A Novel Scalable and Data Efficient Feature Subset Selection Algorithm

Sergio Rodrigues de Morais[1] and Alex Aussem[2]

[1] INSA-Lyon, LIESP, F-69622 Villeurbanne, France
sergio.rodrigues-de-morais@insa-lyon.fr
[2] Université de Lyon 1, LIESP, F-69622 Villeurbanne, France
aaussem@univ-lyon1.fr

Abstract. In this paper, we aim to identify the minimal subset of discrete random variables that is relevant for probabilistic classification in data sets with many variables but few instances. A principled solution to this problem is to determine the *Markov boundary* of the class variable. Also, we present a novel scalable, data efficient and correct Markov boundary learning algorithm under the so-called *faithfulness* condition. We report extensive empiric experiments on synthetic and real data sets scaling up to 139,351 variables.

1 Introduction

The identification of relevant subsets of random variables among thousands of potentially irrelevant and redundant variables with comparably smaller sample sizes is a challenging topic of pattern recognition research that has attracted much attention over the last few years [1, 2, 3]. By relevant subsets of variable, we mean the variables that conjunctly prove useful to construct an efficient classifier from data. This contrasts with the suboptimal problem of ranking the variables individually. Our specific aim is to solve the feature subset selection problem with thousands of variables but few instances using Markov boundary learning techniques. The Markov boundary of a variable T, denoted by $\mathbf{MB}_T$, is the minimal subset of $\mathbf{U}$ (the full set) that renders the rest of $\mathbf{U}$ independent of T.

Having to learn a Bayesian network G in order to learn a Markov boundary of T can be very time consuming for high-dimensional databases. This is particularly true for those algorithms that are asymptotically correct under faithfulness condition, which are the ones we are interested in. Fortunately, there exist algorithms that search the Markov boundary of a target without having to construct the whole Bayesian network first [3]. Hence their ability to scale up to thousands of variables. Unfortunately, they miss many variables due to the unreliability of the conditional independence tests when the conditioning set is large. Hence the need to increase data-efficiency of these algorithms, that is, the ability of the algorithm to keep the size of the conditional set as small as possible during the search.

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pp. 298–312, 2008.

In this paper, we discuss a divide-and-conquer method in order to increase the data-efficiency and the robustness of the Markov boundary (MB for short) discovery while still being scalable and correct under the faithfulness condition. The proposed method aims at producing an accurate MB discovery algorithm by combining fast, rough and moderately inaccurate (but correct) MB learners. The proposed method is compared against two recent powerful constraint-based algorithms PCMB [3], IAMB [4] and Inter-IAMB [5]. We call our algorithm MBOR, it stands for "Markov Boundary search using the OR condition". MBOR was designed with a view to keep the conditional test sizes of the tests as small as possible.

The experiments on the synthetic databases focus on the accuracy and the data efficiency of MBOR, whereas the experiments on real data also addresses its scalability. The benchmarks used for the empiric test are: ASIA, INSURANCE, INSULINE, ALARM, HAILFINDER and CARPO. We report the average number of missing and extra variables in the output of MBOR with various sample sizes. The method is proved by extensive empirical simulations to be an excellent trade-off between time and quality of reconstruction. To show that MBOR is scalable, experiments are conducted on the THROMBIN database which contains 139,351 features [6].

The paper is organized as follows. In Section 2, we briefly discuss the lack of reliability of the conditional independence tests. In Section 3, we present and discuss our proposed algorithm called MBOR. Synthetic and real data sets from benchmarks are used in section 4 to evaluate MBOR against PCMB and InterIAMB.

2 Preliminaries

For the paper to be accessible to those outside the domain, we recall first the principle of Bayesian networks (BN), Markov boundaries and constraint-based learning BN methods. A BN is a tuple $< \mathcal{G}, P >$, where $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ is a directed acyclic graph (DAG) with nodes representing the random variables $\mathcal{V}$ and P a joint probability distribution on $\mathcal{V}$. In addition, $\mathcal{G}$ and P must satisfy the Markov condition: every variable, $X \in \mathcal{V}$, is independent of any subset of its non-descendant variables conditioned on the set of its parents. We denote the conditional independence of the variable X and Y given $\mathbf{Z}$, in some distribution P with $X \perp_P Y|\mathbf{Z}$. The independence constraints implied by the Markov condition necessarily hold in the joint distribution represented by any Bayesian network with structure $\mathcal{G}$. They can be identified by the d-separation criterion (See Pearl 1988). We use $X \perp_{\mathcal{G}} Y|\mathbf{Z}$ to denote the assertion that the DAG $\mathcal{G}$ imposes the constraint, via d-separation, that for all values z of the set Z, X is independent of Y given $Z = z$. We say that P is *faithful* with respect to $\mathcal{G}$ iff the d-separations in the DAG identify *all and only* the conditional independencies in P.

A Markov blanket $\mathbf{M}_T$ of the T is any set of variables such that T is conditionally independent of all the remaining variables given $\mathbf{M}_T$. A Markov boundary, $\mathbf{MB}_T$, of T is any Markov blanket such that none of its proper subsets is a

Markov blanket of T. Suppose $< \mathcal{G}, P >$, satisfies the faithfulness condition, then for all variable T, the set of parents of T, the children of T, and parents of children of T, is the **unique** Markov boundary of T. We denote by $\mathbf{PC}_T$, the set of parents and children of T in $\mathcal{G}$, and by $\mathbf{SP}_T$, the set of *spouses* of T in $\mathcal{G}$. The *spouses* of T are the parents of the children of T. We denote by $\mathbf{dSep}(X)$, the set that d-separates X from the (implicit) target T.

The identification of variable Markov boundary is a challenging topic of pattern recognition research. In recent years, there has been great interest in automatically inducing the Markov boundary from data using constraint-based (CB for short) learning procedures. The correctness, scalability and data efficiency of these methods have been proved and also illustrated by extensive experiments [3]. By correct (or sound), we mean that, under the assumptions that independence test are reliable and that the learning database is a sample from a distribution P faithful to a DAG $\mathcal{G}$, the algorithm returns the correct Markov boundary. The (ideal) assumption that the independence tests are reliable means that they decide (in)dependence iff the (in)dependence holds in P. Despite their great efficiency and scalability, these CB methods suffer from several drawbacks as we will see in the next section.

3 Problems with Constraint-Based MB Discovery

CB methods have the advantage of possessing clear stopping criteria and deterministic search procedures. On the other hand, they are prone to several instabilities: namely if a mistake is made early on in the search, it can lead to incorrect edges which may in turn lead to bad decisions in the future, which can lead to even more incorrect edges. This instability has the potential to cascade, creating many errors in the final graph [7]. We discuss next two well-known sources of test failure.

3.1 Conditional Independence Test Failures with Sparse Data

CB procedures systematically check the data for independence relationships to infer the structure. The association between two variables X and Y given a conditioning set $\mathbf{Z}$ is a measure of the strength of the dependence with respect to D. It is usually implemented with a statistical measure of association. Typically, the algorithms run a χ^2 independence test in order to decide on dependence or independence, that is, upon the acceptance or rejection of the null hypothesis of conditional independence. If the p-value of the test is smaller than a specific significance level, we reject the null hypothesis and consider that the two variables in the test are dependent. Insufficient data presents a lot of problems when working with statistical inference techniques like the independence test mentioned earlier. This occurs typically when the expected counts in the contingency table are small. The decision of accepting or rejecting the null hypothesis depends implicitly upon the degree of freedom which increases exponentially with the number of variables in the conditional set. So the larger the size of the

conditioning test, the less accurate are the estimates of conditional probabilities and hence the less reliable are the independence tests.

3.2 Almost Deterministic Relationships

Another difficulty arises when true- or almost-deterministic relationships (ADR) are observed among the variables. Loosely speaking, a relationship is said to be almost deterministic (and denoted by $\mathbf{X} \Rightarrow Y$) when the fraction of tuples that violate the deterministic dependency is at most equal to some threshold. True DR are source of unfaithfulness but the existence of ADR among variables doesn't invalidate the faithfulness assumption. The existence of ADR in the data may arise incidentally in smaller data samples. To remedy the problem, the variables that are almost-deterministically related to others may simply be excluded from the discovery process. However, if they are to be excluded, they first need to be identified before the DAG construction. This yields two problems. First, the identification is already exponentially complex. Second, a variable may have both deterministic and probabilistic relationships with other variables. On the other hand if we neither exclude deterministic variables nor handle appropriately the problem, then the unfaithful nature of deterministic nodes brings missing or extra edges to the acquired structure.

3.3 Practical Alternatives

Several proposals have been discussed in the literature in order to reduce the cascading effect of early errors that causes many errors to be present in the final graph. The general idea is to keep the size of the conditional sets as small as possible in the course of the learning process. For instance, Fast-IAMB [4] conducts significantly fewer conditional tests compared to IAMB [5] while MMMB [8] and PCMB [3] are more data efficient because $\mathbf{MB}_T$ can be identified by conditioning on sets much smaller than those used by IAMB. Another solution is to determine if enough data is available for the test to be deemed reliable. Following the approach in [9] Inter-IAMB considers a test to be reliable when the number of (complete) instances in D is at least five time the number of degrees of freedom *df* and skips it otherwise. This means that the number of instances required by the test is at least exponential in the size of the conditional set, because *df* is exponential in the size of the conditional set. In [10], they consider that 80% of the cells should have expected values greater than five. If the test is not considered as reliable, the variables are assumed to be independent without actually performing the test. This rule is rather arbitrary and errors may occur well before the lack of data is detected.

As a heuristic, the idea is generally to reduce the degree of freedom of the statistical conditional independence test by some ways. The aim is twofold: to improve the data efficiency and to allow an early detection of ADR. Various strategies exist to reduce the degrees of freedom [11]. In [12] for instance, if the reduced degree of freedom is small, then an ADR $\mathbf{Z} \Rightarrow X$ is suspected, a "safe choice" is taken : dependence is assumed $X \perp_P Y | \mathbf{Z}$ for all Y. Similarly, in

[13, 14], association rules miners are used to detect ADR and in [15, 16], the ADR are detected during the MB discovery. Once the ADR are detected, any CB algorithm can be used to construct a DAG such that, for every pair X and Y in $\mathbf{V}$, (X, Y) is connected in $\mathcal{G}$ if X and Y remains dependent conditionally on every set $\mathbf{S} \subseteq \mathbf{V} \setminus \{X, Y\}$ such that $\mathbf{S} \not\Rightarrow X$ and $\mathbf{S} \not\Rightarrow Y$.

4 New Method

In this section, we present in detail our learning algorithm called MBOR. We recall that MBOR was designed in order to endow the search procedure with the ability to: 1) handle efficiently data sets with thousands of variables but very few instances, 2) be correct under faithfulness condition, 3) handle implicitly some approximate deterministic relationships (ADR) without detecting them. We discuss next how we tackle each problem.

First of all, MBOR scales up to hundreds of thousands of variables in reasonable time because it searches the Markov boundary of the target without having to construct the whole Bayesian network first. Like PCMB [3] and MMMB [8], MBOR takes a divide-and-conquer approach that breaks the problem of identifying $\mathbf{MB}_T$ into two subproblems : first, identifying $\mathbf{PC}_T$ and, second, identifying the parents of the children (the spouses $\mathbf{SP}_T$) of T. According to Peña et al., this divide-and-conquer approach is supposed to be more data efficient than IAMB [5] and its variants, e.g., Fast-IAMB [10] and Interleaved-IAMB [4], because $\mathbf{MB}_T$ can be identified by conditioning on sets much smaller than those used by IAMB. Indeed, IAMB and its variants seek directly the minimal subset of $\mathbf{U}$ (the full set) that renders the rest of $\mathbf{U}$ independent of T, given $\mathbf{MB}_T$. Moreover, MBOR keeps the size of the conditional sets to the minimum possible without sacrificing the performance as discussed next.

The advantage of the divide-and-conquer strategy in terms of data efficiency does not come without some cost. MMMB [8] and PCMB [3] apply the "AND condition" to prove correctness under faithfulness condition. In other words, two variables X and Y are considered as neighbors if $Y \in PC_X$ AND $X \in PC_Y$. We believe this condition is far too severe and yields too many false negatives in the output. Instead, MBOR stands for "Markov Boundary search using the OR condition". This "OR condition" is a major difference between MBOR and all the above mentioned correct divide-and-conquer algorithms: two variables X and Y are considered as neighbors with MBOR if $Y \in PC_X$ OR $X \in PC_Y$. Clearly, the OR condition makes it easier for true positive nodes to enter the Markov boundary, hence the name and the practical efficiency of our algorithm. Moreover, the OR condition is a simple way to handle some ADR. For illustration, consider the sub-graph $X \Rightarrow T \rightarrow Y$, since $X \Rightarrow T$ is an ADR, $T \perp Y|X$ so Y will not be considered as a neighbor of T. As Y still sees T in its neighborhood, Y and T will be considered as adjacent. The main difficulty was to demonstrate the correctness under the faithfulness condition despite the OR condition. The proof is provided in the next section.

MBOR (Algorithm 1) works in three steps and it is based on four subroutines called *PCSuperset*, *SPSuperset* and *MBtoPC* (Algorithms 2-4). Before we describe the algorithm step by step, we recall that the general idea underlying MBOR is to use a weak MB learner to create a stronger MB learner. By weak learner, we mean a simple and fast method that may produce many mistakes due to its data inefficiency. In other words, the proposed method aims at producing an accurate MB discovery algorithm by combining fast and moderately inaccurate (but correct) MB learners. The weak MB learner is used in *MBtoPC* (Algorithm 4) to implement a correct Parents and Children learning procedure. It works in two steps. First, the weak MB learner called *CorrectMB* is used at line 1 to output a candidate MB. *CorrectMB* may be implemented by any algorithm of the IAMB family. In our implementation, we use Inter-IAMB for its simplicity and performance [5]. The key difference between IAMB and Inter-IAMB is that the shrinking phase is interleaved into the growing phase in Inter-IAMB. The second step (lines 3-6) of *MBtoPC* removes the spouses of the target.

In phase I, MBOR calls *PCSuperset* to extract **PCS**, a superset for the parents and children, and then calls *SPSuperset* to extract **SPS**, a superset for the target spouses (parents of children). Filtering reduces as much as possible the number of variables before proceeding to the MB discovery. In *PCSuperset* and *SPSuperset*, the size of the conditioning set $\mathbf{Z}$ in the tests is severely restricted: $card(\mathbf{Z}) \leq 1$ in *PCSuperset* (lines 3 and 10) and $card(\mathbf{Z}) \leq 2$ in *SPSuperset* (lines 5 and 11). As discussed before, conditioning on larger sets of variables would increase the risk of missing variables that are weakly associated to the target. It would also lessen the reliability of the independence tests. So the MB superset, **MBS** (line 3), is computed based on a scalable and highly data-efficient procedure. Moreover, the filtering phase is also a way to handle some ADR. For illustration, consider the sub-graph $Z \Rightarrow Y \rightarrow T \Leftarrow X$, since $X \Rightarrow T$ and $Z \Rightarrow Y$ are ADRs, $T \perp Y|X$ and $Y \perp T|Z$, Y would not be considered as a neighbor of T and vice-versa. The OR-condition in Phase II would not help in this particular case. Fortunately, as Phase I filters out variable Z, Y and T will be considered as adjacent

Phase II finds the parents and children in the restricted set of variables using the OR condition. Therefore, all variables that have T in their vicinity are included in $\mathbf{PC}_T$ (lines 7-8).

Phase III identifies the target's spouses in **MBS** in exactly the same way PCMB does [3]. Note however that the OR condition is not applied in this last phase because it would not be possible to prove its correctness anymore.

5 Proof of Correctness Under Faithfulness Condition

Several intermediate theorems are required before we demonstrate MBOR's correctness under faithfulness condition. Indeed, as **MBS** is a subset of $\mathbf{U}$, a difficulty arises: a marginal distribution $P^{\mathbf{V}}$ of $\mathbf{V} \subset \mathbf{U}$ may not satisfy the faithfulness condition with any DAG even if $P^{\mathbf{U}}$ does. This is an example of embedded faithfulness, which is defined as follow:

Algorithm 1. *MBOR*

Require: T: target; D: data set ($\mathbf{U}$ is the set of variables)
Ensure: [**PC**,**SP**]: Markov boundary of T

Phase I: *Find MB superset* (**MBS**)
$[\mathbf{PCS}, \mathbf{dSep}] = PCSuperSet(T, D)$
$\mathbf{SPS} = SPSuperSet(T, D, \mathbf{PCS}, \mathbf{dSep})$
$\mathbf{MBS} = \mathbf{PCS} \cup \mathbf{SPS}$
$\mathcal{D} = \mathcal{D}(\mathbf{MBS} \cup T)$ *i.e., remove from data set all variables in* $\mathbf{U}/\{\mathbf{MBS} \cup T\}$

Phase II: *Find parents and children of the target*
$\mathbf{PC} = MBtoPC(T, \mathcal{D})$
for all $X \in \mathbf{PCS} \setminus \mathbf{PC}$ **do**
  if $T \in MBtoPC(X, \mathcal{D})$ **then**
    $\mathbf{PC} = \mathbf{PC} \cup X$
  end if
end for

Phase III: *Find spouses of the target*
$\mathbf{SP} = \emptyset$
for all $X \in \mathbf{PC}$ **do**
  for all $Y \in MBtoPC(X, D) \setminus \{\mathbf{PC} \cup T\}$ **do**
    Find minimal $\mathbf{Z} \subset \mathbf{MBS} \backslash \{T \cup Y\}$ such that $T \perp Y | \mathbf{Z}$
    if $(T \not\perp Y | \mathbf{Z} \cup X)$ **then**
      $\mathbf{SP} = \mathbf{SP} \cup Y$
    end if
  end for
end for

Definition 1. *Let P be a distribution of the variables in $\boldsymbol{V}$ where $\boldsymbol{V} \subset \boldsymbol{U}$ and let $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$ be a DAG. $< \mathcal{G}, P >$ satisfies the embedded faithfulness condition if $\mathcal{G}$ entails all and only the conditional independencies in P, for subsets including only elements of $\boldsymbol{V}$.*

We obtain embedded faithfulness by taking the marginal of a faithful distribution as shown by the next theorem:

Theorem 1. *Let P be a joint probability of the variables in $\boldsymbol{U}$ with $\boldsymbol{V} \subseteq \boldsymbol{U}$ and $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$. If $< \mathcal{G}, P >$ satisfies the faithfulness condition and $P^{\boldsymbol{V}}$ is the marginal distribution of $\boldsymbol{V}$, then $< \mathcal{G}, P^{\boldsymbol{V}} >$ satisfies the* embedded *faithful condition.*

The proof can be found in [17]. Note that every distribution doesn't admit an embedded faithful representation. This property is useful to prove the correctness of our MBOR under the faithfulness condition. Let $\mathbf{PC}_X^{\mathbf{U}}$ denote the variables $Y \in \mathbf{U}$ such that there is no set $\mathbf{Z} \in \mathbf{U} \backslash \{X, Y\}$ such that $X \perp_P Y | \mathbf{Z}$. If $< \mathcal{G}, P >$ satisfies the faithfulness condition, $\mathbf{PC}_X^{\mathbf{U}}$ are the parents and children of X in $\mathbf{U}$. Otherwise, $\mathbf{PC}_X^{\mathbf{U}}$ is the unique set of the variables that remains dependent on X conditioned on any set $\mathbf{Z} \in \mathbf{U} \setminus \{X, Y\}$.

Algorithm 2. *PCSuperSet*

Require: T: target; D: data set (**U** is the set of variables)
Ensure: **PCS**: PC superset of T; **dSep**: d-separation set;

Phase I: *Remove X if $T \perp X$*
PCS $= \mathbf{U} \setminus T$
for all $X \in$ **PCS do**
  if $(T \perp X)$ **then**
    PCS $=$ **PCS** $\setminus X$
    dSep$(X) = \emptyset$
  end if
end for

Phase II:*Remove X if $T \perp X|Y$*
for all $X \in$ **PCS do**
  for all $Y \in$ **PCS** $\setminus X$ **do**
    if $(T \perp X \mid Y)$ **then**
      PCS $=$ **PCS** $\setminus X$
      dSep$(X) = Y$
    end if
  end for
end for

Algorithm 3. *SPSuperSet*

Require: T: target; D: data set (**U** is the set of variables); **PCS**: PC superset of T; **dSep**: d-separation set;
Ensure: **SPS**: SP superset of T;

SPS $= \emptyset$
for all $X \in$ **PCS do**
  SPS$_X = \emptyset$
  for all $Y \in \mathbf{U} \setminus \{T \cup$ **PCS**$\}$ **do**
    if $(T \not\perp Y|$**dSep**$(Y) \cup X)$ **then**
      SPS$_X =$ **SPS**$_X \cup Y$
    end if
  end for
  for all $Y \in$ **SPS**$_X$ **do**
    for all $Z \in$ **SPS**$_X \setminus Y$ **do**
      if $(T \perp Y|X \cup Z)$ **then**
        SPS$_X =$ **SPS**$_X \setminus Y$
      end if
    end for
  end for
  SPS $=$ **SPS** $\cup$ **SPS**$_X$
end for

Algorithm 4. *MBtoPC*

Require: T: target; D: data set
Ensure: **PC**: Parents and children of T;

1: $\mathbf{MB} = CorrectMB(T, D)$
2: $\mathbf{PC} = \mathbf{MB}$
3: **for all** $X \in \mathbf{MB}$ **do**
4: **if** $\exists \mathbf{Z} \subset (\mathbf{MB} \setminus X)$ such that $T \perp X \mid \mathbf{Z}$ **then**
5: $\mathbf{PC} = \mathbf{PC} \setminus X$
6: **end if**
7: **end for**

Theorem 2. *Let $\boldsymbol{U}$ be a set of random variables and $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$. If $< \mathcal{G}, P >$ satisfies the faithfulness condition, then every target T admits a unique Markov boundary $\boldsymbol{MB}_T^{\boldsymbol{U}}$. Moreover, for all $\boldsymbol{V}$ such that $\boldsymbol{MB}_T^{\boldsymbol{U}} \subseteq \boldsymbol{V} \subseteq \boldsymbol{U}$, T admits a unique Markov boundary over $\boldsymbol{V}$ and $\boldsymbol{MB}_T^{\boldsymbol{V}} = \boldsymbol{MB}_T^{\boldsymbol{U}}$.*

Proof: If $\mathbf{MB}_T^{\mathbf{U}}$ is the Markov boundary of T in $\mathbf{U}$, then T is independent of $\mathbf{V} \setminus \{\mathbf{MB}_T^{\mathbf{U}} \cup T\}$ conditionally on $\mathbf{MB}_T^{\mathbf{U}}$ so $\mathbf{MB}_T^{\mathbf{U}}$ is a Markov blanket in $\mathbf{V}$. Moreover, none of the proper subsets of $\mathbf{MB}_T^{\mathbf{U}}$ is a Markov blanket of T in $\mathbf{V}$, so $\mathbf{MB}_T^{\mathbf{U}}$ is also a Markov boundary of T in $\mathbf{V}$. So if it is not the unique MB for T in $\mathbf{V}$ there exists some other set $\mathbf{S}_T$ not equal to $\mathbf{MB}_T^{\mathbf{U}}$, which is a MB of T in $\mathbf{V}$. Since $\mathbf{MB}_T^{\mathbf{U}} \neq \mathbf{S}_T$ and $\mathbf{MB}_T^{\mathbf{U}}$ cannot be a subset of $\mathbf{S}_T$, there is some $X \in \mathbf{MB}_T^{\mathbf{U}}$ such that $X \notin \mathbf{S}_T$. Since $\mathbf{S}_T$ is a MB for T, we would have $T \perp_P X | \mathbf{S}_T$. If X is a parent or child of T, we would not have $T \perp_{\mathcal{G}} X | \mathbf{S}_T$ which means we would have a conditional independence which is not entailed by d-separation in $\mathcal{G}$ which contradicts the faithfulness condition. If X is a parent of a child of T in $\mathcal{G}$, let Y be their common child in $\mathbf{U}$. If $Y \in \mathbf{S}_T$ we again would not have $T \perp_{\mathcal{G}} X | \mathbf{S}_T$. If $Y \notin \mathbf{S}_T$ we would have $T \perp_P Y | \mathbf{S}_T$ because $\mathbf{S}_T$ is a MB of T in $\mathbf{V}$ but we do not have $T \perp_{\mathcal{G}} Y | \mathbf{S}_T$ because T is a parent of Y in G. So again we would have a conditional independence which is not a d-separation in $\mathcal{G}$. This proves that there can not be such set $\mathbf{S}_T$. □

Theorem 3. *Let $\boldsymbol{U}$ be a set of random variables and T a target variable. Let $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$ be a DAG such that $< \mathcal{G}, P >$ satisfies the faithfulness condition. Let $\boldsymbol{V}$ be such that $\boldsymbol{MB}_T^{\boldsymbol{U}} \subseteq \boldsymbol{V} \subseteq \boldsymbol{U}$ then, $\boldsymbol{PC}_T^{\boldsymbol{V}} = \boldsymbol{PC}_T^{\boldsymbol{U}}$.*

Proof: Clearly $\mathbf{PC}_T^{\mathbf{U}} \subseteq \mathbf{PC}_T^{\mathbf{V}}$ as $\mathbf{MB}_T^{\mathbf{U}} \subseteq \mathbf{V} \subseteq \mathbf{U}$. If $X \in \mathbf{PC}_T^{\mathbf{V}}$ and $X \notin \mathbf{PC}_T^{\mathbf{U}}$, $\exists \mathbf{Z} \subset \mathbf{MB}_T^{\mathbf{U}} \setminus X$ such that $T \perp_P X | \mathbf{Z}$ because all non adjacent nodes may be d-separated in $\mathcal{G}$ by a subset of its Markov boundary. As $\mathbf{MB}_T^{\mathbf{U}} = \mathbf{MB}_T^{\mathbf{V}}$ owing to Theorem 2, so X and T can be d-separated in $\mathbf{V} \setminus \{X, T\}$. Therefore, X cannot be adjacent to T in $\mathbf{V}$. □

Theorem 4. *Let $\boldsymbol{U}$ be a set of random variables and T a target variable. Let $\mathcal{G} =< \mathbf{U}, \mathbf{E} >$ be a DAG such that $< \mathcal{G}, P >$ satisfies the faithfulness condition. Let $\boldsymbol{V}$ be such that $\boldsymbol{MB}_T^{\boldsymbol{U}} \subseteq \boldsymbol{V} \subseteq \boldsymbol{U}$. Under the assumption that the independence*

tests are reliable, MBtoPC(T, $\boldsymbol{V}$*) returns* $\boldsymbol{PC}_T^{\boldsymbol{U}}$*. Moreover, let* $X \in \boldsymbol{V} \setminus T$*, then* T *is in the output of MBtoPC*$(X, \boldsymbol{V}, \mathcal{D})$ *iff* $X \in \boldsymbol{PC}_T^{\boldsymbol{U}}$*.*

Proof: We prove first that MBtoPC(T,$\mathbf{V}$) returns $\mathbf{PC}_T^{\mathbf{U}}$. In the first stage of MBtoPC, CorrectMB(T,$\mathbf{V}$) seeks a minimal set $\mathbf{S}_T \in \mathbf{V} \setminus T$ that renders $\mathbf{V} \setminus \mathbf{S}_T$ independent of T conditionally on $\mathbf{S}_T$. This set is unique owing to Theorem 2, therefore $\mathbf{S}_T = \mathbf{MB}_T^{\mathbf{V}} = \mathbf{MB}_T^{\mathbf{U}}$. In the backward phase, MBtoPC removes the variables $X \in \mathbf{MB}_T^{\mathbf{V}}$ such that $\exists \mathbf{Z} \subset (\mathbf{MB}_T^{\mathbf{V}} \setminus X)$ for which $T \perp X \mid \mathbf{Z}$. These variables are the spouses of T in $\mathcal{G}$, so MBtoPC(T,$\mathbf{V}$) returns $\mathbf{PC}_T^{\mathbf{U}}$. Now, if $X \notin \mathbf{PC}_T^{\mathbf{U}}$ then $X \notin \mathbf{PC}_T^{\mathbf{V}}$ owing to Theorem 3. So there is a set $\mathbf{Z} \subset \mathbf{V} \setminus \{X, Y\}$ such that $T \perp X \mid \mathbf{Z}$. Therefore, X cannot be in the output of MBtoPC(T,$\mathbf{V}$). □

Theorem 5. *Under the assumptions that the independence tests are reliable and that the database is a sample from a probability distribution* P *faithful to a DAG* $\mathcal{G}$*, MBOR*(T) *returns* $\boldsymbol{MB}_T^{\boldsymbol{U}}$*.*

Proof: Let **MBS** be the MB superset constructed at line 3 of *MBOR*. It is straightforward to show that $\mathbf{MB}_T^{\mathbf{U}} \subset \mathbf{MBS}$. So the Markov boundary of T in **MBS** is that of **U** owing to Theorem 2 so the problem is well defined. In Phase II at line 7, if T is in the output of MBtoPC$(X, \mathbf{V}, \mathcal{D})$ then X should be in the output of MBtoPC$(T, \mathbf{V}, \mathcal{D})$ owing to Theorem 4. So phase II ends up with the $\mathbf{PC}_T^{\mathbf{U}}$. In Phase III, lines 11-19 identify all and only the spouse of T in $\mathcal{G}$ when the faithfulness condition is asssumed as shown in [3]. When the assumption doesn't hold anymore for $< \mathcal{G}, P^{`}\mathbf{V} >$, we need to show that a fake spouse will not enter the set **SP**. In phase III line 12, it is easy to see that MBtoPC$(X, \mathbf{V}, \mathcal{D})$ returns a set $\mathbf{PC}_X^{\mathbf{V}}$ that may differ from $\mathbf{PC}_X^{\mathbf{U}}$. Suppose $Y \notin \mathbf{PC}_X^{\mathbf{U}}$ and Y is in the output of MBtoPC$(X, \mathbf{V}, \mathcal{D})$. This means that there exists at least one active path between X and Y in $\mathcal{G}$ that contains a node in $\mathbf{U} \setminus \mathbf{V}$. At lines 13-14, Y is considered as spouse of T if there is a set $\mathbf{Z} \subset \mathbf{MBS} \setminus \{T \cup Y\}$ so that $T \perp Y | \mathbf{Z}$ and $T \not\perp Y | \mathbf{Z} \cup X$. Therefore, this path in $\mathcal{G}$ should necessarily be of the type $T \rightarrow X \leftarrow A \leftrightsquigarrow B \leftrightsquigarrow Y$ where $\leftrightsquigarrow$ denotes an active path otherwise we would not have $T \not\perp Y | \mathbf{Z} \cup X$. As A is a spouse of T, $A \in \mathbf{MB}_T^{\mathbf{U}}$ and so A is in $\mathbf{V}$. Suppose B is not in $\mathbf{V}$, then A still d-separates X and Y so Y cannot be in the output of MBtoPC$(X, \mathbf{V}, \mathcal{D})$ since we found a set $\mathbf{Z} \subseteq \mathbf{V}$ such that $X \not\perp_P Y | \mathbf{Z}$. So Y is included in **S**P at line 16 iff Y is a spouse of T in **U**. □.

6 Experimental Validation

In this section, we compare the performance of InterIAMB, PCMB and MBOR through experiments on synthetic and real databases with very few instances compared to the number of variables. They are written in MATLAB and all the experiments are run on a Intel Core 2 Duo T77500 with 2Gb RAM running Windows Vista. To implement the conditional independence test, we calculate the G^2 statistic as in [11], under the null hypothesis of the conditional independence. The significance level of the test in all compared algorithms is 0.05 except on the high-dimensional THROMBIN data where it is 0.0001. All three

algorithms are correct under the faithfulness condition and are also scalable. We do not consider MMMB and HITON-MB because we are not interested in any algorithm that does not guarantee the correctness under faithfulness assumption. We do not consider GS because IAMB outperforms it [5]. Even if PCMB was also shown experimentally in [3] to be more accurate than IAMB and its variants, we consider InterIAMB because it is used as a subroutine in MBOR.

It might very well happen that several variables have the same association value with the target in data sets with very few instances. In this particular case, somewhat arbitrary (in)dependence decisions are taken. This can be seen as a source of randomness inherent to all CB procedures. To handle this problem, our implementation breaks ties at random: a random permutation of the variables is carried out before MBOR is run. This explains the variability of MBOR with very few instances and/or extremely large number of variables (e.g., THROMBIN).

6.1 Synthetic Data

Figure 1 illustrates the results of our experiments on six common BN benchmarks : BREAST-CANCER or ASIA (8 nodes/8 arcs), INSURANCE (27/52), INSULINE (35/52), ALARM (37/46), HAILFINDER (56/66) and CARPO (61/74). These benchmarks are available from the UCI Machine Learning Repository. All three algorithms have been run on each variable for all data sets. Figure 1 (upper part) summarizes graphically the results in terms of missing and extra nodes in the output of the MB averaged over 10 runs for 200, 500 and 1000 i.i.d. samples. The upper part shows the average false positive and and lower part shows the false negative rates. The overall accuracy is very similar for nodes with Markov boundaries with less than 4 variables. For larger MBs, however, the advantages of MBOR against the other two algorithms are far more noticeable. For instance, MBOR consistently outperforms the other algorithms on variable *IPA* in the INSULINE benchmark as may be seen in Table 2. Figure 2 (lower part) show the performance for nodes with more than 4 variables. Results are averaged over all the above mentioned benchmarks. As observed, MBOR reduces drastically the average number of false negatives compared to PCMB and InterIAMB (up to 40% on INSULINE). This benefit comes at very little expense: the false positive rate is slightly higher. This is not a surprise as PCMB makes it harder for true positives to enter the output.

Table 1. INSULINE benchmark : number of extra and missing variables for PCMB, Inter-IAMB and MBOR for variable *IPA* run on 1000 instances. Results are averaged over 10 runs.

Algorithm	false positive	false negative
PCMB	0.4	11.8
InterIAMB	0	12.6
MBOR	2.1	2.1

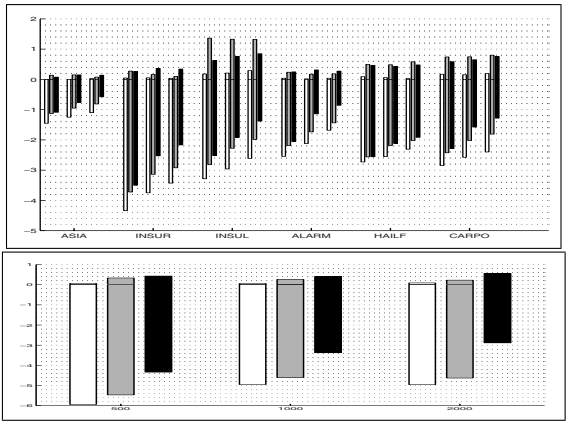

Fig. 1. Upper plot : average missing (lower part) and extra (upper part) variables for learning Markov boundaries of all variables of ASIA, INSURANCE, INSULINE, ALARM, HAILFINDER and CARPO networks. The results of PCMB, InterIAMB and MBOR are shaded in white, gray and black respectively. For each benchmark the bars show the results on 200, 500 and 1000 instances respectively. All results are averaged over 10 runs. Lower plot : results are averaged over all benchmarks for nodes that have a MB with more than 4 variables in the MB, for 500, 1000 and 2000 i.i.d. samples.

6.2 Real Data

In this section, we assess the performance of the probabilistic classification using the feature subset output by MBOR. To this purpose, we consider several categorical data bases from the UCI Machine Learning Repository in order to evaluate the accuracy of MBOR against InterIAMB and PCMB. The database description and the results of the experiments with the Car Evaluation, Chess, Molecular Biology, SPECT heart, Tic-Tac-Toe, Wine and Waveform are shown in Table 1. Performance is assessed by hit rate (correct classification rate), relative absolute error (R.A.E.), and Kappa Statistics obtained by 10-fold cross-validation. Kappa can be thought of as the chance-corrected proportional agreement, and possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement). As may be seen, the classification performance by naive Bayes classifier on the features selected by MBOR has always outperformed that of InterIAMB and PCMB by a noticeable margin, especially on the Molecular Biology database.

6.3 Real Data : Thrombin Database

Our last experiments demonstrate the ability of MBOR to solve a real world FSS problem involving thousands of features. We consider the THROMBIN database which was provided by DuPont Pharmaceuticals for KDD Cup 2001. It is exemplary of a real drug design [6]. The training set contains 1909 instances characterized by 139,351 binary features. The accuracy of a Naive Bayesian classifier

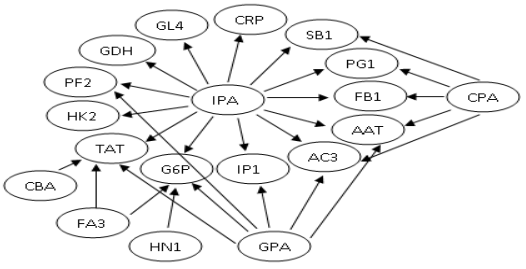

Fig. 2. INSULINE benchmark : Markov boundary of the variable *IPA*

Table 2. Feature selection performance using several UCI datasets in terms of classification performance using 10-fold cross-validation by naive Bayes classifier on the features selected by InterIAMB, PCMB and MBOR. By "hit rate", we mean correct classification rate.

Data Sets	*Inst.*	*Attr.*	*Accuracy*	*Algorithms*		
				InterIAMB	PCMB	MBOR
Car Evaluation	1728	6	Hit Rate	79.11%	79.11%	85.36%
			Kappa	0.5204	0.5204	0.6665
			Rel. abs. error.	56.59%	56.59%	49.88%
Chess (King-Rook vs. King-Pawn)	3196	36	Hit Rate	94.37%	92.43%	93.15%
			Kappa	0.8871	0.8479	0.8624
			Rel. abs. error	45.09%	47.23%	44.51%
Molecular Biology (Splice-junction Gene Sequences)	3190	61	Hit Rate	74.01%	74.01%	95.61%
			Kappa	0.5941	0.5941	0.9290
			Rel. abs. error	56.94%	56.94%	10.27%
SPECT Heart	267	22	Hit Rate	76.40%	79.40%	84.27%
			Kappa	0.2738	0	0.4989
			Rel. abs. error	77.83%	93.92%	71.71%
Tic-Tac-Toe Endgame	958	9	Hit Rate	67.95%	67.95%	72.44%
			Kappa	0.1925	0.1951	0.3183
			Rel. abs. error	85.62%	85.75%	82.53%
Wine	178	13	Hit Rate	94.38%	94.38%	98.88%
			Kappa	0.9148	0.9148	0.9830
			Rel. abs. error	19.08%	19.08%	2.67%
Waveform - Version 1	5000	21	Hit Rate	76.42%	76.42%	81.32%
			Kappa	0.6462	0.6462	0.7196
			Rel. abs. error	44.50%	44.50%	29.43%

was computed as the average of the accuracy on true binding compounds and the accuracy on true non-binding compounds on the 634 compounds of the test set. As the data is unbalanced, the accuracy is calculated as the average of true positive rate and the true negative rate. A significance level of 0.0001 avoids better than 0.01 the spurious dependencies that may exist in the data due to

the large number of features. MBS with $\alpha = 0.0001$ returns a set of 21 variables, SPS select 39 variables and MBOR outputs 5 variables on average in about 3h running time.

Note shown here, the 10 runs return each time a different MB, all of them containing 5 features. They mostly differ by one or two variables. MBOR scores between 36% (really bad) to 66% with an average 53% on average which seems really deceiving compared to PCMB and IAMB that achieves respectively 63% and 54% as shown in [3]. Nonetheless, MBOR is highly variable and was able to identify 3 different MBs that outperform those found by IAMB and 90% of those by PCMB. For instance, the MB which scores 66% contains the two variables 20973, 63855. These two variables, when used conjunctly, score 66,9% which is impressive according to [6, 3] for such a small feature set. Note that a MB with the four features obtained by the winner of KDD cup 2001 scores 67% accuracy.

The execution time was not reported as it is too dependent on the specific implementation. We were unable to run PCMB on the Thrombin database in reasonable time with *our* MATLAB implementation. On synthetic data, MBOR runs (say) 30% faster than PCMB.

7 Discussion and Conclusion

We discussed simple solutions to improve the data efficiency of current constraint-based Markov boundary discovery algorithms. We proposed a novel approach called MBOR that combines the main advantages of PCMB and IAMB while still being correct under faithfulness condition. Our experimental results show a clear benefit in several situations: densely connected DAGs, weak associations or approximate functional dependencies among the variables. Though not discussed here, a topic of considerable interest would be to ascertain the data distributions for which MBOR, PCMB or the stochastic variant of IAMB termed KIAMB proposed in [3], is most suited. This needs further substantiation through more experiments and analysis.

References

[1] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)

[2] Nilsson, R., Peña, J., Bjrkegren, J., Tegnr, J.: Consistent feature selection for pattern recognition in polynomial time. Journal of Machine Learning Research 8, 589–612 (2007)

[3] Peña, J., Nilsson, R., Bjrkegren, J., Tegnr, J.: Towards scalable and data efficient learning of markov boundaries. International Journal of Approximate Reasoning 45(2), 211–232 (2007)

[4] Yaramakala, S., Margaritis, D.: Speculative markov blanket discovery for optimal feature selection. In: ICDM, pp. 809–812 (2005)

[5] Tsamardinos, I., Aliferis, C.F., Statnikov, A.R.: Algorithms for large scale markov blanket discovery. In: FLAIRS Conference, pp. 376–381 (2003)

[6] Cheng, J., Hatzis, C., Hayashi, H., Krogel, M., Morishita, S., Page, D., Sese, J.: KDD Cup 2001 Report. In: ACM SIGKDD Explorations, pp. 1–18 (2002)
[7] Dash, D., Druzdzel, M.J.: Robust independence testing for constraint-based learning of causal structure. In: UAI, pp. 167–174 (2003)
[8] Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing bayesian network structure learning algorithm. Machine Learning 65(1), 31–78 (2006)
[9] Spirtes, P., Glymour, C., Scheines, R.: Causation, prediction, and search. Springer, Heidelberg (1993)
[10] Yaramakala, S.: Fast markov blanket discovery. In MS-Thesis. Iowa State University (2004)
[11] Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd edn. MIT Press, Cambridge (2000)
[12] Yilmaz, Y.K., Alpaydin, E., Akin, H.L., Bilgiç, T.: Handling of deterministic relationships in constraint-based causal discovery. In: Probabilistic Graphical Models (2002)
[13] Luo, W.: Learning bayesian networks in semi-deterministic systems. In: Canadian Conference on AI, pp. 230–241 (2006)
[14] Kebaili, Z., Aussem, A.: A novel bayesian network structure learning algorithm based on minimal correlated itemset mining techniques. In: IEEE Int. Conference on Digital Information Management ICDIM 2007, pp. 121–126 (2007)
[15] Aussem, A., de Morais, S.R., Corbex, M.: Nasopharyngeal carcinoma data analysis with a novel bayesian network skeleton learning. In: Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) AIME 2007. LNCS (LNAI), vol. 4594, pp. 326–330. Springer, Heidelberg (2007)
[16] Rodrigues de Morais, S., Aussem, A., Corbex, M.: Handling almost-deterministic relationships in constraint-based bayesian network discovery: Application to cancer risk factor identification. In: 16th European Symposium on Artificial Neural Networks ESANN 2008, pp. 101–106 (2008)
[17] Neapolitan, R.E.: Learning Bayesian Networks. Prentice-Hall, Englewood Cliffs (2004)