MEDINFO 2007
K. Kuhn et al. (Eds)
IOS Press, 2007

A New Machine Learning Classifier for High Dimensional Healthcare Data

Rema Padman[a] , Xue Bai[a,b] and Edoardo M. Airoldi[b]

[a]The H. John Heinz III School of Public Policy and Management, Carnegie Mellon University, U.S.A.
[b]Center for Automated Learning and Discovery, Carnegie Mellon University, U.S.A.

Abstract

Data sets with many discrete variables and relatively few cases arise in health care, ecommerce, information security, and many other domains. Learning effective and efficient prediction models from such data sets is a challenging task. In this paper, we propose a new approach that combines Metaheuristic search and Bayesian Networks to learn a graphical Markov Blanket-based classifier from data. The Tabu Search enhanced Markov Blanket (TS/MB) procedure is based on the use of restricted neighborhoods in a general Bayesian Network constrained by the Markov condition, called Markov Blanket Neighborhoods. Computational results from two real world healthcare data sets indicate that the TS/MB procedure converges fast and is able to find a parsimonious model with substantially fewer predictor variables than in the full data set. Furthermore, it has comparable or better prediction performance when compared against several machine learning methods, and provides insight into possible causal relations among the variables.

Keywords:

Markov Blanket, Bayesian Networks, machine learning, Tabu Search, health care decision support.

Introduction

The deployment of comprehensive information systems and online databases has made extremely large collections of real-time data readily available. In many domains such as genetics, clinical diagnoses, direct marketing, finance, and on-line business, data sets arise with thousands of variables and a small ratio of cases to variables. Such data present dimensional difficulties for classification of a target variable, and identification of critical predictor variables [1]. Furthermore, they pose even greater challenges in the determination of actual influence, i.e., causal relationships between the target variable and predictor variables. The problem of identifying essential variables is critical to the success of decision support systems and knowledge discovery tools due to the impact of the number of variables on the speed of computation, the quality of decisions, operational costs, and understandability and user acceptance of the decision model. For example, in medical diagnosis and healthcare decision support, the elimination of redundant tests may reduce the risks to patients and lower healthcare costs [2]. In this study, we address this problem of efficiently identifying a small subset of predictor variables from among a large number, and estimating the causal relationship between the selected variables and the target variable, using *Markov Blanket* (MB) and *Tabu Search* (TS) approaches.

We propose a two-stage Tabu Search enhanced Markov Blanket procedure that finds a parsimonious MB Directed Acyclic Graph (DAG). This two-stage algorithm generates an MB DAG in the first stage as a starting solution; in the second stage, the Tabu Search metaheuristic strategy is applied to improve the effectiveness of the MB DAG as a classifier, with conventional Bayesian updating. Classification using the Markov Blanket of a target variable in a Bayesian Network has important properties: it specifies a statistically efficient prediction of the probability distribution of a variable from the smallest subset of variables; it provides accuracy while avoiding over-fitting due to redundant variables; and it provides both a classifier and some insight into causal relations between a reduced set of predictors and the target variable. The TS/MB procedure proposed in this paper allows us to move rapidly through the search space of Markov Blanket structures and escape from local optima, thus learning a more robust structure.

Background knowledge

A *Bayesian Network* is a graphical representation of the joint probability distribution of a set of random variables. A Bayesian Network for a set of variables $X = \{X_1, ..., X_n\}$ consists of: (i) a directed acyclic graph (DAG) S that encodes a set of conditional independence assertions among variables in X; (ii) a set $P = \{p_1, ..., p_n\}$ of local conditional probability distributions associated with each node and its parents.

P satisfies the *Markov condition* [3] for S if every node X_i in S is independent of its non-descendants and non-parents in S, conditional on its parents. The Markov Condition implies that the joint distribution p can be factorized as a product of conditional probabilities, by specifying the distribution of each node conditional on its parents. In particular, for a given structure S, the joint probability distribution for X can be written as

$$p(X) = \prod_{i=1}^{n} p_i(x_i \mid pa_i) \tag{1}$$

where pa_i denotes the set of parents of X_i; this is called a Markov factorization of *P* according to *S*.

Given the set of variables *X* and target variable *Y*, a *Markov Blanket* (MB) for *Y* is the smallest subset *Q* of variables in *X* such that *Y* is independent of *X\Q*, conditional on the variables in *Q*. *P* is *faithful* to the graph *S* with the vertex set *X* if and only if there are no conditional independence relations in *P* other than those entailed by satisfying the Markov condition for *S*. If *P* is faithful to the graph *S*, then given a Bayesian Network (*S*, *P*), **there is** a unique Markov Blanket for *Y* consisting of the set of parents of *Y*; the set of children of *Y*; and the set of parents of children of *Y*.

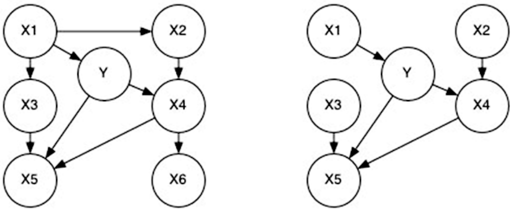


Figure 1 - (left) A Bayesian Network (S, P), and (right) the Markov Blanket for the variable Y

For example, given the two DAGs in Figure 1, the factorization of *p* entailed by the Bayesian Network (*S*, *P*) is

$$p(Y,X_1,...,X_6) = p(Y \mid X_1) \cdot p(X_4 \mid X_2, Y) \cdot p(X_2 \mid X_1) \cdot \\ p(X_5 \mid X_3, X_4, Y) \cdot p(X_3 \mid X_1) \cdot p(X_6 \mid X_4) \cdot p(X_1) \quad (2)$$

The factorization of the conditional probability for the Markov Blanket of *Y*: $p(Y \mid X_1, \ldots ,X_6)$ is the product of those (local) factors in equation above that contain the term *Y*:

$$p(Y \mid X_1,...,X_6) = C \cdot p(Y \mid X_1) \cdot p(X_4 \mid X_2, Y) \cdot \\ p(X_5 \mid X_3, X_4, Y) \cdot, \quad (3)$$

where *C* is a normalizing constant independent of *Y*.

Recent research in classification problems has tried to identify the Markov Blanket variables of the target class variable by filtering predictor variables using statistical tests of conditional independence [4]. Few of the proposed methods, however, have been used to generate the Markov Blanket from real-world data, and tests on real-world problems have usually involved only a small number of variables. Furthermore, while all the previous studies use the notion of Markov Blanket as a minimal set of dependent variables, none of them actually generate and retain the graphical structure of the Markov Blanket corresponding to a specific data set, nor have they used the structure for Bayesian inference in order to perform the classification. Our algorithm addresses all these limitations. Details are available in [5].

Tabu Search is a powerful meta-heuristic strategy that helps local search heuristics to explore the solution space by guiding them out of local optima [6]. Its strategic use of memory and responsive exploration is based on selected concepts that cut across the fields of artificial intelligence and operations research. It has been applied successfully to a wide variety of continuous and combinatorial optimization problems, capable of reducing the complexity of the search process and accelerating the rate of convergence. In its simplest form, Tabu Search starts with a feasible solution and chooses the *best move* according to an evaluation function while taking steps to ensure that the method does not re-visit a solution previously generated. This is accomplished by introducing *tabu restrictions* on possible moves to discourage the reversal, and, in some cases, repetition of selected moves. The *tabu list* that contains these forbidden move attributes is known as the short term memory function. It operates by modifying the search trajectory to exclude moves leading to new solutions that contain attributes (or attribute mixes) belonging to solutions previously visited within a time horizon governed by the short term memory. Intermediate and long-term memory functions may also be incorporated to intensify and diversify the search.

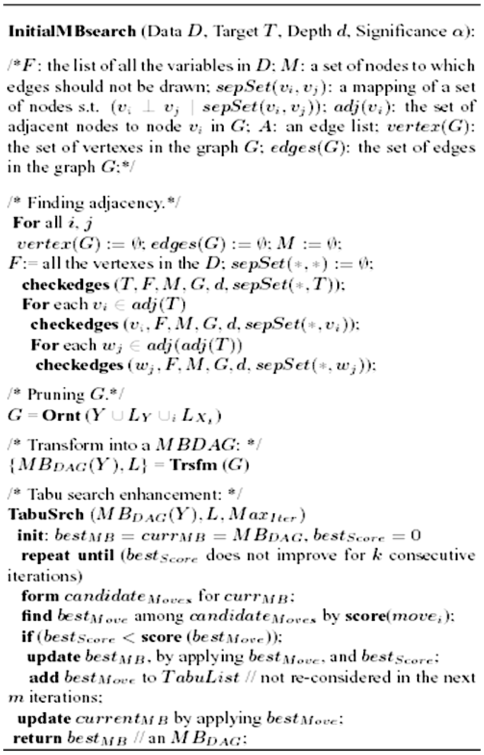
TS/MB model

A sketch of the algorithm is presented in Figure 2. The detailed algorithm is presented in [5]. Our algorithm first generates an *initial Markov Blanket* for the target variable. However, the initial MB may be highly suboptimal due to the application of repeated conditional independence tests and propagation of errors in causal orientation [5, 7]. Therefore, Tabu Search is applied to improve the initial MB. Four kinds of moves are considered in the procedure: edge addition, edge deletion, edge reversal and edge reversal with node pruning. At each stage, for each allowed move, the resulting Markov Blanket is computed, factored, its predictions scored, and the current MB modified with the best move. The algorithm stops after a fixed number of iterations or a fixed number of non-improving iterations.

Computational results

We tested our algorithm on two biomedical data sets [8, 9]. Table 1 provides a brief characterization of the data sets. Prostate cancer (PCA) data set concerns diagnosis of prostate cancer from mass spectroscopy of human sera [8]. Arrhythmia data set concerns classification of subjects into 8 disease categories from clinical and EKG data [9].

The parameters in our experiments are: data-splits, scoring criteria, starting solution structure, the depth of conditional independence search (***d***), and significance level (α). We use a nested, stratified cross-validation scheme [10]. In the inner layer, the procedure trains and optimizes the Markov Blanket on training data for each parameter configuration. The configuration that yields the best MB according to the scoring criterion is chosen as the best configuration. The outer layer of cross-validation estimates the performance of the optimized Markov Blanket classifier on the testing data. We report both the AUC and prediction accuracy on the testing set to evaluate the classification performance of the generated models.



```
InitialMBsearch (Data $D$, Target $T$, Depth $d$, Significance $\alpha$):

/*$F$: the list of all the variables in $D$; $M$: a set of nodes to which
edges should not be drawn; $sepSet(v_i, v_j)$: a mapping of a set
of nodes s.t. $(v_i \perp v_j \mid sepSet(v_i, v_j))$; $adj(v_i)$: the set of
adjacent nodes to node $v_i$ in $G$; $A$: an edge list; $vertex(G)$:
the set of vertexes in the graph $G$; $edges(G)$: the set of edges
in the graph $G$;*/

/* Finding adjacency.*/
 For all $i, j$
  $vertex(G) := \emptyset$; $edges(G) := \emptyset$; $M := \emptyset$;
$F$:= all the vertexes in the $D$; $sepSet(*, *) := \emptyset$;
   checkedges ($T, F, M, G, d, sepSet(*, T)$);
   For each $v_i \in adj(T)$
     checkedges ($v_i, F, M, G, d, sepSet(*, v_i)$);
     For each $w_j \in adj(adj(T))$
      checkedges ($w_j, F, M, G, d, sepSet(*, w_j)$);

/* Pruning $G$.*/
$G$ = Ornt ($Y \cup L_Y \cup_i L_{X_i}$)

/* Transform into a $MBDAG$: */
$\{MB_{DAG}(Y), L\}$ = Trsfm ($G$)

/* Tabu search enhancement: */
TabuSrch ($MB_{DAG}(Y), L, Max_{Iter}$)
  init: $best_{MB} = curr_{MB} = MB_{DAG}, best_{Score} = 0$
   repeat until ($best_{Score}$ does not improve for $k$ consecutive
iterations)
   form $candidate_{Moves}$ for $curr_{MB}$;
   find $best_{Move}$ among $candidate_{Moves}$ by score($move_i$);
   if ($best_{Score} <$ score ($best_{Move}$));
    update $best_{MB}$, by applying $best_{Move}$, and $best_{Score}$;
     add $best_{Move}$ to $TabuList$ // not re-considered in the next
$m$ iterations;
  update $current_{MB}$ by applying $best_{Move}$;
 return $best_{MB}$ // an $MB_{DAG}$;
```

Figure 2 - A sketch of the TS/MB algorithm

Table 1 - Characteristics of Data Sets

Task	Vars.	Samp.	Var. Type	Target Var.
PCA diagnosis	779	326	Discretized	Binary
Arrhythmia diagnosis	279	417	Ordinal	8 Categories

Table 2 presents the average best-fitting classification results for PCA data and the comparison against several state-of-the-art classifiers in three different ways. Details of the classifiers and the motivation for their choice in this study are discussed in [5]. We report both the AUC[1] and prediction accuracy on the testing set as well as the size of reduction in the set of variables. The size reduction was evaluated based on the fraction of variables in the resulting models. All metrics (variable size reduction, AUC, and accuracy) were averaged over cross-validation splits.

Comparison I (columns 1 and 2) presents the results when using the *full set* of variables as input. Comparison II (columns 3 and 4) uses the *same number* of variables as identified by TS/MB, as input for all the other classifiers. These variables for the classifiers in the comparison set are selected using information gain (IG) criterion [11]. Comparison III (columns 5 and 6) uses the *exact same* variables identified by TS/MB as input variables for all classifiers. Comparisons II and III are used to test the source of the differential in the observed accuracy and AUC values in the full data set.

[1] Since AUC is only applicable to binary classification problems, we report only the accuracy for Arrhythmia data.

*Table 2 - Five-fold cross validation results of various classifiers on 779 peaks; on 19 peaks selected by information gain; and on the exact same 19 peaks selected by the TS/MB classifier for the PCA data. Best performance figures are in **bold***

Input	**All Peaks**		**Peaks selected by IG**		**Peaks selected by TSMB**		
Method	**AUC %**	**Accuracy %**	**AUC %**	**Accuracy%**	**AUC %**	**Accuracy%**	**# Peaks selected**
MB	95.3	87.1	95.3	87.1	95.3	87.1	19
TS/MB	98.3	90.3	98.3	90.3	98.3	90.3	19
Naïve Bayes	97.5	89.3	67.5	63.2	77.4	69.4	
SVM	97.1	98.5	63.3	62.0	72.6	69.9	
Voted Perceptron	73.9	58.0	65.2	59.2	75.4	67.8	
Max. Entropy	87.4	98.8	64.7	64.1	74.9	70.9	
K-NN	96.3	88.6	65.6	58.6	72.1	65.3	
Logistic Reg.	Failed	Failed	73.6	56.4	98.1	90.0	

The results clearly indicate that TS/MB dominates on the AUC metric and performs well on the accuracy measure, except in one instance. In addition, there is an almost 98% reduction in the number of variables used for this prediction. It is also interesting to note that all competing classifiers perform better with the variables provided by TS/MB than using the IG criterion (Comparison II vs. Comparison III) on both accuracy and AUC. Figure 3 shows the MB DAG learned from PCA data that achieves the best accuracy on the testing data. All the directed edges are robust over almost all cross validation runs, with very small variation. Further study and interaction with clinicians is necessary to identify the clinical significance of these variables in actual settings.

Table 3 presents the average best-fitting classification results for Arrhythmia data and compares them against the results obtained from several state-of-the-art classifiers. Figure 4[2] shows the MB DAG that achieves the best accuracy on the testing set of Arrhythmia data. The classifier in [9] achieved an accuracy of 62%, which obviously is not sufficiently good for clinical use. The best result from TS/MB achieved average accuracy of 96.8% with a 95% reduction in the number of variables required for the prediction, a significant improvement over the earlier study.

Discussion and conclusion

On average, the TS/MB classifier reduces the set of predictor variables by at least 95% from the full set of variables. In some cases it is reduced to a sufficiently small set for entry into hand calculators, or paper and pencil decision procedures that are easy to use in clinical and other decision settings. At the same time, when compared to the state-of-the-art classification methods, the TS/MB classifier procedures excellent classification results, especially in real world applications where the cost of misclassification has significant implications. Moreover, the algorithm generates a graphical structure that represents the relationships between the variables and provides additional insight into causal discovery. These experiments, as well as more results we have obtained on data sets from other domains, such as Internet marketing and sentiment extraction, suggest that for problems where the ratio of samples to the number of the variables is small or the independence assumption is not appropriate, the two-stage MB classifier is superior in terms of the prediction performance, effectiveness in identifying critical predictors, and robustness [5].

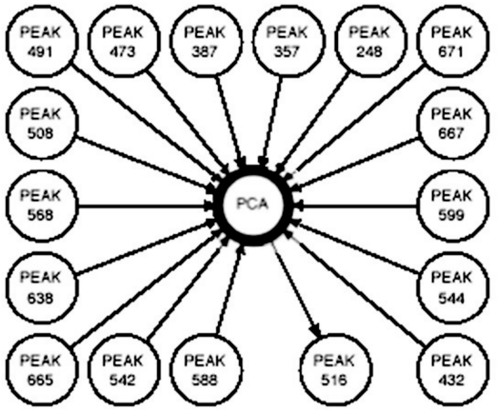


Figure 3 - The best fitting MB DAG for the PCA

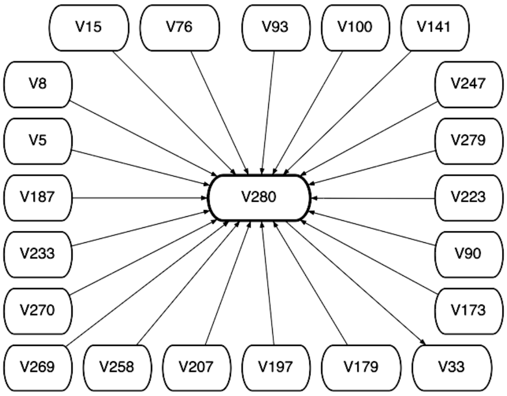


Figure 4 - The best fitting MB DAG for arrhythmia

It is possible that different Markov Blanket graphical structures that are consistent with the TS/MB classifier output would give slightly different classification results. Because any undirected and bi-directed edges are deleted after the edge orientation step, these deletions might result in suboptimal decisions. Tabu Search iteratively investigates alternative orientations and further edge additions to minimize the extent of sub-optimality. On the other hand, theoretically, two DAGs that have the same Markov factorization are Markov equivalent. In our case, two Markov Blankets can be Markov equivalent even if some edges are oriented in a different, but statistically non-differentiable way.

The results of this study need to be validated against clinicians' interpretation of the variables and their value in actual diagnosis settings. This research can also be extended to address the interesting problem of simultaneously building classifiers for all variables in a large variable data set, discovering a causal model for all variables in such data, or automatically classifying medical documents into categories. Future research will explore these issues.

[2] The class variable "V280" has 8 categories, encoding 8 different diseases.

Table 3 - Five-fold cross validation results of various classifiers using 279 variables; using 21 variables selected by information gain; and using the exact same 21 variables selected by the TS/MB classifier for the Arrythmia data. Best performance figures are in ***bold***

Input	**All vars.**	**Vars. Selected by IG**	**Vars. Selected by TS/MB**	
Method	**Accuracy %**	**Accuracy %**	**Accuracy %**	**# peaks selected**
MB	77.6	77.6	77.6	**20**
TS/MB	**96.8**	**96.8**	**96.8**	**21**
Naïve Bays	57.0	57.0	57.0	
SVM	93.3	69.4	72.6	
Voted Perception	72.3	71.5	71.3	
Max. Entropy	96.4	74.7	76.4	
K-NN	71.3	76.6	75.1	
Logistic Reg.	80.9	72.0	74.0	

Acknowledgements

We thank Professor Fred Glover of the University of Colorado at Boulder, Professor Clark Glymour and Professor Peter Spirtes of Carnegie Mellon University, and Dr. Frank Wimberly at Institute for Human and Machine Cognition for valuable suggestions throughout the project; and Professor Constantin Aliferis of Vanderbilt University for providing us with the PCA data set.

References

[1] Berry M and Linoff G. Data Mining Techniques: For Marketing, Sales, and Customer Support. Wiley, 1997.

[2] Cooper GF, Aliferis C, Aronis J, Buchanan B, Caruana R, Fine M, Glymour C, Gordon G, Hanusa B, Janosky J, Meek C, Mitchell T, Richardson T, Spirtes P. An evaluation of machine-learning methods for predicting pneumonia mortality. Artificial Intelligence in Medicine 1992; 9: 107–139.

[3] Pearl J. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2000.

[4] Aliferis C, Tsamardinos I, Statnikov A. Hiton, a novel markov blanket algorithm for optimal variable selection. Proceedings of the American Medical Informatics Association Annual Symposium, 2003b, 21–25.

[5] Bai X, Padman R. Tabu search enhanced markov blanket classifier for high dimensional data sets. In Proceedings of INFORMS Computing Society, pp. 337-354. Kluwer Academic Publisher, 2005.

[6] Glover F. Tabu Search. Kluwer Academic Publisher, 1997.

[7] Spirtes P, Glymour C and Scheines R. Causation, Prediction, and Search. MIT Press, 2000.

[8] Adam B, Qu Y, Davis J, Ward M, Clements M, Cazares L, Semmes O, Schellhammer P, Yasui Y, Feng Z, Wright G. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Research 2002; 62:3609–14.

[9] Güvenir HA, Acar A, Demiröz G, Cekin A. A supervised machine learning algorithm for arrhythmia analysis. Computers in Cardiology 1997; 24: 433-36.

[10] Weiss S. and Kulikowski C. Computer Systems That Learn. Morgan Kaufmann, 1991.

[11] Mitchell T. Machine Learning. McGraw-Hill. 1997.

Address for correspondence

Rema Padman,
The Heinz School,
Carnegie Mellon University,
Pittsburgh, PA 15213, USA.
Email: rpadman@cmu.edu