# PCX: Markov Blanket Classification for Large Data Sets with Few Cases

Xue Bai, Clark Glymour,
Rema Padman, Joseph Ramsey, Peter Spirtes, Frank Wimberly

Center for Automated Learning and Discovery
March 01, 2004
CMU-CALD-04-102

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**Abstract**

Data sets with many discrete variables and relatively few cases arise in many domains. Several studies have sought to identify the Markov Blanket (MB) of a target variable by filtering variables using statistical decisions for conditional independence and then applying a classifier using the MB predictors. Other studies have applied the PC algorithm or heuristic procedures, to estimate a DAG model of the MB and classify by Bayesian updating. The PC output is not a DAG or MB, and how a DAG representation of the MB is formed in these studies is not specified. Using a filter from the HITON feature selection procedure, we find a Markov equivalence class using the PC algorithm, provide an explicit algorithm for converting the output to a graphical Markov Blanket, and classify by Bayesian updating. We apply this procedure (PCX) to five empirical data sets from different domains, and compare it with results from HITON, which applies several state-of-the-art classifiers. The PCX classifier has fewer variables than those found by the HITON procedure, and gives comparable classification accuracy while supplying insight into possible causal relations among the variables.

**Keywords**
PCX, PC algorithm, Bayesian Networks, Markov Blanket, Markov Blanket Bayesian Classifier

## 1. INTRODUCTION

In genetics, proteomics, clinical diagnosis, and many other domains, data sets arise with a very small ratio of cases to variables. Such data present familiar dimensional difficulties for classification of a target variable, and even more difficulty for the determination of those variables that actually influence, or are influenced by, a target variable. Classification that relies on large numbers of variables is often inapplicable, for example in clinical diagnosis problems; the use of inessential variables tends to increase the variance of classification estimates; and classification with large number of variables provides no insight into causal relationships, insight that can be important in guiding further empirical research. Hence, the twin problems arise of finding among a large number of variables a small subset essential to and sufficient for, classification, and of estimating the causal relations relating those variables to the target variable.

Recent work by Aliferis, et al. [2], has provided an important approach to the first of these problems. In a two-stage procedure, Aliferis, et al., find a subset of variables estimated to constitute the Markov Blanket of the target variable, i.e., the smallest set of predictor variables conditional on which all other variables in the dataset are independent of the target variable. They then use non-Bayesian classifiers with the reduced variable set, finding comparable accuracy to the results of applying the classifiers to the full original variable set. If the joint distribution of the full set of variables satisfies the Markov property for a directed acyclic graph (DAG), and a converse property, faithfulness, both of which are specified below, and the assumption that all common causes of variables in the original data set are also in that data set, and any probability constraints assumed by the classifier also hold, their procedure is guaranteed to find the correct Markov Blanket with probability 1 in the large sample limit.

The notion of a Markov Blanket (MB) for a variable *X* in a dataset *D* has two senses: it is the minimal set of variables conditional on which all other variables in *D* are independent of *X*, and it is also a DAG of that minimal set together with the target node. When the parameters of a MB DAG are estimated, the result is a Bayesian network, which, using standard Bayesian updating procedures, itself provides the basis of a classifier that assumes only a multinomial distribution of the discrete variables. Bayesian networks also have a causal interpretation: a directed edge from one variable to another, *X* -> *Y*, represents the claim that *X* is a direct cause of *Y* with respect to other variables in a DAG, i.e., if other variables were to be held fixed at appropriate values, and *X* were varied by an intervention (e.g., randomization), *X* and *Y* would covary [1, 3,]. An MB DAG can thus provide both a classifier and some insight into causal relations between a reduced set of predictors and the target variable.

We describe a two stage algorithm that finds an MB DAG and uses it as a classifier with conventional Bayesian updating. Our procedure uses the variables selected by the first stage of the HITON algorithm to find a reduced set of variables. Our second stage, the PCX algorithm, finds the MB DAG, further reducing the number of prediction variables needed, estimates the parameters by maximum likelihood, and classifies cases. Using five data sets employed by Aliferis, et al. [2], we show that the procedure results in a smaller number of predictors in all cases than does the full HITON algorithm, while providing comparable classification accuracy and yielding hypotheses about the causal structure of the system.

## 2. REPRESENTATION AND BACKGROUND ALGORITHMS

A Bayesian network is a DAG whose nodes are random variables with a joint probability distribution that factors according to the product of distributions of each variable conditional on its parents (i.e., variables with edges directed into it) in the graph. Equivalently, the distribution and graph satisfy the local Markov condition: each variable is independent of its non descendants conditional on its parents. A probability distribution is faithful to a DAG if and only if all conditional independence relations in the distribution are consequences of the local Markov condition applied to the DAG. Two DAGs are Markov equivalent if they imply the same conditional independence relations by the local Markov condition. The Markov Equivalence class of a DAG ***G***, **ME**(***G***) is the set of all DAGs Markov equivalent to ***G***. For variable set **V** and variable *X* in **V**, the set of Markov Blanket variables in **V** for *X*, **MB**(V, *X*) is the smallest subset of **V** not containing *X* such that *X* is independent of **V** \ **MB**(**V**,*X*). Given a DAG ***G*** with vertex set **V** and a probability distribution P on **V** locally Markov for ***G***, the Markov Blanket Bayesian network for **V**, *X*, ***G***, **MB**(***G***, **V**, *X*) is the subgraph of ***G*** on vertices **MB**(**V**,*X*) (less the edges between parents of *X* and between parents of children of *X*) and the marginal of P on that subset. The edge structure of **MB**(***G***, **V**, *X*) consists of the directed edges into *X* from the parents of *X* in ***G***, the directed edges from *X* into the children of *X* in ***G***, and the directed edges from the parents of the children of *X* in ***G*** into the children of *X* in ***G*** [4]. Classification by updating with **MB**(***G***, **V**, *X*) [4] imposes no restrictions on the probability distribution of the target variable conditional on the variables in the Markov Blanket beyond those implicit in the discretization and the conditional independence constraints implied by the graphical structure and the Markov condition.

In discussing previous literature and in our own procedure, we refer to the PC algorithm. The orientation rules of the original presentation of the algorithm [1] are incomplete and the complexity is sensitive to the implementation of the orientation procedure; our implementation employs a complete set of orientation rules [14]. Assuming i.i.d. samples from a probability distribution faithful to a DAG for the initial variable set, the PC algorithm converges probability 1 to a graphical object called a pattern or essential graph that represents **ME**(***G***) in the large sample limit [1], a property we will refer to as soundness.

Some features of the PC output should be noted. The output of PC is a mixed graph with undirected edges, directed edges, and possibly doubly directed edges. Variables adjacent—in the output—i.e., connected by an edge of some kind—represent adjacencies common to all DAGs in the conjectured Markov equivalence class. Doubly directed edges can arise because statistical decisions yield combinations of conditional independence relations inconsistent with the Markov and faithfulness assumptions for any DAG on the specified variable set. For example, if there are unrecorded variables that influence two or more recorded variables, even if the joint distribution on the unrecorded and recorded variables were faithful to a DAG, the marginal probability distribution on the recorded variables may not be Markov and faithful for any DAG. In forming an estimated Markov Blanket DAG from PC output it becomes essential to direct undirected edges and remove one orientation in doubly directed edges. Further, the PC output will typically have directed edges that may be in the underlying DAG, but are not in the MB DAG of the target.

Complexity is dominated by the adjacency search. For a DAG whose vertices each have degree k and n variables the adjacency stage of the algorithm requires $C(2, n)\ 2^k$ statistical tests (for k < n-

2). The actual degree of the DAG, if any exists, is of course unknown in advance. The PC algorithm requires the user to set one parameter, used as the alpha value in tests of conditional independence. Although the procedure has been criticized on this ground, the algorithm can be implemented with decisions about conditional information measures, but there is little to be gained thereby, since application of these measures likewise requires a threshold. Finally, the results of PC are asymptotically correct in the following senses: edges not in the output are not in the DAG; oriented edges in the output are in the DAG; orientations in the output are in the DAG [5].

## 3. SURVEY

There are many studies applying Markov Blanket classifiers and comparing their accuracy with a variety of alternatives, but fewer studies that generate the Markov Blanket from data, and those usually for small numbers of variables. Theoretically correct Bayesian algorithms [4] for finding DAGs are now known, but have not been applied to the problem of finding MBs for data sets with large variables.

An exception is the work of Koller and Sahami [5] who use a heuristic procedure to find the Markov Blanket variables in datasets with large numbers of variables. The heuristic is based on two (not always true) assumptions, that the target influences the predictors, and that the variables most strongly associated with the target are in its Markov Blanket. No classifier is studied. In Kohler and Sahami's experiments with large variable sets, one hundred or more predictor variables remain.

Two algorithms similar to ours, GS [6] and IAMBnPC [7] have been proposed. GS uses a measure of association with the target variable and conditional independence tests to find a reduced set of variables estimated to be the Markov Blanket; it then applies an algorithm to produce an MB graph and classifier. The second stage of the procedure is unsound. IAMBnC uses a dynamical variant of the variable selection filter, followed by PC. How a graphical Markov Blanket is obtained from PC output is not explained. A variant interIAMBnPC, [9] interleaves the PC algorithm with the filter. On the standard ALARM network test example, PC applied directly performs better than any of these algorithms, although PC is known to be inferior on this structure to a Bayesian algorithm that searches over **ME** sets [8]. On simulated data with 1000 variables, interIAMBnPC performed best; PC was not applied directly. A final procedure, HITON, [2] on which we rely in this paper, supplements the dynamic variable filter of IAMBnPC with a "wrapper" using any of several non-Bayesian classifiers, and then classifies the target with the non-Bayesian classifier. A graphical MB is not produced. The results are compared on five empirical data sets from a variety of domains each with a very large ratio of variables to cases. We adapted the first stage of the HITON algorithm, which is described as Figure 1.

Our study, using an algorithm PCX, similar to IAMBnPC, differs in three ways from these valuable precedents. (1) our implementation of PC reduces the runtime of orientation rules and maximizes orientation information; (2) we apply our procedure to the large empirical data sets used in the Aliferis, et al. paper and compare the results; (3) the papers cited above give no indication of how PC output, which for reasons given in the previous section is not itself an MB

or even a DAG, and is typically consistent with several such structures, is converted into an MB. Any such conversion is somewhat arbitrary; we give an explicit algorithm for the conversion.

**HITON-PC**(Data *D*, Target *T*)

"returns parents and children of *T*"

*CurrentPC* = {}

**Repeat**

Find variable *Vi* ( CurrentPC that maximizes association(Vi T) and admit Vi into CurrentPC

If there is a variable X and a subset S of CurrentPC s.t. □(X : T | S))

remove X from CurrentPC;

mark X and do not consider it again in phase I

**Until** no more variables are left to consider

**Return** *CurrentPC*

**HITON-MB**(Data *D*, Target *T*)

"returns a set of candidate Markov Blanket nodes of *T*"

*PC* = parents and children of *T* returned by HITON-PC(*D, T*)

*PCPC* = parents and children of the parents and children of *T*

$CurrentMB = PC \cup PCPC$

// Retain only parents of common children

$\forall$ potential spouse X $\in$ *CurrentMB* and $\forall$ $Y \in PC$:

**if** $\neg \exists\ S \subseteq \{Y\} \cup \boldsymbol{V} - \{T, X\}$ so that $\perp\ (T ; X \mid S)$

**then** retain *X* in *CurrentMB*

**else** remove it

Return *CurrentMB*

**Figure 1**: Pseudo-code for the first stage of HITON

## 4. DATA

The data we use were kindly provided to us by Aliferis et al. The data sets are described in Table 1. The thrombin problem concerns identification of biomolecules that bind to thrombin and have potential as anti-clotting properties. [9] Prediction variables are molecular structural properties.

Arrhythmia data concern classification of subjects into 8 disease categories from clinical and EKG data [10]. Ohsumed data concern identification of Medline documents relevant to neonatal diseases [11]. The lung cancer problem requires diagnosis of squamus vs. adenocarcinoma from gene expression data. [12] The prostate cancer problem concerns diagnosis of prostate cancer from mass spectroscopy of human sera.

**Table 1. Dataset Characteristics**

| **Dataset** | **Thrombin** | **Arrhythmia** | **Ohsumed** | **Lung Cancer** | **Prostate Cancer** |
|---|---|---|---|---|---|
| **# Variable** | 139,351 | 279 | 14,373 | 12,600 | 779 |
| **Variable Types** | binary | nominal/ordinal/continuous | continuous | continuous | continuous |
| **Target** | binary | nominal | binary | binary | binary |
| **Sample Size** | 2,543 | 417 | 5,000 | 160 | 326 |
| **C. V. Folds** | 1 | 10 | 1 | 5 | 10 |

## 5. PROCEDURE

For each of five data sets, we use data and the initial variable filter of the HITON algorithm to obtain a reduced set of variables relevant to the target. The algorithm PCX is then applied to obtain a graphical MB, and the MB is tested on the data with varying cross validation folds (chosen, for comparability, to be the same as in the HITON study), and the classification results are given as confusion matrices (choosing the most probable value of the target for each case), along with the area under the Receiver Operating Characteristic (ROC) curve (AUC). The variables selected by the HITON filter and cross-validation samples were kindly provided us by Aliferis, et al., and we use their projections of continuous variables to categorical values[1]. Aliferis et al. apply HITON with several state-of-the-art classifiers, selected differently for different data sets. Classifier parameters in their study were adjusted for each specific cross-validation run. In our experiments, PCX has two adjustable parameters: the significance (or alpha level) used in all independence tests in the algorithm and the depth of search used in the

[1] The discretization process was done by first normalizing the data. After normalization, the discretization routine uses the null hypothesis to determine significance, with a=0.05. If not significantly associated, then discretize according to: 0 (less then -1 standard deviation); 1 between -1 and 1 standard deviation); 2 (greater than 1 standard deviation). If significant, and (1) a binary chisquare test is done (tests for significance after ordering and testing, dividing at all possible points on the ordered set, assigning values 0,1); (2) a ternary kruskalwallis test is done (using a sliding window of varying width to assign values after ordering of 0,1,2), the parameters for the best of binary and ternary are used to discretize the feature, and values are assigned 0,1 or 0,1,2 respectively.

PC algorithm which is called by PCX. Depth of search in the PC component can be limited, but we used unlimited depth in our experiments. Except in the case of a data set for lung cancer, in applying PCX we pretest for alpha level and fix the alpha level in all cross validation runs. The lung cancer data is an extreme case. It has 12,000 variables but only 160 samples. The performance metrics become less stable at a fixed alpha level. The default alpha level used in three of the cross validation runs produced no positives in the other two. In these two runs the alpha level was adjusted upwards.

The PCX algorithm is described in Figure 2: The input parameters are $D_1$: a training data set with m variables and n samples; $D_T$, a test or prediction data set that has the same variables as $D$; $T$: the classification variable; $d$: the maximum size of condition sets for the conditional independence tests in PC search; and $\alpha$ is the significance level. The output is the graphical Markov Blanket structure ($MB$) for $T$ and the confusion matrix $M$.

For each classification problem, we use the variables from the HITON filter and the data as inputs to the PCX algorithm. The classification procedure is tested with the same n fold cross validation for each data set as used by Aliferis et al., training on 90% of the data and testing on the remaining 10% of the data. We choose the alpha level for each problem from a preliminary sample of the data. For each data set, the alpha level is constant for all cross-validation runs. The $\chi^2$ statistic with a significance level is used to test for statistical independence in the PC portion of PCX. The $p$- value is used to select and order the associated nodes in the RAMSEY subprocedure of PCX.

**PCX** (Data $D_1$, Data $D_2$, Target $T$, Depth $d$, Alpha $\alpha$)

"returns a graphical Markov Blanket structure (*MB*) for $T$ : MB(T) and the confusion matrix *M*"

V ($T$) = RAMSEYProcedure($T$)

**Repeat**

V ($Vi$) = RAMSEYProcedure($Vi$), where $Vi \in$ V ($T$)

**Until** no more variables are left to consider

V = V ($T$) $\cup$ V($v_i$) ($i$=1…m)

Run the PC algorithm over V

//The result is a pattern $P$, possibly with double headed or undirected edges

$\forall$ double headed edge $Vi \leftrightarrow T \in P$, or undirected edge $Vi - T \in P$ :

replace the edge with $T \rightarrow Vi$

Delete all edges adjacent to parents of $T$, except for the edges from the parent to $T$

$\forall$ double headed edge $Vi \leftrightarrow Vj \in P$, or undirected edge $Vi - Vj \in P$ :

**If** $Vj$ is a child of $T$ and $Vi$ is not

replace the edge with an edge $Vi \rightarrow Vj$

**else if** ($Vj$ and $Vi$ are both the children of $T$) or (neither $Vj$ nor $Vi$ are the children of $T$)

delete this edge

Delete all edges into parents of $T$ or parents of children of T

Delete all edges out of children of $T$

For each remaining node that is neither a parent of $T$, nor a child of $T$, nor a parent of a child of $T$, delete the node.

//the resulting graph is a Markov Blanket of *T: MB(T)*

Classify cases by Bayesian updating using *MB(T)*

Return *MB(T)* and *M*.

**RAMSEYProcedure**(Target $X$)

"returns a set of associated nodes to $X$: V($X$)"

For $X$, find the set of variables V ($X$) that are associated with $X$; Order them by their strength of association with $X$ by the associated $p$- value.

For each $Vi$ in V ($X$),

**if** ( $\exists Vj$ in V($X$), so that $\perp$ ($Vi$ ; $X$ | $Vj$ ) ) or ( $\exists Vj$ and $Vk$ in V($X$), so that $\perp$ ($Vi$ ; $X$ | $Vj$ , $Vk$ ) )

remove $Vi$ from V($X$).

//Test the variables with lowest association with $X$ first, and condition first on variables with highest association with $X$.

Return V ($X$)

**Figure 2**: Pseudo-code for PCX

## 6. RESULTS

In Table 2 We show the task-specific performance averaged over cross-validation runs. The results by PCX are compared with the results by HITON algorithm averaged over the classifiers. We give the average AUC, the average prediction accuracy and the number of predictor variables.

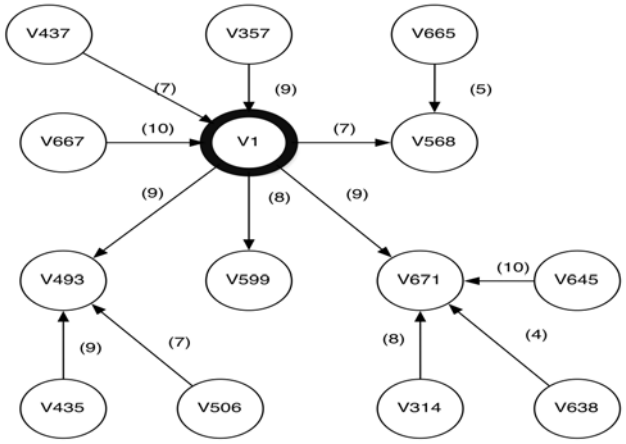
We show the best fitting MB DAG for the five experiments (figure 3). The number of times each edge occurs in all repeated cross validations is shown in parentheses.

**Table 2. Average performance comparison (PCX / HITON over the state-of-the-art classifiers[2])**

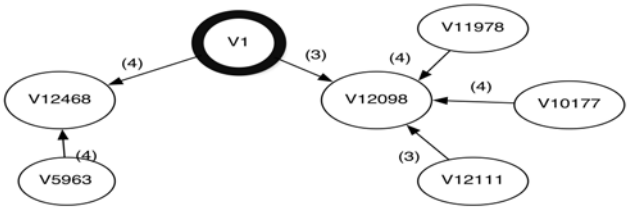
| Data sets | Thrombin | Arrhythmia | Ohsumed | Lung Cancer | Prostate Cancer |
|---|---|---|---|---|---|
| **Accuracy** | 94.50 / NA | 63.39 / 65.85 | 89.5 / NA | 94.21 / NA | 93.76 / NA |
| **AUC** | 82.18 / 92.70 | NA/NA | 81.22/ 83.04 | 92.28 / 97.60 | 94.45/ 96.14 |
| **#Predictor Variables** | 23 / 32 | 12 / 63 | 22 / 34 | 7 / 16 | 13 /16 |
| **Variable Reduction[3]** | 6059 / 4354 | 23.3 / 4.4 | 653.3 / 422.7 | 1800 / 787.5 | 59.9 / 48.7 |
| **C. V. Folds** | 1 | 10 | 1 | 5 | 10 |

[2] Classifiers include polynomial-kernel, Support Vector Machines, K-Nearest Neighbors , Feed-forward Neural Networks, Decision Trees, Naïve Bayes Classifier

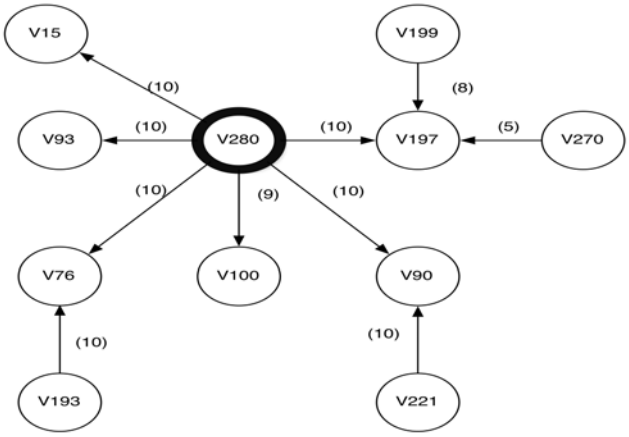
[3] Variable reduction = the original number of the variables / the number of the predictor variables




(a) Markov Blanket DAG for Prostate Cancer

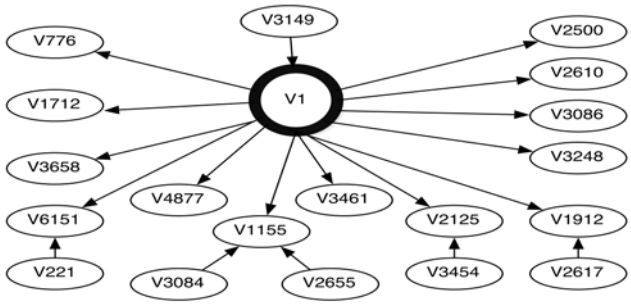



(c) Markov Blanket DAG for Lung Cancer

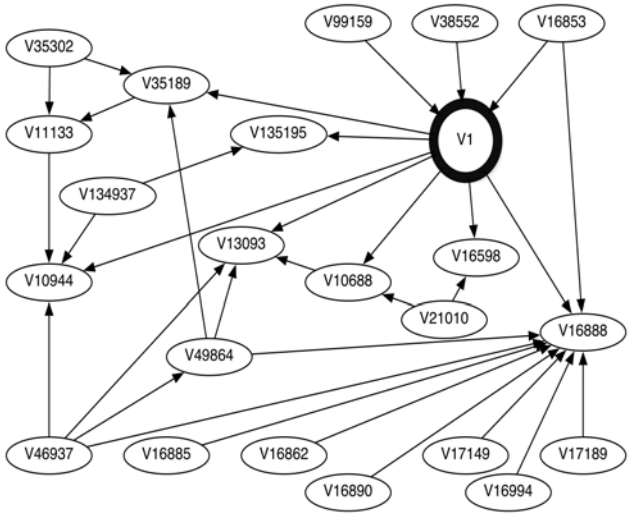



(b) Markov Blanket DAG for Arrhythmia

**Figure 3: MB DAGs**




(d) Markov Blanket DAG for Ohsumed




(e) Markov Blanket DAG for Thrombin

**Figure 3(cont.): MB DAGs**

## 7. DISCUSSION

On average PCX reduces the set of predictor variables to 46% of those used in HITON, in some cases to a sufficiently small set for entry into hand calculators or paper and pencil decision procedures in clinical settings and simplifying genetic marker identification. On the four of the five empirical data sets it procedures excellent classification results using the most probable value of the target variable as classification criterion. The Arrhythmia data set is difficult for all classifiers considered. The AUC results for PCX are slightly poorer than the average of the classifiers used in the Aliferis, et al. study. The worst one, the thrombin data set, yields 10.5% lower AUC than HITON. In each data set some directed edges are robust over almost all cross validation runs, but there is considerable variation. Heuristic Bayesian search procedures could provide probability values assigned to each edge, but nothing seems to be known about the

calibration of such probabilities—i.e., how frequently an edge with a given probability occurs in searches over datasets from a wide sample of DAGs and with varying sample sizes.

It is possible that different graphical MBs consistent with the PC output would give slightly different classification results. The rules used in converting PC output to a graphical MB are chiefly arbitrary in these respects (1) Undirected and bidirected edges between a child of the target and another non-target, non-child variable are replaced by edges directed into the child. This shows a preference for including selected variables in the final MB. (2) Undirected and bidirected edges adjacent to the target variable are replaced by directed edges out of the target variable. This shows a preference for a smaller number of parameters. (3) Edges between children are deleted, principally to avoid the trouble of checking for cycles. In any particular case these choices might be suboptimal decisions and an iterative post-search that investigates alternative orientations and further edge additions among the final variables might be preferable.

The removal of bidirected edges from the graphical MB involves a loss of potentially important information about causal structure, since with PC such edges indicate that unrecorded variables contribute to the association between the two adjacent measured variables. Hence the causal claims implied by the MB DAGs shown in our figures cannot be taken literally in many cases. A sound algorithm, FCI, is available for identifying aspects of causal structure when latent variables may be present, and could be substituted for PC in the PCX algorithm, but it is considerably slower than PC.

It should be noted that the part of the HITON procedure we have used does not in general correctly identify variables adjacent to the target, although in principle it correctly identifies the *union* of the set of variables adjacent to the target and the set of variables adjacent to those variables.

The PCX algorithm could easily be improved in several ways. (1) the speed of the algorithm could be increased by giving the PC algorithm information about which edges were removed by the HITON procedure--those removals are sound, and the PC algorithm accepts such background knowledge; (2) the procedure could be followed by a heuristic search that changes the directions of some edges, particularly making children of the target into parents, and possibly adds edges; (3) it is impossible to simultaneously estimate parameters in MB DAG models in which there are many multivalued parents because the contingency tables become too large to store, a limitation that could be overcome by dynamical maximum likelihood estimation of the conditional probability for each test instance as it arises; (4) a few cases in the test set for which the predictors have values that do not occur in the training set are simply passed in the present implementation—they could be estimated uninformatively with uniform Dirichelet priors instead of maximum likelihood, but that would essentially give the same prediction for every such case; better, in keeping with (3), the probability of the target for such could instead be estimated by the average of the probability of the target on the nearest similar cases represented in the training set, weighted by the frequency of their occurrence. We have not implemented any of these improvements for lack of time.

This work, and previous work on producing a graphical Markov Blanket for classification, does not address the interesting problem of simultaneously building classifiers for all variables in a large variable data set, or the problem of discovering a causal model for all variables in such

data. A variety of heuristic procedures have been proposed, but no sound procedure that is faster than PC seems to be known as yet.

## 8. ACKNOWLEDGMENTS

The authors thank Constantin Aliferis and Alexander Statnikov from the Department of Biomedical informatics, Vanderbilt University for their generous and invaluable assistance.

## 9. REFERENCES


[1] Spirtes, P., C. Glymour, and R. Scheines, *Causation, Prediction, and Search.* Springer Lecture Notes in Statistics, 1993; 2nd edition, The MIT Press, 2000

[2] Aliferis C.F., I. Tsamardinos, Statnikov A., HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection, Technical report DSL-03-08, Vanderbilt University, 2003; also, *AMIA*, 2003.

[3] Pearl, J. *Causality*. Oxford University Press, 2000.

[4] D.M. Chickering (2002). Learning Equivalence Classes of Bayesian-Network Structures. *Journal of Machine Learning Research*, 2:445-498. *of Machine Learning Research*, 2:445-498.

[5] Koller, D. and M. Sahami, Towards Optimal Feature Selection, *Proceedings of the 13th International Conference on Machine Learning (ML)*, Bari, Italy, July 1996, pages 284--292.

[6] Margaritis, D. and S.Thrun, Bayesian Network Induction via Local Neighborhoods., Advances in Neural Information Processing System (NIPS) 12, 1999.

[7] Aliferis, C. and I. Tsamardinos, Algorithms for Large-Scale Local Causal Discovery in the Presence of Small Sample or Large Causal Neighborhoods, Technical report DSL-02-08, Vanderbilt University, 2002.

[8] Spirtes, P., and Meek, C. "Learning Bayesian Networks with Discrete Variables from Data", in *Proceedings of The First International Conference on Knowledge Discovery and Data Mining*, ed. by Usama M. Fayyad and Ramasamy Uthurusamy, AAI Press, pp. 294-299, 1995.

[9] Cheng, J. et al. KDD CUP 2001 Report. SIGKDD Explorations, 3 (2): 1 – 18, 2002.

[10] Guvenir, H. et al. A Supervised Machine Learning Algorithm for Arrhythmia Analysis. Proceedings of Computers in Cardiology, Lund Sweden, 1997.

[11] Hersh, W., et al. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. Proceedings of the 17th Annual ACM SIGR Conference on Research and Development in Information Retrieval, 192-201, 1994

[12] Bhattacharjee, A., et al. Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses. Proceedings of the National Academy of Science, 98 (24(: 13790-5, 2001.

[13] Adam, B. et al., Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men. Cancer Research 62, 3609-3614, 2002.

[14] Meek, C., Strong completeness and faithfulness in Bayesian *networks*, Uncertainty in AI, 1995.