods on real, known disease outbreaks in the future, as well as a variety of other datasets.

References

[1] A. Bronstein, J. Das, M. Duro, R. Friedrich, G. Kleyner, M. Mueller, S. Singhal, and I. Cohen. Bayesian networks for detecting anomalies in internet-based services. In *Intl. Symposium on Integrated Network Mgmt.*, 2001.

[2] Kaustav Das and Jeff Schneider. Detecting anomalous records in categorical datasets. In *Proc. ACM Knowledge Discovery and Data Mining*, Aug 2007.

[3] Kaustav Das, Jeff Schneider, and Daniel Neill. Anomaly pattern detection in categorical datasets. In *Proc. ACM Knowledge Discovery and Data Mining*, Aug 2008.

[4] Shih Dong-Her, Chiang Hsiu-Sen, Chan Chun-Yuan, and Binshan Lin. Internet security: malicious e-mails detection and protection. *Industrial Mgmt. and Data Sys.*, 104:613 – 623, Sep 2004.

[5] William R. Hogan, Gregory F. Cooper, Garrick L. Wallstrom, Michael M. Wagner, and Jean-Marc Depinay. The bayesian aerosol release detector: An algorithm for detecting and characterizing outbreaks caused by an atmospheric release of bacillus anthracis. *Statistics in Medicine*, 26:5225–5252, Sep 2007.

[6] KDDCup. The third international knowledge discovery and data mining tools competition, kdd cup 1999. In *The Fifth International Conference on Knowledge Discovery and Data Mining*, 1999.

[7] M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, pages 1481–1496, 1997.

[8] M. Kulldorff and N. Nagarwalla. Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14:799–810, 1995.

[9] Andrew Moore and Weng-Keen Wong. Optimal reinsertion: A new search operator for accelerated and more accurate bayesian network structure learning. In *20th Intl. Conf. on Machine Learning*, pages 552–559, Aug 2003.

[10] Daniel B. Neill and Andrew W. Moore. Anomalous spatial cluster detection. In *Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection*, August 2005.

[11] Weng Keen Wong, Andrew Moore, Gregory Cooper, and Michael Wagner. Bayesian network anomaly pattern detection for disease outbreaks. In *Twentieth Intl. Conf. on Machine Learning*, pages 808–815, Aug 2003.

[12] Nong Ye and Mingming Xu. Probabilistic networks with undirected links for anomaly detection. In *IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, pages 175–179, June 2000.

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Carnegie Mellon