Speculative Markov Blanket Discovery for Optimal Feature Selection

Sandeep Yaramakala
Department of Computer Science
Iowa State University
Ames, IA 50011, USA
ysandeep@cs.iastate.edu

Dimitris Margaritis
Department of Computer Science
Iowa State University
Ames, IA 50011, USA
dmarg@cs.iastate.edu

Abstract

In this paper we address the problem of learning the Markov blanket of a quantity from data in an efficient manner. Markov blanket discovery can be used in the feature selection problem to find an optimal set of features for classification tasks, and is a frequently-used preprocessing phase in data mining, especially for high-dimensional domains. Our contribution is a novel algorithm for the induction of Markov blankets from data, called Fast-IAMB, that employs a heuristic to quickly recover the Markov blanket. Empirical results show that Fast-IAMB performs in many cases faster and more reliably than existing algorithms without adversely affecting the accuracy of the recovered Markov blankets.

1. Introduction

It is often the case that an engineer or researcher is interested in one particular attribute in a set of observations. To analyze and possibly predict the value of this attribute, he or she needs to first ascertain which of the other attributes in the domain affect it. This task is frequently referred to as the *feature selection problem*. A solution to this problem is often non-trivial, and can be infeasible when the domain is defined over a large number of attributes.

A principled solution to the feature selection problem is to determine a subset of attributes that can "shield" (render independent) the attribute of interest from the effect of the remaining attributes in the domain. Koller and Sahami [4] first showed that the Markov blanket of a given target attribute is the theoretically optimal set of attributes to predict its value.

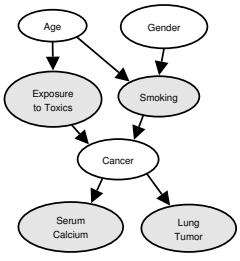
Because the Markov blanket of a target attribute T renders it statistically independent from all the remaining attributes (see the Markov blanket definition below), all information that may influence its value is stored in the values of the attributes of its Markov blanket. Any attribute from the feature set outside its Markov blanket can be effectively ignored from the feature set without adversely affecting the performance of any classifier that predicts the value of T.

Definition 1 (Markov blanket). *A Markov blanket* $\mathbf{B}(T)$ *of an attribute* $T \in \mathcal{U}$ *is any subset* $\mathbf{S}$ *of attributes for which*

$$(T \perp\!\!\!\perp \mathcal{U} - \mathbf{S} - \{T\} \mid \mathbf{S}) \text{ and } T \notin \mathbf{S}. \quad (1)$$

A set is called a ***Markov boundary*** *of* T *if it is a minimal Markov blanket of* T*,* i.e.*, none of its proper subsets satisfy Equation (1).*

In this paper, we identify the Markov blanket of an attribute with its Markov boundary. We use capitals (*e.g.*, X, Y, etc) to indicate attributes and bold letters to indicate sets of attributes *e.g.*, $\mathbf{Z}$. T is the target attribute whose blanket we are seeking, while $\mathbf{B}$(T) is its Markov blanket (a set of attributes). The notation $(X \perp\!\!\!\perp Y \mid \mathbf{Z})$ denotes that X and Y are conditionally independent given $\mathbf{Z}$. Similarly, $(X \not\perp\!\!\!\perp Y \mid \mathbf{Z})$ denotes conditional dependence. r_X and $r_{\mathbf{Z}}$ denote the number of values that attribute X and the variables in set $\mathbf{Z}$ (jointly) take, respectively. The number of records in the data set is denoted as N. $\mathcal{U}$ is the set of all attributes in the domain.



In this paper we assume that the data were generated by a single faithful directed graphical model (a Bayesian network), that models the domain. A Bayesian network is a statistical model that is capable of graphically representing independencies that hold in a domain [6]. The existence of a *faithful* Bayesian network (see definition of faithfulness below) implies that the Markov blanket of any attribute in the domain is unique, and can be easily "read off" the network structure: The Markov blanket of an attribute is the set of *parents*, *children* and *spouses* (*i.e.*, parents of common children) as encoded by the graph structure of the Bayesian network. For example, the figure at the beginning of this paragraph shows a Bayesian network consisting of five attributes. The Markov blanket of

Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

the attribute *Cancer* is the set {*Exposure to Toxics*, *Smoking*, *Serum Calcium*, *Lung Tumor*} (the gray nodes in the figure). This set shields *Cancer* from the effects of those attributes outside it.

Definition 2 (Faithfulness). *A Bayesian network B and a joint distribution P are faithful to to one another iff every conditional independence entailed by the graph of B is also present in P* i.e.*,*

$$(\mathbf{X} \perp\!\!\!\perp_B \mathbf{Y} \mid \mathbf{Z}) \iff (\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z})$$

Theorem 1. *If a Bayesian network B is faithful, then for every attribute T, $\mathbf{B}(T)$ is unique and is the set of parents, children and spouses of T.*

The goal of this paper is to develop a fast algorithm for discovering Markov blankets from data. We emphasize that we do not address Bayesian network structure discovery here—Markov blankets are discovered without determining the structure of the underlying Bayesian network.

2. Related Work

Margaritis and Thrun [5] presented the first provably correct algorithm that discovers the Markov blanket of an attribute from data, under assumptions (see below). The Grow-Shrink Markov Blanket algorithm (GSBN), a Bayesian network structure induction algorithm, invokes the Grow-Shrink Markov blanket algorithm (called **GSMB**) for every attribute in the domain as a first step. It then utilizes knowledge of the Markov blankets recovered to make the actual Bayesian network structure discovery more efficient. As implied by its name, the GS Markov blanket algorithm contains two phases: a growing phase and a shrinking phase.

The GSMB algorithm has the desirable property that, under assumptions, is provably sound *i.e.*, it can recover the *exact* Markov blanket of any given attribute in the domain. The assumptions made are: (i) the existence of a faithful Bayesian network for the domain under consideration (this implies the existence and uniqueness of the blanket, see Theorem 1 above) and, (ii) the assumption that the conditional independence tests are correct.

Tsamardinos, Aliferis, and Statnikov [7] describe a number of variants of GSMB that aim at improved speed and reliability. We evaluate here the Incremental Association Markov Blanket (**IAMB**) and Interleaved IAMB (**Inter-IAMB**) algorithms. Like GSMB, the IAMB and Inter-IAMB algorithms also use a two-phase approach for discovering Markov blankets. However, they reorder the set of attributes each time a new attribute enters the blanket in the growing phase. This reordering is done using an information-theoretic heuristic function h (conditional mutual information). The motivating idea is that IAMB and its variants might perform better because (hopefully) fewer false positives will be added during the growing phase (that would have had to be removed during the shrinking phase).

3. The Fast-IAMB Algorithm

In this section we present a new algorithm for Markov blanket discovery, called **Fast-IAMB**. The Fast-IAMB algorithm is shown in Fig. 1.

```
 1: B(T) ← ∅
 2: S ← {A | A ∈ U − {T} and A ⫫̸ T}
 3: while (S ≠ ∅) do
 4:    ⟨X_1, ..., X_|S|⟩ ← S sorted according to h
 5:    insufficient_data ← FALSE
 6:    /* Growing phase. */
 7:    for i = 1 to |S| do
 8:       if N / (r_{X_i} × r_T × r_{B(T)}) ⩾ k then
 9:          B(T) ← B(T) ∪ {X_i}
10:       else
11:          insufficient_data ← TRUE
12:          goto 15    /* Insufficient data. */
13:       end if
14:    end for
15:    /* Shrinking phase. */
16:    for each attribute A ∈ B(T) do
17:       if (A ⫫ T | B(T) − {A}) then
18:          B(T) ← B(T) − {A}
19:       end if
20:    end for
21:    if insufficient_data = TRUE and [no attributes were re-
       moved in the shrinking phase] then
22:       halt
23:    else
24:       S ← {A | A ∈ U − {T} − B(T) and
                  (A ⫫̸ T | B(T))}
25:    end if
26: end while
```

Figure 1. The Fast-IAMB algorithm.

Similar to GSMB, IAMB, and Inter-IAMB, Fast-IAMB contains a "growing" phase (in which it attempts to add attributes to the blanket $\mathbf{B}$(T)), followed a "shrinking" phase (in which it attempts to remove as many irrelevant attributes as possible). During the growing phase of each iteration, it sorts the attributes that are candidates for admission to $\mathbf{B}(T)$ from most to least conditionally dependent, according to a heuristic function h. (This is similar to IAMB and Inter-IAMB; however, Fast-IAMB uses the more statistically appropriate significance of a G^2 conditional statistical test for h rather than the raw conditional information value, as IAMB and Inter-IAMB do). Each such sorting step is potentially expensive, since it involves the calcula-

tion of the G^2 test value between T and each member of $\mathbf{S}$: Each such calculation is equivalent to a conducted conditional independence test. The *key idea* behind Fast-IAMB is to reduce the number of such tests *by adding not one, but a number of attributes at a time* after each reordering of the remaining attributes following a modification of the Markov blanket. Fast-IAMB *speculatively* adds one or more attributes of highest G^2 test significance without re-sorting after each modification as IAMB and Inter-IAMB do, which (hopefully) adds more than one true members of the blanket. Thus, the cost of re-sorting the remaining attributes after each Markov blanket modification can be amortized over the addition of multiple attributes.

The Fast-IAMB algorithm is sound in that it discovers the exact Markov blanket under the same set of assumptions used by existing algorithms *viz.* the existence of a faithful (though not necessarily known) Bayesian network to the domain under consideration and the assumption that the conditional independence tests performed by the algorithm are correct. The proof of soundness is the same as that given by [7].

A natural question is to determine how many attributes should be added to the blanket at each iteration. We use the following heuristic: *we add dependent attributes as long as the conditional independence tests are reliable i.e.*, we have enough data for conducting them. For this purpose, we use a numeric parameter k that denotes the minimum average number of instances per cell of a contingency table that should be present for a conditional independence test to be deemed reliable. Let the next attribute that we consider for addition to $\mathbf{B}$(T) be X. To perform a reliable conditional independence test between T and X given $\mathbf{B}$(T), the average number of instances per cell of the contingency table of $\{X, T\} \cup \mathbf{B}(T)$ must be at least k, *i.e.*,

$$\frac{N}{r_T \times r_{\mathbf{B}(T)} \times r_X} \geqslant k. \qquad (2)$$

In all our experiments we choose $k = 5$ because, as suggested by Agresti [1], this is the minimum average number of instances per cell for the G^2 statistic to have a χ^2 distribution, which is a requirement for a significance (p-value) to be calculated. Also, note that in lines 7–14 no conditional independence tests are actually performed. The average number of instances per cell (line 8) calculation can be done in constant time.

One practical question remains: What is to be done if the average number of instances per cell for each remaining attributes is less than k? Tsamardinos and Aliferis [7] do not refer to this important practical question in their description of IAMB and Inter-IAMB. One has two choices: assume dependence or assume independence. While assuming dependence might seem to be the "safe" choice, in practice this would result in large blankets that are hard to justify and of little practical use. We therefore assume independence when the condition in Eq. (2) fails and halt (line 22) returning the current blanket. This does not adversely impact the performance of Fast-IAMB compared to IAMB and Inter-IAMB, as our experiments confirm.

4. Experimental Results

In order to empirically compare the performance of Fast-IAMB with the other Markov blanket discovery algorithms, we conducted a number of experiments on both synthetic and real-world data sets, listed below.

Dataset	No. of attributes	No. of records
HAILFINDER20K	56	20,000
ADULT	9	45,222
CENSUS-INCOME	34	142,521

HAILFINDER20K is a synthetic data set, while both ADULT [3] and CENSUS-INCOME [2] are well-known real-world data sets, containing demographic information.

The confidence level of each independence test was set to 95% ($\alpha = 0.05$). For each experiment, we report the number of conditional independence tests conducted to discover the blanket of each attribute of the domain, the total execution time taken by each algorithm, the distribution of the conditioning set sizes of the tests conducted, and a distance measure that indicates the "fitness" of the discovered blanket. We define the latter to be the average, over all attributes X outside the blanket, of the expected KL-divergence between $\Pr(T \mid \mathbf{B}(T))$ and $\Pr(T \mid \mathbf{B}(T) \cup \{X\})$. We can expect this measure to be close to zero when $\mathbf{B}(T)$ is an approximate blanket. This measure is similar to the one proposed by [4].

Fig. 2 (top row) shows that, in almost all cases, Fast-IAMB requires fewer conditional independence tests than either IAMB or Inter-IAMB. The number of tests directly influences the execution time of each algorithm (as expected), shown in Fig. 3. From this figure one can verify that Fast-IAMB executes faster than both IAMB and Inter-IAMB for all data sets: The running time of Fast-IAMB ranges from 68% to 82% of the execution time of IAMB, and 52% to 72% of Inter-IAMB.

Fig. 2 (middle row) shows that the blankets discovered by Fast-IAMB are approximately as good as IAMB and Inter-IAMB, as measured by the expected conditional KL-divergence between T and all attributes outside the blanket. This allows the blankets that are discovered Fast-IAMB to be used in comparable situations as IAMB and Inter-IAMB.

Fig. 2 (bottom row) shows the distribution of the sizes of the conditioning sets, where size is measured as the number of attributes in the conditioning set. In general, conditioning is undesirable since it typically results in less reliable independence tests. As can be seen from the figure, while the numbers of unconditional tests that all three algorithms are comparable, Fast-IAMB conducts significantly fewer con-

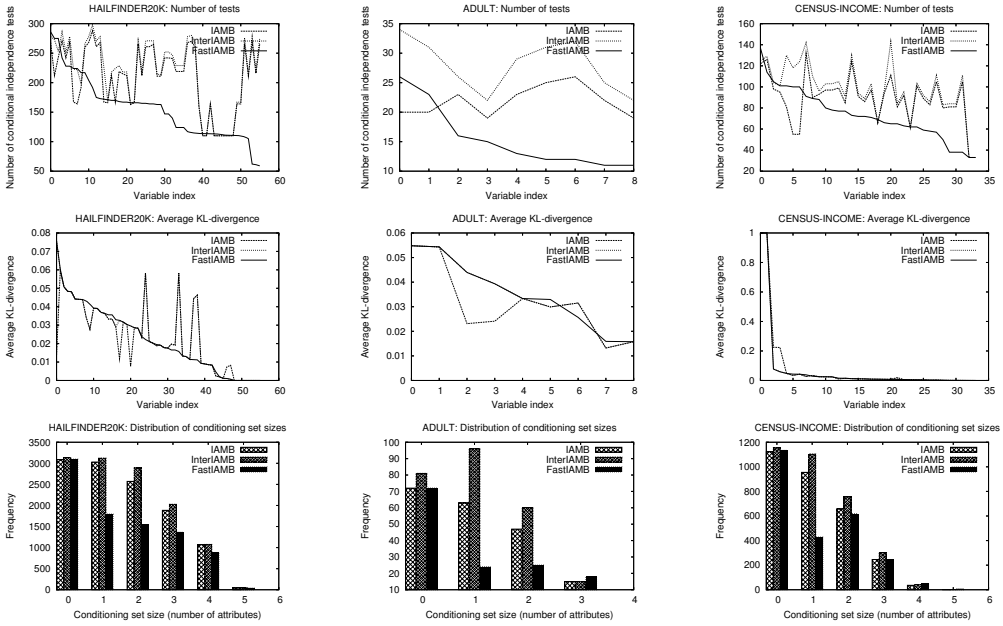
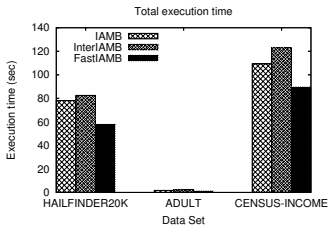


Figure 2. (*Top row*): Number of conditional independence tests for each attribute's blanket discovery in each data set. (*Middle row*): Average expected conditional KL-divergence, measuring the fitness of the recovered Markov blankets. (*Bottom row*): Distribution of conditioning set sizes of conditional tests conducted.



Algorithm	Exec. time (sec)
HAILFINDER20K data set	
IAMB	78.09
Inter-IAMB	82.52
Fast-IAMB	**57.83**
ADULT data set	
IAMB	1.71
Inter-IAMB	2.21
Fast-IAMB	**1.16**
CENSUS-INCOME data set	
IAMB	109.40
Inter-IAMB	123.17
Fast-IAMB	**89.58**

Figure 3. Total execution times for each data set.

ditional tests compared to IAMB and Inter-IAMB, which indicates improved test reliability.

5. Conclusion and Future Research

The main contribution of this paper is a novel algorithm for the induction of Markov blankets from data, called Fast-IAMB, that employs speculation to recover the Markov blanket faster. Our empirical results show that Fast-IAMB performs often faster and more reliably than existing algorithms, without adversely affecting the accuracy of the recovered Markov blankets. A direction of potential future research is relaxing the requirement of existence of a faithful underlying Bayesian network (which can be difficult to ascertain in practice) while maintaining the theoretical optimality of the recovered Markov blanket with respect to feature selection.

References

[1] A. Agresti. *Categorical Data Analysis*. New York: John Wiley and Sons Inc., 1990.

[2] S. Hettich and S. D. Bay. The UCI KDD archive, 1999. [http://kdd.ics.uci.edu/] UC Irvine, Dept. of ICS.

[3] S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases, 1998. [http://www.ics.uci.edu/∼mlearn/MLRepository.html] UC Irvine, Dept. of ICS.

[4] D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[5] D. Margaritis and S. Thrun. Bayesian network induction via local neighborhoods. In S. Solla, T. Leen, and K.-R. Müller, editors, *Proceedings of Conference on Neural Information Processing Systems (NIPS-12)*. MIT Press, 1999.

[6] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers Inc., 1988.

[7] I. Tsamardinos, C. Aliferis, and A. Statnikov. Algorithms for large scale Markov blanket discovery. In *The 16th International FLAIRS Conference*, St. Augustine, Florida, USA, 2003.