Fast Markov blanket discovery

by

Sandeep Yaramakala

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Computer Science

Program of Study Committee:
Dimitris Margaritis, Major Professor
Vasant Honavar
Tapabrata Maiti

Iowa State University

Ames, Iowa

2004

Graduate College
Iowa State University

This is to certify that the Master's thesis of

Sandeep Yaramakala

has met the thesis requirements of Iowa State University

Major Professor

For the Major Program

DEDICATION

To my parents Smt. Chagari Sasikala and Sri Yaramakala Raghuram Reddy and my sister Divya without whose support I would not have been able to complete this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this thesis, we address the problem of learning the Markov blanket of a quantity from data in an efficient manner. The discovery of the Markov blanket is useful in the feature subset selection problem and in discovering the structure of a Bayesian network. Empirical results show that the algorithm performs reasonably faster than existing algorithms. In addition, the results also show that the speedup does not adversely affect the accuracy of the recovered Markov blankets.

CHAPTER 1. Introduction

We live in a world that is deluged with data and information. The great progress made in the fields of science and technology has rapidly increased our capability to generate and collect data in the last several decades. Some of the chief sources of data are a direct result of the computerization of almost every walk of life: online banking, e-commerce, RFID tagging of many types of commodities and the ever expanding Internet are a rich source of data and information, to name a few. Other sources that generate data include scientific experiments in various fields that have tremendously benefited from the advancement of science and scientific tools to study and analyze several different types of things that were previously difficult or hard to understand. Our insatiable thirst for knowledge is a driving force that creates a need to study the enormous amount of data generated by these processes and extract information and knowledge from it. In this thesis we are interested in looking at data domains that are well-defined in that we precisely know the set of attributes the domain is defined over and the kinds of values these attributes can assume.

When examining a specific domain of interest, it is often the case that a particular attribute in this vast pool of data intrigues us and makes us keen on studying it. To study and analyze this attribute, we usually try to ascertain how the influence of other attributes in the domain affect it. For example, an engineer at an automobile manufacturing plant might be interested in studying the amount of hydrocarbon emission of the automobile being built. The engineer might try to determine the factors affecting the emission of hydrocarbons and then try to find an optimal set of values for these factors so as to reduce the amount of emission. As another example, a biologist studying an organism might be interested in determining which attributes help in accurately predicting the structure and function of genes in that organism. In all these

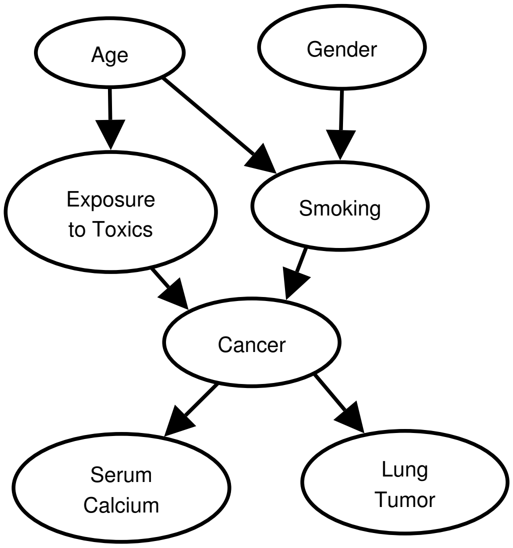

Figure 1.1 A Bayesian network for lung cancer.

tasks the goal is to study how the attribute under consideration "behaves" under the effect of other attributes in the domain. However, the domain might be defined over a very large number of attributes, making this task non-trivial, and, in some cases infeasible. Moreover, interpretation of the resulting analysis becomes exceedingly difficult when there are a number of attributes to consider. A solution to this problem is to determine a set of attributes that can shield the attribute of interest from the effect of other attributes in the domain. This set of attributes is called a *Markov blanket* and the task of finding such a set is the focus of this thesis.

In this thesis we assume that there exists a faithful Bayesian network that models the domain. Bayesian networks are graphical models for succinct representation of joint probability distributions. Normally, a Bayesian network only models the independencies in the data set. A faithful Bayesian network on the other hand also models the dependencies in the data. Using a faithful Bayesian network, the Markov blanket of any attribute in the domain 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. The following example illustrates how one can determine the Markov blanket of an attribute given a Bayesian network.

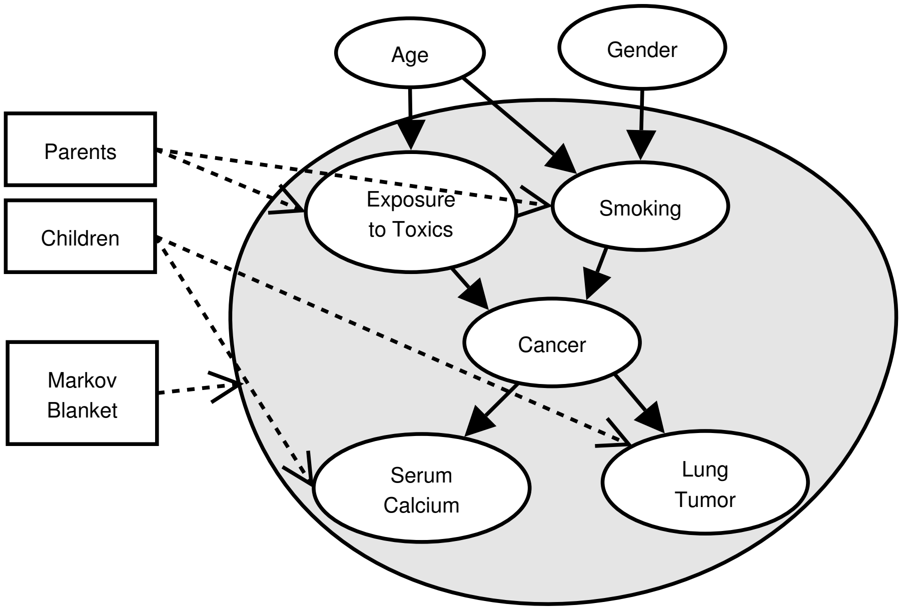

Figure 1.2 The Markov blanket of *Cancer*.

Figure 1.1 shows a Bayesian network consisting of five attributes of a person: *Age*, *Gender*, *Exposure to Toxics*, *Smoking* (denoting whether the person has a past history of smoking), *Cancer* (denoting whether or not the person has lung cancer), *Serum Calcium* (denoting the level of serum calcium) and *Lung Tumor* (denoting the presence or absence of a lung tumor).

If *Cancer* is the attribute of interest, then one might wish to determine the value of this attribute given some assignment of values to the other attribute in the domain. For example, one might know that a particular person is 30 years old, is male, has not had heavy exposure to toxic substances, is a smoker, has high levels of serum calcium and a lung tumor. One can readily estimate the most probable value of the attribute *Cancer* given this information about the person. However, the Markov blanket of this attribute renders some of the available information irrelevant.

The set of parents of *Cancer* is {*Exposure to Toxics*, *Smoking*} and the set of children is {*Serum Calcium*, *Lung Tumor*}. There are no spouses. Therefore, the Markov blanket of the variable *Cancer* is the set {*Exposure to Toxics*, *Smoking*, *Serum Calcium*, *Lung Tumor*}. This blanket is depicted in Figure 1.2. For the attribute *Cancer*, knowledge about the age and gender of a person become irrelevant if we know {*Exposure to Toxics, Smoking, Serum*

Calcium, Lung Tumor}, because the blanket shields *Cancer* from the effects of those attributes outside it.

The general idea is that, to study a particular attribute one only needs to consider those attributes in its blanket. The goal of this thesis is to develop a fast algorithm for discovering Markov blankets. However, we do not address Bayesian network structure discovery here — Markov blankets are discovered without determining the structure of the underlying Bayesian network.

The remainder of this thesis is organized as follows: In **Chapter 2** we review some preliminaries and give some Bayesian network formalisms that allow us to connect the notion of a Markov blanket to a Bayesian network. In **Chapter 4** we present the necessary statistical background on testing for independence of variables and describe the issues that one might encounter in the presence of insufficient data. In **Chapter 3** we review past work related to Markov blanket discovery. In particular, we review a number of sound blanket discovery algorithms and weight their advantages and disadvantages. In **Chapter 5** we present a novel algorithm called FAST-IAMB for discovering Markov blankets and also give empirical results comparing this algorithm with earlier work. We conclude with some thoughts on possible future research directions.

CHAPTER 2. Preliminaries

In this chapter we formally define the notion of a Markov blanket and see how it relates to a Bayesian network. The organization of this chapter is as follows: In Section 2.1 we introduce the notations used throughout this thesis. We formally define the notion of a Markov blanket in Section 2.2. In Section 2.3, we present formal definitions of some Bayesian network concepts and describe how the concept of a Markov blanket relates to a Bayesian network.

2.1 Notations

Table 2.1 lists the symbols used in this thesis. We use italicized capital letters at the end of the alphabet to denote variables and bold-face capitals to denote sets of variables. In this thesis we only deal with categorical data *i.e.* data in which all variables are discrete valued. However, the ideas presented here readily extend to continuous or hybrid domains. We use the phrase "a configuration of $\mathbf{Z}$" to mean one possible instantiation of the variables in the set $\mathbf{Z}$.

The notation $X \perp Y$ denotes that X and Y are unconditionally independent of each other. Similarly, $X \not\perp Y$ denotes unconditional dependence. The notation $X \perp Y \mid \mathbf{Z}$ denotes that X and Y are conditionally independent given Z. Likewise, $X \not\perp Y \mid \mathbf{Z}$ denotes conditional dependence: this means that there exists at least one configuration of $\mathbf{Z}$ in which X and Y are unconditionally dependent. For a probabilistic definition of conditional independence, see Definition 1; for a statistical definition, refer to Chapter 4.

The words "variable", "attribute" and "feature" are all used interchangeably throughout this thesis.

Table of symbols	
$\mathcal{D}$	Main data set
N	Number of points in the data set *i.e.* $\lvert\mathcal{D}\rvert$
$X, Y, Z, \ldots$	One-dimensional variables
$x, y, z, \ldots$	Values of corresponding variables $X, Y, Z, \ldots$
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \ldots$	Sets of variables
$\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots$	Values of sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \ldots$
$\mathcal{U}$	Universe: set of all variables in the domain
n	Number of variables in $\mathcal{U}$
T	Target: variable for which we find the Markov blanket
C	Class variable in a supervised-learning task
$\mathbf{B}(T)$	Markov blanket of T
r_X	Number of values taken on by variable X
$r_{\mathbf{X}}$	Number of configurations of the set $\mathbf{X}$
h	Heuristic function used in Markov blanket discovery algorithms
$D(p(x) \parallel q(x))$	The KL divergence between pmf's $p(X)$ and $q(X)$

Table 2.1 Symbols used in this thesis.

2.2 The Notion of a Markov Blanket

Definition 1 (Conditional Independence) *Let* $\Pr(\cdot)$ *be a joint probability function over the variables in* $\mathcal{U}$*, and let* $\mathbf{X}$*,* $\mathbf{Y}$*, and* $\mathbf{Z}$ *stand for any three subsets of variables in* $\mathcal{U}$*.* $\mathbf{X}$ *and* $\mathbf{Y}$ *are said to be conditionally independent given* $\mathbf{Z}$ *if and only if*

$$\Pr(\mathbf{X} = \mathbf{x} \mid \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) = \Pr(\mathbf{X} = \mathbf{x} \mid \mathbf{Z} = \mathbf{z}) \text{ whenever } \Pr(\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) > 0. \tag{2.1}$$

Similarly, we can define unconditional independence as follows: $\mathbf{X}$ *and* $\mathbf{Y}$ *are said to be* ***unconditionally independent*** *iff*

$$\Pr(\mathbf{X} = \mathbf{x} \mid \mathbf{Y} = \mathbf{y}) = \Pr(\mathbf{Y} = \mathbf{y}) \text{ whenever } \Pr(\mathbf{Y} = \mathbf{y}) > 0.$$

Intuitively, if $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ then $\mathbf{Y}$ gives us no information about $\mathbf{X}$ beyond what is already in $\mathbf{Z}$.

Definition 2 (Markov Blanket) *A Markov blanket* $\mathbf{B}(T)$ *of an element* $T \in \mathcal{U}$ *is any subset* $\mathbf{S}$ *of elements for which*

$$T \perp \mathcal{U} - \mathbf{S} - \{T\} \mid \mathbf{S} \text{ and } T \notin \mathbf{S}. \tag{2.2}$$

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 2.2.*

The concept of the Markov blanket of a variable is central to the algorithms presented in the paper. Intuitively, the Markov blanket of a variable shields the variable from the influence of the other variables in the set. Therefore, a variable's value can be probabilistically determined from the values of the variables in its blanket alone — the value assignments of the variables outside the blanket become irrelevant. In this thesis, we identify the Markov blanket of a variable with its Markov boundary and use the notation $\mathbf{B}(T)$ to denote the Markov blanket of variable T.

2.3 Bayesian Network Formalisms

A Bayesian network is a graphical model for efficiently representing a joint probability distribution defined over a set of variables. It is denoted by $\langle D, P \rangle$ where

1. D is a directed acyclic graph defined over a set of variables $\mathcal{U}$; the graph encodes independence relationships among the variables in $\mathcal{U}$.

2. P denotes a set of local probability distributions, one for each variable conditioned on its parents

For any three variables A, B and C in $\mathcal{U}$, if there is a directed edge $A \rightarrow B$ in a Bayesian network defined over $\mathcal{U}$, then A is called a **parent** of B and B is called a **child** of A. If there exist two edges such that $A \rightarrow C \leftarrow B$ (*i.e.* C is a common child of A and B), then A is a **spouse** of B (and vice versa).

An example network is shown in Figure 2.1 [Friedman and Goldszmidt, 1998]. The network is defined over a domain that has five variables, all of them binary valued: *Earthquake* (E) that represents the event that an earthquake has occurred, *Burglary* (B) that represents the event that a burglary has occurred, *Alarm* (A) that represents the event that the alarm has gone off, *Radio* (R) that represents a radio announcement and *Call* (C) that represents the event that a neighbor has called. Table 2.2 shows one of the many conditional probability distributions that form P.

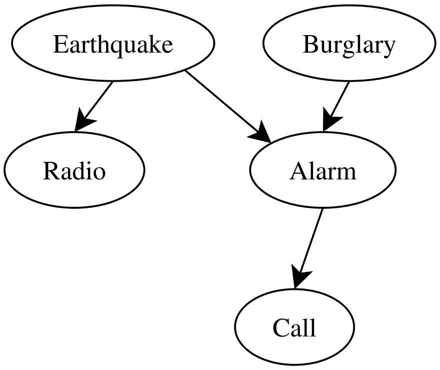

Figure 2.1 A Bayesian network with five nodes.

E	B	$\Pr(A = \text{Yes} \mid E, B)$
Yes	Yes	0.9
Yes	No	0.2
No	Yes	0.9
No	No	0.01

Table 2.2 An example local probability distribution for the variable *Alarm*.

Given the Bayesian network of Figure 2.1, we can make some statements about the independencies that are implied by it. For example, we can say that the occurrence of an *Earthquake* does not depend on whether a *Burglary* has taken place or not when nothing else is known about the values of the other variables. Another statement implied by the network is that, given that there is an *Earthquake*, a *Radio* announcement can be predicted regardless of whether the *Alarm* sounds. In fact, all independencies represented by a Bayesian network can be "read" from its DAG structure by using the rules of d-separation, the definition of which we give below.

The following definitions are from [Pearl, 1988]. They formalize a number of Bayesian network concepts.

Definition 3 **(D-separation)** *If* $\mathbf{X}$, $\mathbf{Y}$, *and* $\mathbf{Z}$ *are three disjoint subsets of nodes in a DAG* D, *then* $\mathbf{Z}$ *is said to d-separate* $\mathbf{X}$ *from* $\mathbf{Y}$, *denoted* $\mathbf{X} \perp_D \mathbf{Y} \mid \mathbf{Z}$, *if there is no path between a* *node in* $\mathbf{X}$ *and a node in* $\mathbf{Y}$ *along which the following two conditions hold:*

1. *every node with converging arrows is in* $\mathbf{Z}$ *or has a descendant in* $\mathbf{Z}$ *and*

2. *every other node is outside* $\mathbf{Z}$*.*

Informally, two sets of variables $\mathbf{X}$ and $\mathbf{Y}$ are said to be d-separated by a third set $\mathbf{Z}$ if all undirected paths between any two variables in $\mathbf{X}$ and $\mathbf{Y}$ are **blocked**. An undirected path is said to be blocked if for some set of variables A, B and C in the path, at least one of the following three holds:

1. A, B, C are connected as shown in Figure 2.2(a) and C is in $\mathbf{Z}$.

2. A, B, C are connected as shown in Figure 2.2(b) and C is in $\mathbf{Z}$.

3. A, B, C are connected as shown in Figure 2.2(c) and neither C nor any of its descendants are in $\mathbf{Z}$.

If $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ then $\mathbf{X}$ and $\mathbf{Y}$ are said to be condintionally independent given $\mathbf{Z}$ in the DAG D. On the other hand, if $\mathbf{X}$ and $\mathbf{Y}$ are not d-separated then $\mathbf{X}$ and $\mathbf{Y}$ are said to be conditionally dependent in D given $\mathbf{Z}$.

The following is a list of some of the independencies that can be inferred from the network in Figure 2.1 using the above three rules of d-separation:

- $R \perp A \mid E$ because node E blocks the path $R - E - A$ (case 2).

- $E \perp B$ because node A blocks the path $E - A - B$ (case 3).

- $B \perp C \mid A$ because node A blocks the path $B - A - C$ (case 1).

- $B \perp R \mid E, A$ because node E blocks the path $B - A - E - R$ (case 2).

Definition 4 (Dependency Model) *Any model* M *of a set of variables* $\mathcal{U}$ *from which one can determine whether* $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ *is true for all possible subsets* $\mathbf{X}$,$\mathbf{Y}$ *and* $\mathbf{Z}$ *is called a dependency model.*

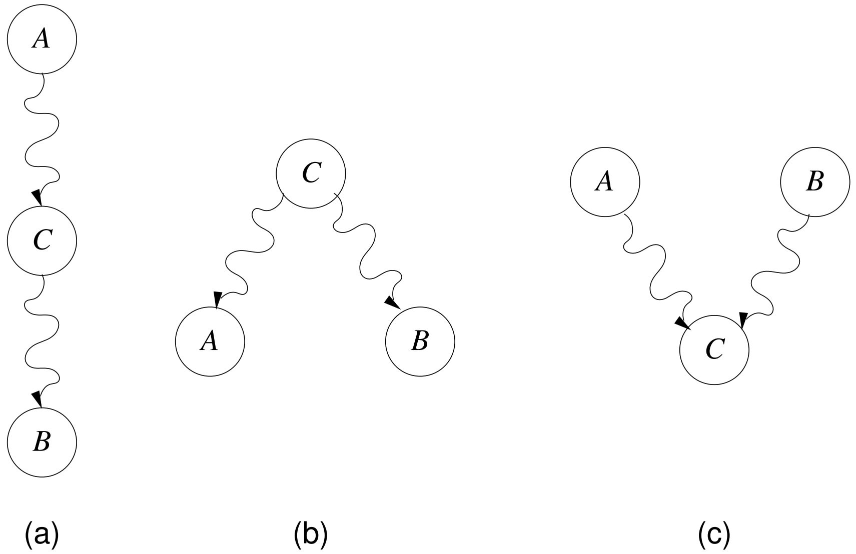

Figure 2.2 The path $A-C-B$ is blocked in the figure on the left provided C is in the conditioning set $\mathbf{Z}$. The path $A-C-B$ is blocked in the center figure if C is in the conditioning set $\mathbf{Z}$. The path $A-C-B$ is blocked in the figure on the right if neither C nor any of its descendants are in $\mathbf{Z}$.

A joint probability distribution P is a dependency model of the domain over which it is defined.

Definition 5 (I-map) *A DAG D is said to be an I-map of a dependency model M if every d-separation condition displayed in D corresponds to a valid conditional independence relationship in M,* i.e., *if for every three disjoint sets of vertices $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ we have*

$$\mathbf{X} \perp_D \mathbf{Y} \mid \mathbf{Z} \implies \mathbf{X} \perp_M \mathbf{Y} \mid \mathbf{Z}$$

A DAG is a ***minimal*** *I-map of M if none of its arrows can be deleted without destroying its I-map property.*

Intuitively, a DAG D is an I-map of a dependency model M if every independence statement derived from D using the rules of d-separation is also valid in M.

Definition 6 (Bayesian network) *Given a probability distribution P on a set of variables $\mathcal{U}$, a DAG D is called a Bayesian network of P iff D is a minimal I-map of P.*

This definition states a DAG D is called a Bayesian network of a joint probability distribution P if and only if every independence relation derived from D is also present in P.

Definition 7 (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 G is also present in P* i.e.

$$\mathbf{X} \perp_B \mathbf{Y} \mid \mathbf{Z} \iff \mathbf{X} \perp_P \mathbf{Y} \mid \mathbf{Z}$$

Intuitively, the above definition states that a Bayesian network B is faithful to a joint distribution P if every independency *and* dependency (lack of independence) implied by the network is also present in P. A Bayesian network is said to be **faithful** if it is faithful to the probability distribution it represents.

We now give an important theorem that relates the concept of a Markov blanket to a Bayesian network. This theorem allows us to precisely determine the Markov blanket of a variable in a network by "reading" its structure.

Theorem 1. *If a Bayesian network B is faithful, then for every variable T, $\mathbf{B}(T)$ is unique and is the set of parents, children and spouses of T (proof in [Tsamardinos and Aliferis, 2003]).*

The theorem shows how the notion of a Markov blanket is closely related to a Bayesian network. Using the above theorem, the Markov blanket of any given variable can be readily "read" off the structure of the corresponding faithful Bayesian network. The Markov blankets of the variables in Figure 2.1 are given below:

- $\mathbf{B}(E) = \{R, A, B\}$ (R, A are children; B is a spouse).
- $\mathbf{B}(B) = \{A, E\}$ (A is a child; E is a spouse).
- $\mathbf{B}(R) = \{E\}$ (E is a parent).
- $\mathbf{B}(A) = \{E, B, C\}$ (E, B are parents; C is a child).
- $\mathbf{B}(C) = \{A\}$ (A is a parent).

The rest of this thesis will describe algorithms that can efficiently recover the Markov blanket of a given variable by making statistical decisions about the independencies present in the data set under consideration.

This concludes this chapter on the basics. In the next chapter, we shall see why the discovery of Markov blankets is of much importance and we shall also review past work related to Markov blanket discovery. All algorithms assume that a single Bayesian network can faithfully represent the distribution of the domain, which implies (by Theorem 1) that there exists a unique Markov blanket for every variable in the domain.

CHAPTER 3. Review of Related Work

3.1 Feature selection

We first define the feature selection problem, discuss the two approaches that attempt to solve it and then show how Markov blankets can be useful in determining an optimal solution for algorithms that take one of the approaches.

3.1.1 The feature selection problem

A large number of problems in Machine Learning deal with supervised learning. In this setting, a learner induces a model from a data set that is defined over a set of features. Each instance in this set is labeled with a *class* label that identifies the category or class to which the instance belongs. The learnt model is used to classify an input pattern or instance, typically represented by a set of value assignments to features, into one of many classes. The goal of a learner is to learn a model that yields a good classification accuracy on unseen data, called generalization performance.

Some real-world data sets give us a wealth of attributes and/or data to use for learning. While having abundant data is usually very helpful, the same cannot be said about having too many attributes. In fact, most inductive methods generalize worse given too many attributes than given a subset of those attributes [Caruana and Freitag, 1994]. The attribute selection problem (also called feature selection or feature subset selection problem) refers to the task of finding a "good" subset of features. [Langley, 1994] and [Dash and Liu, 2000] survey a number of algorithms that attempt to solve this problem. As pointed out in [Dash and Liu, 2000], feature selection is defined by many authors by looking at it from various angles. The following lists those that cover a range of definitions:

Idealized Find a minimally sized feature subset that is necessary and sufficient to find the target concept.

Classical Find a subset of m features from a set of n features, $m < n$, such that the value of a criterion function is optimized over all subsets of size m.

Improving prediction accuracy Find a subset of features that helps improve accuracy of prediction or helps decrease the model size without significantly lowering prediction accuracy.

Approximating original class distribution Find a subset that approximates the original class distribution as closely as possible.

3.1.2 Existing approaches

Feature selection algorithms can be classified into two categories [John et al., 1994]: *filter* and *wrapper* methods. In the filter method, feature selection is done independently of the learning algorithm used to construct the classifier. The selected feature set is then used to induce a classfier. The wrapper method on the other hand uses the learning algorithm for selecting a good subset of features. At each stage of the wrapper method a feature set is evaluated by using the learning algorithm to build a classifier from the feature set. The accuracy of the resulting classifier is used to compare different sets of features and the set that yields the maximum accuracy is selected.

The set of features selected by the filter approach may not be optimal in terms of the predictive accuracy of the classifier because the optimal set may be dependent on the inductive and representational biases of the learning algorithm that is used to construct the classifier [John et al., 1994]. However, the filter approach is computationally faster. This is because the wrapper approach induces a new model for each feature subset it evaluates. This technique makes the wrapper approach feasible only if the learning algorithm is relatively fast. So, how well can we do using filter methods alone? As it turns out, [Koller and Sahami, 1996] and [Tsamardinos and Aliferis, 2003] show that Markov blanket discovery is particularly useful in the feature subset selection problem.

Before we review related work, we need some definitions. An attribute is said to be *irrelevant* if it is unconditionally independent of everything. An attribute is said to be *redundant* with respect to some attribute T if its value is fully determined (or even approximately determined) by some set of features S that does not include T. It should be noted that both irrelevant attributes and redundant attributes pose problems to a learning algorithm. However, some of the early work on filter methods concentrated only on eliminating irrelevant attributes [Langley, 1994]. Figure 3.1 gives examples of irrelevant and redundant attributes under the assumption that the data under consideration was generated by a Bayesian network faithful to it.

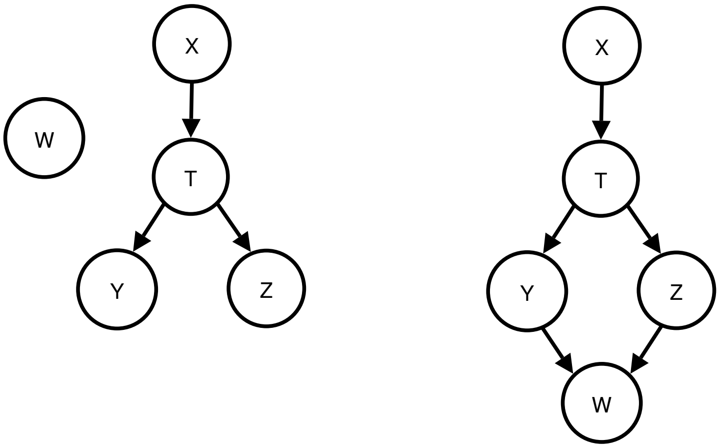

Figure 3.1 W is an irrelevant attribute in the figure on the left whereas it is a redundant attribute in the figure with respect to T on the right.

[Singh and Provan, 1996] and [Koller and Sahami, 1996] describe some of the first filter methods that use information-theoretic approaches to solve the attribute selection problem. The use of information-theoretic approaches is of interest to us because these techniques can filter out irrelevant *and* redundant attributes. Singh and Provan describe the Info-AS algorithm for inducing *selective* Bayesian network classifiers. In the attribute selection phase, the algorithm starts out with the empty set of attributes. It then greedily adds attributes, adding at each step the attribute that maximizes some information-theoretic metric (like Conditional

Informational Gain or Conditional Gain Ratio, for example). Even though the Info-AS algorithm seems similar to the KS algorithm (described in section 3.1.3), [Koller and Sahami, 1996] give a convincing argument that Info-AS's *forward selection* procedure does not guarantee finding a good subset of features.

3.1.3 Optimal filter methods

It was Koller and Sahami's paper [Koller and Sahami, 1996], that first created a framework for defining the theoretically optimal[1] filter method for this problem. Because the Markov blanket of an attribute subsumes its information content (*i.e.* that attribute does not give any additional information about the rest of the variables, including the class variable, beyond what is already in its blanket), it can be shown using this framework that such an attribute can be removed from the feature set without affecting the classifier's performance. Of course, this holds only if the attribute's blanket does not contain the class variable. Otherwise, the attribute should not be removed. Using this framework, we can reason that attributes that are either irrelevant or redundant will be removed from the feature set. An irrelevant feature will be removed based on a Markov blanket consisting of the empty set of features whereas an redundant feature will be removed by using S, the set of features that determine its value, as its Markov blanket. If this process is repeated a number of times, all irrelevant and redundant features will eventually be removed from the domain.

[Koller and Sahami, 1996] give the following algorithm (Figure 3.2) to find an approximate solution using this framework. Intuitively, the algorithm first determines candidate blankets of fixed size for each feature X in $\mathcal{U} - \{C\}$, where C is the class variable. It then removes the feature for which the candidate blanket is closest, in the KL-divergence sense, to being a Markov blanket for that feature. The distance "metric" used to determine how close a candidate blanket $\mathbf{M}$ of a feature X is to the actual Markov blanket is given below:

$$\delta(X \,|\, \mathbf{M}) \;=\; \sum_{\mathbf{m},x} \Pr(\mathbf{M}=\mathbf{m}, X=x) \cdot D(\Pr(C \,|\, \mathbf{M}=\mathbf{m}, X=x) \,\|\, \Pr(C \,|\, \mathbf{M}))$$

[1] A theoretically optimal filter method does not mean that the selected feature set gives the best predictive accuracy when used by a classification algorithm.

This process is repeated until a prespecified number of features are removed. One of the drawbacks of this algorithm is that the number of features to be removed is often not known a priori. In addition, the use of fixed-size sets as blankets does not ensure that a blanket subsumes all the information content of the removed feature, since the set is only an approximation to the actual Markov blanket. [Tsamardinos et al., 2003] provide some experimental results and describe some problems of this algorithm mostly due to a number of naïve assumptions it makes.

```
KS($\mathcal{D}, K, L$)
    ▷ $K$ is the size of candidate blankets, $L$ is the number of features to be removed
    for each pair $(X, Y)$ such that $X, Y \in \mathcal{U}$ do
        Compute $\gamma_{XY} = \sum_{x,y} \Pr(X = x, Y = y) \cdot D(\Pr(C \mid X = x, Y = y) \parallel \Pr(C \mid Y = y))$
    end for
    $\mathbf{G} \leftarrow \mathcal{U}$
    repeat
        for each feature $X \in \mathbf{G}$ do    ▷ Compute candidate blankets
            Let $\mathbf{M}_X$ be the set of $K$ features $Y \in \mathbf{G} - \{X\}$
                    such that $\gamma_{XY}$ is smallest
            Compute $\delta(X \mid \mathbf{M}_X)$
        end for
        $W \leftarrow \arg\min_{X \in \mathbf{G}} \delta(X \mid \mathbf{M}_X)$
        $\mathbf{G} \leftarrow \mathbf{G} - \{W\}$    ▷ Remove feature $W$
    until $L$ features are removed
    return $\mathbf{G}$
```

Figure 3.2 The KS algorithm.

[Tsamardinos and Aliferis, 2003] also study the use of Markov blankets for feature selection. Examined in this work are the concepts of feature relevancy, the feature selection problem and the distinction between wrapper methods and filter methods. They conclude, under certain assumptions, that the Markov blanket of the class variable is the optimal feature set for classification problems and present the IAMB algorithm for Markov blanket discovery. We will look at IAMB and other blanket discovery algorithms in detail in the next section.

3.2 Markov Blanket Discovery Algorithms

3.2.1 Introduction

In today's world, many scientific fields benefit from decision-theoretic techniques and many systems use such techniques to make decisions. Some examples of such systems include medical diagnosis systems, expert systems and systems used in automated troubleshooting. Decision theoretic techniques are useful because they allow explicit management of uncertainty and trade-offs.

A Bayesian network is handy tool for decision making involving uncertainties. It is a graphical model that encodes probabilistic relationship among variables of interest [Heckerman, 1995]. As noted in [Heckerman, 1995], a Bayesian network offers a number of advantages for data analysis, some of which are given below:

1. The model can handle situations where some data entries are missing because it encodes dependencies among all variables.
2. It also allows us to infer causal relationships among variables.

For these reasons, there is a tremendous amount of interest in automatically discovering the structure of Bayesian networks from data. In this respect, two approaches to learning Bayesian network structure have emerged [Margaritis and Thrun, 1999a]:

- those that employ independence properties of the underlying network that produced the data in order discover parts of its structure
- those that learn a local maximum likelihood network structure for representing the data, disregarding independencies in it.

The first of these approaches is of much interest to us because of some recent work by Margaritis and Thrun.

3.2.2 The Markov Blanket GS algorithm

[Margaritis and Thrun, 1999a] present the first algorithm that makes use of Markov blankets for Bayesian network structure discovery. As a first step, the structure discovery algorithm invokes the Grow-Shrink (GS) algorithm (also presented in this paper) on each variable in the domain, in order to discover its Markov blanket. The algorithm then utilizes the blanket information for inducing the structure of the Bayesian net. The GS algorithm is given in Figure 3.3.

```
GS($\mathcal{D}, T$)
$\mathbf{B}(T) \leftarrow \emptyset$
while $\exists\, X \in \mathcal{U} - \{T\}$ such that $X \not\perp T \mid \mathbf{B}(T)$ do    ▷ Growing phase
        $\mathbf{B}(T) \leftarrow \mathbf{B}(T) \cup \{X\}$
end while
while $\exists\, X \in \mathbf{B}(T)$ such that $X \perp T \mid \mathbf{B}(T) - \{X\}$ do ▷ Shrinking phase
        $\mathbf{B}(T) \leftarrow \mathbf{B}(T) - \{X\}$
end while
return $\mathbf{B}(T)$
```

Figure 3.3 The GS algorithm.

The GS algorithm has certain nice properties that make it extremely useful for discovering Markov blankets. First of all, under some assumptions, the algorithm is provably sound *i.e.* it can recover the *exact* Markov blanket of the variable under consideration. The assumptions made are:

(i) the existence and faithfulness of a Bayesian network to the data set under consideration — this implies the existence and uniqueness of the blanket, and

(ii) the assumption that the conditional independence tests are reliable.

Refer to [Margaritis and Thrun, 1999b] for a formal proof of correctness.

Another nice property is that the algorithm is scalable — it requires only $O(n)$ conditional independence tests for discovering a Markov blanket. Before we move on, it should be mentioned that the time complexity of Markov blanket discovery algorithms is often measured by the number of conditional independence tests that are performed in the process.

The $O(n)$ running-time requirement of GS can be proved in the following manner: First, it can be shown that the GS algorithm needs to make at most three passes over the entire set of variables in the growing phase. In the first pass, all the parents and children and possibly some spouses of the target variable T are added. The remaining spouses are added in the second pass. The third one verifies that nothing else needs to be added. Second, the shrinking phase requires just one pass through the variables in $\mathbf{B}(T)$. Therefore, we can conclude that the running-time of this algorithm is $O(n)$. In order to speed up the algorithm, Margaritis and Thrun suggest the use of mutual information to order the variables before the growing phase. The intuition is that this will minimize the chances chances of making a second pass in the growing phase because variables that are close to the target, and those in its blanket, are identified and added to the blanket as early as possible (thereby rendering all the other variables conditionally independent of the target).

Despite having these desired properties, the GS algorithm can sometimes fail to recover the correct Markov blanket because assumption (ii) is violated in certain cases. This occurs when the data set is so small in size so as to make most of the conditional independence tests unreliable. It also occurs whenever the growing phase of the algorithm adds many false positives to the blanket — a condition that will make the tests at the end of the growing phase and those performed in the entire shrinking phase unreliable, even if the available data set is reasonably large due to the large size of the conditioning set. The larger the size of the conditioning set, the less accurate are the estimates of conditional probabilities and hence the independence tests are not reliable. Even the use of mutual information to guide blanket discovery does not totally mitigate this problem because the heuristic used in GS is "static"

[Tsamardinos et al., 2003]; the resulting ordering over the $\mathcal{U} - \{T\} - \mathbf{B}(T)$ can actually lead to the spouses of the target variable being considered very late in the growing phase. Whereas the former problem (lack of data) cannot be handled by any existing blanket recovery technique, we shall see how the latter one can be solved to some extent.

3.2.3 The IAMB and Inter-IAMB algorithms

[Tsamardinos et al., 2003] describe a few variants of GS that attempt that solve the problem mentioned at the end of the previous section. The interesting ones presented are the Incremental Association Markov Blanket (IAMB) and Interleaved-IAMB (Inter-IAMB) algorithms. These algorithms are also sound under the same set of assumptions used by the GS algorithm. They are interesting to study because they are not affected as much by assumption (ii) as is GS. For each algorithm, we first describe the algorithm, examine its pros and cons, and then try to ascertain its running-time requirements.

The IAMB algorithm is given in Figure 3.4. As we noted earlier, this algorithm and Inter-IAMB are variants of the GS algorithm. The difference between GS and the IAMB algorithms is that the IAMB algorithms employ what Tsamardinos et al. call a “dynamic” heuristic.

Just like GS, the IAMB algorithm also uses a two-phase approach for discovering Markov blankets. In the growing phase all variables that belong to the blanket and possibly some false positives enter $\mathbf{B}(T)$. The shrinking phase then identifies these false positives and removes them. However, there is a difference in how IAMB orders variables in the growing phase. IAMB reorders the set of variables each time a new variable enters the blanket. The reordering can be done using an information-theoretic heuristic like mutual information. This dynamic reordering of variables is what makes IAMB and its variants perform better — when the blanket contains spouses of T, the reordering ensures that the spouses of T are also considered early on in the growing phase, unlike what happens in GS. This in turn causes IAMB (and its variants, of course) to add fewer false positives during the growing phase. As noted earlier, fewer false positives lead to more reliable tests of conditional independence, which in turn help

```
IAMB($\mathcal{D}, T, h$)
$\mathbf{B}(T) \leftarrow \emptyset$
while $\mathbf{B}(T)$ has changed do                    ▷ Growing phase
    $S \leftarrow \{A \mid A \in \mathcal{U} - \{T\} - \mathbf{B}(T)\}$
    $X \leftarrow \arg\max_{A} h(A, T \mid \mathbf{B}(T))$
    if $X \not\perp T \mid \mathbf{B}(T)$ then
        $\mathbf{B}(T) \leftarrow \mathbf{B}(T) \cup \{X\}$
    end if
end while
for each attribute $A \in \mathbf{B}(T)$ do             ▷ Shrinking phase
    if $A \perp T \mid \mathbf{B}(T) - \{A\}$ then
        $\mathbf{B}(T) \leftarrow \mathbf{B}(T) - \{A\}$
    end if
end for
return $\mathbf{B}(T)$
```

Figure 3.4 The IAMB algorithm.

the process of Markov blanket discovery.

Now that we have presented the algorithm, we can analyze its time complexity. It is important to note that performing a conditional independence test between two variables is identical to calculating the conditional mutual information between them and then doing a significance test on the value obtained — see Chapter 4 and also Appendix A for a formal proof of why this is so. This allows us to treat both these "queries" as requiring the same amount of time in $O()$ notation and therefore we do not distinguish between them. In the growing phase, IAMB adds at most $n = |\mathcal{U}|$ variables to $\mathbf{B}(T)$. Each addition reorders the remaining variables based on the conditional mutual information between them and the target (the conditioning set being the current blanket). Since each reordering takes $O(n)$ time, the entire growing phase takes $O(n^2)$ time in the worst case. The shrinking phase takes at most

$O(n)$ time. Therefore, the total time complexity is $O(n^2)$. However, this is a very conservative upper-bound. In practice, IAMB often performs better.

Promising as it may seem, even IAMB suffers from the problem of the addition of false positives to the blanket in the growing phase, though to a lesser extent. To see why, consider the Bayesian network in figure 3.5.

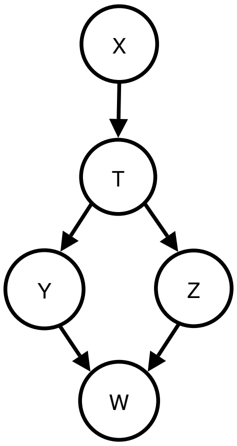

Figure 3.5 Though W is outside the blanket of T, it might be added to the blanket early on in the growing phase.

The true blanket of the target variable T is the set $\{X, Y, Z\}$; the variable W is outside the blanket. For the purpose of illustration, assume that the IAMB algorithm has just entered the growing phase *i.e.* the blanket is empty. In such a scenario, IAMB orders the variables in $\mathcal{U} - \{T\}$ using mutual information as its heuristic[2]. When using an information-theoretic heuristic like mutual information, there is a possibility that the mutual information between T and W is very high; higher than the mutual information between T and Y, and T and Z taken separately. If this were indeed the case, W would be considered first and would be added to the blanket before Y and Z are added. Such possibilities arise whenever there exist multiple paths for the flow of information between the variables being considered. When there is just one path, however, this possibility never arises. In our example, the two paths connecting T and W are $T \rightarrow Y \rightarrow W$ and $T \rightarrow Z \rightarrow W$.

[2] Strictly speaking, the heuristic used is conditional mutual information where the conditioning set is the set of variables added to the blanket thus far.

The Inter-IAMB algorithm partially solves the problem mentioned in the above paragraph. The algorithm is given in Figure 3.6.

```
Inter-IAMB($\mathcal{D}, T, h$)
$\mathbf{B}(T) \leftarrow \emptyset$
while $\mathbf{B}(T)$ has changed do                                  ▷ Growing phase
        $S \leftarrow \{A \mid A \in \mathcal{U} - \{T\} - \mathbf{B}(T)\}$
        $X \leftarrow \arg\max_{A} h(A, T \mid \mathbf{B}(T))$
        if $X \not\perp T \mid \mathbf{B}(T)$ then
                $\mathbf{B}(T) \leftarrow \mathbf{B}(T) \cup \{X\}$
        end if
        for each attribute $A \in \mathbf{B}(T)$ do                  ▷ Shrinking phase
                if $A \perp T \mid \mathbf{B}(T) - \{A\}$ then
                        $\mathbf{B}(T) \leftarrow \mathbf{B}(T) - \{A\}$
                end if
        end for
end while
return $\mathbf{B}(T)$
```

Figure 3.6 The Inter-IAMB algorithm.

The algorithm only partially solves the problem because the false positives might not be detected until late into the growing phase of the algorithm and also because not all of them might be detected.

The key difference between IAMB and Inter-IAMB is that the shrinking phase is interleaved into the growing phase in Inter-IAMB. This has some advantages and disadvantages to it. The advantage of interleaving these two phases is that Inter-IAMB can eliminate some[3] of the false positives in the current blanket as the algorithm progresses during the growing

[3] We say *some* and not *all* because Inter-IAMB uses only the current blanket to determine whether or not a variable is a false positive.

phase, without having to wait until the rowing phase is complete. This is the key to the often superior "accuracy" of Inter-IAMB over IAMB in terms of being able to recover the right Markov blanket. However, this improved accuracy comes with a performance penalty. Since the two phases are interleaved, the Inter-IAMB algorithm makes at most $2n$ mutual information/conditional independence computations for each variable that enters the blanket in the growing phase (as opposed to the at most n computations made by IAMB). The $O()$ notation makes this performance penalty invisible since the time complexity is still $O(n^2)$. However, it is often the case Inter-IAMB conducts at least as many tests to recover the blanket as does IAMB (on the same target variable) and often a lot more.

Another disadvantage of Inter-IAMB comes from an unwanted side-effect of interleaving the two phases: a variable may be added to and removed from the blanket multiple times, though this does not happen very often. This does not happen either in GS or in IAMB because their growing and shrinking phases are not interleaved. The following example will help illustrate this point. Assume that the Bayesian network in figure 3.7 generated the data set under consideration.

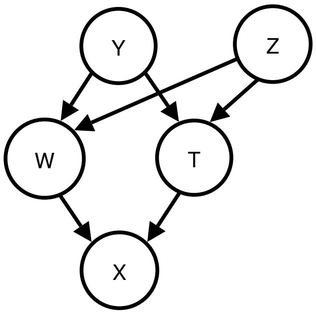

Figure 3.7 W is added twice and removed once by Inter-IAMB.

With this Bayesian network, let us assume that the Inter-IAMB algorithm progresses as follows while calculating the Markov blanket of T.

1. GP[4]: W is added first to $\mathbf{B}(T)$. (As we mentioned before, this could happen whenever there exist multiple paths between the target and the variable under consideration. In

[4] Growing phase

this case, the two paths are $T - Y - W$ and $T - Z - W$.)

2. SP[5]: No variables are removed from the blanket in the shrinking phase since W is unconditionally dependent on T.

3. GP: Y is added next to $\mathbf{B}(T)$.

4. SP: Again, no variables are removed since $W \not\perp T \mid Y$ and $Y \not\perp T \mid W$.

5. GP: Next, Z is added to the blanket.

6. SP: This time however, W is removed from $\mathbf{B}(T)$ since $W \perp T \mid Y, Z$.

7. GP: X is added next to the blanket.

8. SP: No variables are removed since X, Y and Z are all dependent on T (given the other two).

9. GP: W is again added to the blanket.

10. SP: No variables are removed in this iteration either.

Clearly, W was added twice to the blanket and removed once before the algorithm terminated. In fact, W could have been added and removed twice if the network is like the one shown in Figure 3.8.

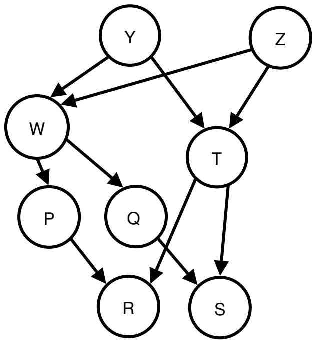

Figure 3.8 W is added twice and removed twice by INTER-IAMB.

[5] Shrinking phase

The reason for this behavior is simple: Inter-IAMB decides whether a variable is a false positive or not based only on the "current" blanket. Sometimes, the current blanket suffices to correctly identify a false positive. In other cases, like the example just described for instance, it raises a false alarm causing Inter-IAMB to delete it first, only to add it later on.

In the next chapter, we describe the statistical techniques used by our algorithm to make conditional independence decisions. The chapter following this one describes our algorithm for Markov blanket discovery.

CHAPTER 4. Testing for Statistical Independence

In this chapter we discuss statistical independence testing in discrete domains which is used in our Markov blanket discovery algorithm. The outline of this chapter is as follows: we first introduce some terminology and give definitions of two important statistics used in statistical independence tests. We then outline our approach for testing for conditional independence. Finally, we describe some of the issues that one might encounter in the presence of insufficient data and outline the steps we take to mitigate this problem.

4.1 Preliminaries

Data sets in discrete domains that contain two or more qualitative variables can naturally be represented as a *cross-tabulation* or a *contingency table*. For example, Table 4.1 shows a data set of 5375 tuberculosis deaths defined over two variables, namely sex and type of tuberculosis causing death. A table such as Table 4.1 is known as a 2 × 2 contingency table because each variables takes on two values. Each cell in this table contains the *count* or *frequency* of the instances in the data set that fall under the corresponding category. In our thesis, we are only interested in contingency tables involving two attributes.

	Sex	
Type ↓	Males	Females
Tuberculosis from respiratory system	3534	1319
Other forms of tuberculosis	270	252

Total: 5375

Table 4.1 A 2×2 contingency table: Deaths from tuberculosis [Everitt, 1977].

Contingency tables are useful in making statistical decisions about the attributes in the data set under consideration. An important question involves deciding whether the variables of interest are independent or not. Such a question is termed a *hypothesis.* The primary hypothesis is called the *null hypothesis* — it indicates our *belief* about the data set. The other hypothesis, against which we test our null hypothesis, is called the *alternative hypothesis.* In this thesis, we always use the same null hypothesis for every statistical independence test that we perform — it is the statement 'the variables are independent'. The alternative hypothesis is simply the opposite: 'the null hypothesis is not true', which, in our case reads 'the variables are dependent'. Having defined the null and the alternate hypotheses, we can describe how to test the truth of the null hypothesis using data. This can be done using a *test statistic.* The test statistic is a numerical quantity calculated from a data set that helps in the decision making process. Informally, the test statistic is compared against a known cut-off value, and if the statistic is found to be greater than the cut-off value we conclude that the null hypothesis is false and accept the alternate hypothesis. A well-known independence tests is Pearson's chi-squared test. The chi-squared test dates back to the early 20th century, having been introduced by Pearson in 1904 to test for the statistical independence of two variables.

4.2 The $\mathcal{X}^2$ and $\mathcal{G}^2$ statistics

The chi-squared test is a simple test that determines whether or not two discrete probability distributions are statistically different. Typically, these are the actual distribution and an "expected" distribution (defined over the same set of variables). The actual distribution is what is obtained from the data set being studied whereas the expected distribution is the distribution we would have obtained if the null hypothesis was true on that data set. For an $r \times c$ contingency table defined over two variables X and Y (*i.e.* a contingency table for X and Y where X takes on r values and Y takes on c values), Pearson's chi-squared test uses the following quantity (also known as the chi-squared statistic):

$$\mathcal{X}^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O(i,j) - E(i,j))^2}{E(i,j)} \tag{4.1}$$

In the above equation, $O(i,j)$ is the frequency or count of instances that have $X = x_i$ and $Y = y_j$, and $E(i,j)$ is the *expected* number of instances that have $X = x_i$ and $Y = y_j$ under the null hypothesis. Since we assume the null hypothesis of independence, the expected number of instances that have $X = x_i$ and $Y = y_j$ is

$$E(i,j) = N \cdot \widehat{\Pr}(X = x_i) \cdot \widehat{\Pr}(Y = y_j) \tag{4.2}$$

where $\widehat{\Pr}(\cdot)$ denotes the maximum-likelihood estimate of $\Pr(\cdot)$. One can compare $\mathcal{X}^2$ to a cut-off value to decide whether or not the variables are indeed independent. Providing none of the expected values are very small, the distribution of $\mathcal{X}^2$ can be shown to be approximately that of χ^2 (the chi-squared distribution) with ν degrees of freedom (in short, $\mathcal{X}^2 \sim \chi^2_\nu$). For an $r \times c$ table and the null hypothesis that we use, it can be shown that ν is equal to $(r-1) \times (c-1)$. If we use a $100(1-\alpha)\%$ confidence level, the cut-off value to compare against would be $\chi^2_{\nu,\alpha}$. Most statistical texts provide tables that list these cut-off values for various values of ν and α. If $\mathcal{X}^2 > \chi^2_{\nu,\alpha}$, we reject the null hypothesis (independence) and accept the alternate hypothesis (dependence). Typically, a 95% confidence level ($\alpha = 0.05$) is used for most statistical tests. The interested reader is referred to [Everitt, 1977, Upton, 1978, Agresti, 1990] for a detailed explanation and the mathematical underpinnings of these concepts.

A simple example will help illustrate the use of the chi-squared test. We will use Table 4.1 for the purpose of this example. It is repeated here as Table 4.2 in a slightly different form. Also given below in Table 4.3 are the expected counts for each cell of Table 4.2.

	Sex		
Type ↓	Males	Females	*Row totals*
Tuberculosis from respiratory system	3534	1319	4853
Other forms of tuberculosis	270	252	522
Column totals	3804	1571	5375

Table 4.2 Observed counts and marginal totals for Table 4.1.

	Sex		
Type ↓	Males	Females	*Row totals*
Tuberculosis from respiratory system	3434.6	1418.4	4853
Other forms of tuberculosis	369.4	152.6	522
Column totals	3804	1571	5375

Table 4.3 Expected counts and marginal totals under the hypothesis of independence for Table 4.1.

The expected counts in Table 4.3 were obtained using maximum-likelihood probability estimates as given by eq. 4.2. For example, the expected count in the first cell of the table is obtained as follows:

$$\begin{aligned} E_{\text{Male, Respiratory TB}} &= N \times \widehat{\Pr}(\text{Male}) \times \widehat{\Pr}(\text{Respiratory TB}) \\ &= N \times \frac{O(\text{Male})}{N} \times \frac{O(\text{Respiratory TB})}{N} \\ &= \frac{3804 \times 4853}{5375} \\ &= 3434.6 \end{aligned}$$

Using the table of expected counts (Table 4.3), we can calculate the chi-squared statistic in the following manner:

$$\begin{aligned} \mathcal{X}^2 &= \frac{(3534 - 3434.6)^2}{3434.6} + \frac{(1319 - 1418.4)^2}{1418.4} + \frac{(270 - 369.4)^2}{369.4} + \frac{(252 - 152.6)^2}{152.6} \\ &= 101.35 \end{aligned}$$

For our data, $r = c = 2$, and therefore the degrees of freedom $v = (2-1)(2-1) = 1$. If we set α at 0.05, the cut-off value is $\chi^2_{1,0.05} = 3.84$. Since $\mathcal{X}^2 > \chi^2_{1,0.05}$, we conclude (with 95% confidence) that the variables Sex and Type of tuberculosis are *dependent*.

An alternative to $\mathcal{X}^2$ is $\mathcal{G}^2$, which is another statistic that also allows us to compare two discrete distributions. For an $r \times c$ contingency table, it is given by

$$\mathcal{G}^2 = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} O(i,j) \ln\left(\frac{O(i,j)}{E(i,j)}\right) \tag{4.3}$$

where $O(i,j)$ and $E(i,j)$ are defined as before. $\mathcal{G}^2$ is sometimes called "likelihood ratio χ^2" or "Maximum Likelihood statistic." $\mathcal{G}^2$, just like $\mathcal{X}^2$, is asymptotically distributed as a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom[1], meaning that the cut-off values for both statistics (for fixed ν, α) are the same.

[Ku and Kullback, 1974] mention that the chi-squared statistic $\mathcal{X}^2$ is actually an approximation to the log-likelihood statistic $\mathcal{G}^2$. (Refer to [Williams, 1976] for a proof.) [Ku and Kullback, 1974] and [Williams, 1976] also mention that the additional advantage of using $\mathcal{G}^2$ is that it is addi-

[1] This holds for an $r \times c$ table and the null hypothesis that we use.

tive whereas $\mathcal{X}^2$ is not. We shall see shortly why this is important to us. For these reasons, we use $\mathcal{G}^2$ instead of $\mathcal{X}^2$ in all our independence tests.

4.3 Testing for Conditional Independence

So far we have seen how the $\mathcal{X}^2$ and $\mathcal{G}^2$ statistics are useful in testing for the (unconditional) independence of two variables. Testing for conditional independence, however, is not as straightforward as the the unconditional case and differs from it in certain aspects. First, the test statistic used in the conditional test is computed using a different approach. Second, the cut-off value that the test statistic is compared against is also obtained in a different manner. Yao and Tritchler [Yao and Tritchler, 1993] give a detailed treatment of testing for conditional independence and suggest a test statistic that can be used for an exact analysis. However, we use a simpler and easier to obtain a test statistic that is similar to the one used by them. We use the $\mathcal{G}^2$ statistic and assume that the cells in the contingency table arise out of the normal approximation to the hypergeometric distribution. In such a case, the test for conditional independence can be carried out as follows.

Assume, for the purpose of discussion, that we want to test for the conditional independence of X and Y given $\mathbf{Z}$ in some data set $\mathcal{D}$, where $\{X, Y\} \cup \mathbf{Z} \subseteq \mathcal{U}$. Then, the test statistic that we use for testing for the conditional independence of $X, Y \,|\, \mathbf{Z}$ is:

$$\mathcal{G}^2(X, Y \,|\, \mathbf{Z}) = \sum_{\mathbf{z} \in \{\text{All configurations of } \mathbf{Z}\}} \mathcal{G}^2(X, Y \,|\, \mathbf{Z} = \mathbf{z}) \tag{4.4}$$

where $\mathcal{G}^2(X, Y \,|\, \mathbf{Z} = \mathbf{z})$ is the Maximum Likelihood statistic for that subset of the data set where $\mathbf{Z} = \mathbf{z}$.

There is an interesting property of χ^2 (chi-squared) probability distributions that helps us compute the distribution that the above test statistic will approach asymptotically. This property is called the *additive property* and it states that:

> *If $X_1, \ldots, X_n$ are independent random variables with $\chi^2_{\nu_1}, \ldots, \chi^2_{\nu_n}$ distributions respectively, then $Y = \sum_{i=1}^{n} X_i$ is a random variable that has a $\chi^2_{\sum_{i=1}^{n} \nu_i}$ distribution.*

We already know from the previous section that the $\mathcal{G}^2(X,Y)$ statistic for testing the unconditional independence of two variables X and Y approaches the chi-squared distribution asymptotically. Let us assume that $\mathcal{G}^2(X,Y \mid \mathbf{Z}=\mathbf{z})$ has $\nu_{\mathbf{z}}$ degrees of freedom. Now, if we use the fact that the $\mathcal{G}^2(X,Y \mid \mathbf{Z}=\mathbf{z})$ statistics are in fact random variables and that they are independent of each other at each configuration $\mathbf{z}$ of $\mathbf{Z}$ (see footnote below), then we can clearly conclude that the test statistic $\mathcal{G}^2(X,Y \mid \mathbf{Z})$ has a chi-squared distribution with $\nu = \sum_{\mathbf{z}} \nu_{\mathbf{z}}$ degrees of freedom. Using the notation in Chapter 2, the degrees of freedom ν are $\nu = (r_X - 1) \times (r_Y - 1) \times r_{\mathbf{Z}}$, and the cut-off value is $\chi^2_{\nu,\alpha}$.

An example will help make things clear. Suppose we want to test for the conditional independence of X and Y given a variable Z using some data set $\mathcal{D}$. Assume that X and Y are binary variables whereas Z takes on three values. Given below are three hypothetical 2×2 contingency tables for X and Y at each value of Z.

	$Y = y_1$	$Y = y_2$
$X = x_1$	164	18
$X = x_2$	3559	1299

$Z = z_1$

	$Y = y_1$	$Y = y_2$
$X = x_1$	16	102
$X = x_2$	191	13012

$Z = z_2$

	$Y = y_1$	$Y = y_2$
$X = x_1$	2	690
$X = x_2$	68	879

$Z = z_3$

Table 4.4 Testing for Conditional Independence of X and Y given Z.
R = 20,000 instances.

Using equation 4.4, the test statistic would be

$$\begin{aligned}
\mathcal{G}^2(X,Y \mid Z) &= \mathcal{G}^2(X,Y \mid Z=z_1) + \mathcal{G}^2(X,Y \mid Z=z_2) + \mathcal{G}^2(X,Y \mid Z=z_3) \\
&= 31.16 + 43.81 + 61.86 \\
&= 136.83
\end{aligned}$$

and it follows a χ^2_ν distribution with $\nu = (2-1) \times (2-1) \times 3 = 3$ degrees of freedom. Assuming $\alpha = 0.05$, the cut-off value to compare this statistic against is $\chi^2_{3,0.05} = 7.81$. Since

[1] This is true due to the assumption that instances in the data set are independently and identically distributed and the fact that the $\mathcal{G}^2_{\mathbf{z}}$ statistics are calculated from disjoint subsets of the data.

$\mathcal{G}^2(X, Y \mid Z) > \chi^2_{3,0.05}$, we reject the null hypothesis of independence and conclude that X and Y are conditionally dependent given Z.

4.4 Dealing with insufficient data

Insufficient data presents a lot of problems when working with statistical inference techniques like the independence tests mentioned earlier. In particular, insufficient data can lead to the problem of zero cell frequencies when we are dealing with contingency tables. Since the $\mathcal{X}^2$ and $\mathcal{G}^2$ statistics rest on the normal approximation to the hypergeometric distribution, these approximations get strained when there are zero frequency cells or when the expected counts are 'very small'. The latter phrase has generally been interpreted to mean that for an independence test to be reliable, all cells have non-zero expected values and at least 80% of the cells should have expected values greater than 5 [Everitt, 1977, Upton, 1978]. Unfortunately, when working on real-world data sets such cases frequently arise.

We now describe some of the heuristic "solutions" that we use to help mitigate these problems proposed by other researchers. The first, proposed by [Brin et al., 1997], tries to overcome the problem of low expected counts by simply ignoring these cells when calculating the $\mathcal{X}^2$ or $\mathcal{G}^2$ sums (equations 4.1, 4.3). Brin et al. argue that if the variables involved in the independence test would have been associated due to the contribution of a very small cell, then the association would involve very rare events. Hence the variables can be thought of as being independent. The second heuristic "solution" uses what is called CT-support for ensuring that a contingency table has enough data to perform a statistical test on it.

The concept of CT-support was also introduced in [Brin et al., 1997]. CT-support is defined in terms of two parameters (p, s): p is a percentage and s is a count. We say that a table has (p, s) CT-support if at least $p\%$ of the cells in the table have counts at least s. We use CT-support in two ways. One, we use it to determine if an unconditional independence test can be carried out reliably. Two, we also use it in conditional independence tests of X, Y given $\mathbf{Z}$ to determine all those configurations $\mathbf{z}$ of $\mathbf{Z}$ that have insufficient data. We then ignore these configurations when we perform the conditional independence test. By this we mean that we do not count their contributions toward the final $\mathcal{G}^2(X, Y \mid \mathbf{Z})$ sum and also do not count their contributions to the degrees of freedom. This method seems to works well in practice as our experiments demonstrate.

An unresolved problem remains: what should be done in those situations tests where all configurations of $\mathbf{Z}$ do not have CT-support, or where a contingency table used in an unconditional test does not have CT-support? Unfortunately, there is not much we can do in these cases and we are forced to rely on assumptions: we can assume that the variables are dependent or we could assume they are independent without actually performing a test. Once we pick an assumption, we use it in all independence tests in an experiment.

CHAPTER 5. The Fast-IAMB Algorithm

5.1 Motivation and Algorithm

A number of Markov blanket discovery algorithms were presented in Chapter 3. We analyzed each algorithm in detail and also described their relative advantages and disadvantages. The following table briefly summarizes the properties of these algorithms.

Algorithm	Advantages	Disadvantages
GS	$O(N)$ running time; fewer than $3N$ tests in total.	Ordering of variables unspecified. Doesn't remove false positives in the Growing phase.
IAMB	Uses a dynamic heuristic to order variables.	$O(N^2)$ running time; less than $N^2 + N$ tests in total. Doesn't remove false positives in the Growing phase.
Inter-IAMB	Uses a dynamic heuristic to order variables. Removes some false positives in the Growing phase.	$O(N^2)$ running time. Slower than IAMB in most cases.

Table 5.1 This table summarizes the key characteristics of the Markov blanket discovery algorithms seen so far.

In this chapter, we present a new algorithm called Fast-IAMB for Markov blanket discovery. The algorithm is sound in that it discovers the exact Markov blanket under the same set of assumptions used by existing algorithms *viz.* the existence and uniqueness of the blanket, the faithfulness of a Bayesian network to the data set under consideration and the assumption that the total number of conditional independence tests performed by the algorithm are reliable. The proof of soundness is the same as that found in [Tsamardinos et al., 2003]. This is because Fast-IAMB is similar to IAMB and Inter-IAMB in how it discovers a Markov blanket. However, there are certain key differences that make Fast-IAMB faster and also more reliable in many cases. We will use the term “reliable” to mean the ability of an algorithm to recover the correct Markov blanket. We shall first examine the properties of GS, IAMB and Inter-IAMB given above to obtain some key insights into their working that will help us understand what constitutes a reliable and fast algorithm. We then describe how we can achieve these goals and present the pseudo code of Fast-IAMB. We conclude this section with a brief note on choosing a good heuristic.

From Table 5.1 it is clearly evident that though GS is faster than IAMB or Inter-IAMB, it is not as reliable. On the other extreme, Inter-IAMB is very reliable but is often slower than GS or IAMB. Fast-IAMB is an algorithm that strives to achieve a balance between these two extremes. In particular, the following characteristics were the motivating idea behind Fast-IAMB:

1. The algorithm should be reliable like Inter-IAMB.

2. The algorithm should also be able to quickly converge to the blanket like GS.

In other words, the algorithm should be able to identify the correct Markov blanket in most cases. Clearly, to be reliable the algorithm should use a dynamic heuristic to order the variables. In addition, the algorithm should also remove some false positives as it progresses in the Growing phase. As we have seen with Inter-IAMB, interleaving the Growing and Shrinking phases helps. However, since we also want the algorithm to be fast, we should somehow reduce the number of conditional independence tests required to converge to the right Markov blanket.

The key idea behind reducing the number of conditional independence tests is to add not one, but a number of variables after each reordering of $\mathcal{U} - \{T\} - \mathbf{B}(T)$. Adding a number of variables that are close to the target (when sorted by conditional mutual information) will very likely add some true members of the blanket. Thus, we amortize the cost of sorting the remaining variables in $\mathcal{U} - \{T\} - \mathbf{B}(T)$, over multiple variables that are added to the blanket without re-sorting.

There is a well-known “folk-theorem” (as [Koller and Sahami, 1996] put it) that states that probabilistic influence or association between variables tends to attenuate over distance. Therefore, mutual information or $(1 - p\text{-value})$ of the $\mathcal{G}^2$ statistic between variables is a natural measure of determining those variables that are close to the target in the original Bayesian network.

Now that we have decided on the above strategy, the natural question is to determine the number of variables that should be added to the blanket at each iteration. We use the following heuristic: we add variables as long as the conditional independence tests are reliable enough *i.e.* we have enough data for conducting them. For this purpose, we use a numeric parameter k that is supplied by the user. This parameter denotes the minimum average number of instances per cell of a contingency table that should be present if we are to assume that the conditional independence tests are reliable. Let the target variable be T, and the current Markov blanket be $\mathbf{B}(T)$. Let the next variable that we consider for addition to $\mathbf{B}(T)$ be X. The aim is to be able to perform a reliable conditional independence test between T and X given the current Markov blanket. For this to happen, we must have

$$\frac{R}{r_T \times r_{\mathbf{B}(T)} \times r_X} \geqslant k$$

The intuition behind the above formula is this: There are R instances in the entire data set, $r_T \times r_X$ cells in each contingency table, and there are as many contingency tables as there are configurations of the current blanket $\mathbf{B}(T)$ *i.e* $r_{\mathbf{B}(T)}$. Therefore, if the average number of cells is greater than or equal to k, a conditional independence test performed between T and X given $\mathbf{B}(T)$ would be reliable and hence we can determine whether we should or should not add X to the blanket. This process is repeated as long as the average number of instances per cell is greater than or equal to k. Using the ideas just presented, we obtain the algorithm shown in Figure 5.1.

Note that while adding variables in lines $6 - 13$ no conditional independence tests are actually performed. The only value that is computed after the addition of each variable is the average number of instances per cell (line 7), which can be done in constant time.

The algorithm works well if k instances per cell on average are sufficient to discover the Markov blanket. Of course, when working with real world data sets, data is scarce and more than k instances on average might be necessary. Therefore some modification must be made to the algorithm to handle such cases. In Chapter 4 we mentioned that one can assume inde-

```
Fast-IAMB($\mathcal{D}, T, h, k$)
 1  $\mathbf{B}(T) \leftarrow \emptyset$
 2  $S \leftarrow \{A \mid A \in \mathcal{U} - \{T\}$ and $A \not\perp T\}$
 3  while $S \neq \emptyset$ do                                  ▷ Growing phase
 4          $\langle X_1, \ldots, X_{|S|} \rangle \leftarrow S$ sorted according to $h$
 5          flag = FALSE
 6          for i = 1 to $|S|$ do
 7                  if $\frac{R}{r_{X_i} \times r_T \times r_{\mathbf{B}(T)}} \geqslant k$ then
 8                          $\mathbf{B}(T) \leftarrow \mathbf{B}(T) \cup \{X_i\}$
 9                  else
10                          flag = TRUE
11                          goto 14
12                  end if
13          end for
14          for each attribute $A \in \mathbf{B}(T)$ do            ▷ Shrinking phase
15                  if $A \perp T \mid \mathbf{B}(T) - \{A\}$ then
16                          $\mathbf{B}(T) \leftarrow \mathbf{B}(T) - \{A\}$
17                  end if
18          end for
19          $S \leftarrow \{A \mid A \in \mathcal{U} - \{T\} - \mathbf{B}(T)$ and $A \not\perp T \mid \mathbf{B}(T)\}$
20  end while
21  return $\mathbf{B}(T)$
```

Figure 5.1 The Fast-IAMB algorithm.

pendence or dependence without actually performing a statistical test. A similar workaround can be applied to this situation to yield the following code:

```
if [flag = TRUE] and [no variables were removed in the shrinking phase]
and [assuming dependence] then
        return $\mathcal{U} - \{T\}$
else
        $S \leftarrow \{A \mid A \in \mathcal{U} - \{T\} - \mathbf{B}(T)$ and $A \not\perp T \mid \mathbf{B}(T)\}$
end if
```

If there are fewer than k instances per cell on average, the algorithm attempts to shrink the current Markov blanket before adding more variables. If the shrinking phase does not remove any variable from the blanket and we assume that variables are dependent when data is insufficient to reliably conduct a statistical test, then the algorithm returns the set of all the variables except the target as the blanket. On the other hand, if there are more than k instances on average, or a variable is removed in the shrinking phase, or we assume independence of variables when data is insufficient, the algorithm proceeds as usual.

Before we conclude this section, there is one minor but important remark to be made with regard to the IAMB and INTER-IAMB algorithms. For the sake of convenience, steps 4–7 of both these algorithms are listed once again below:

```
4  $X \leftarrow \arg\max_{A} h(A, T \mid \mathbf{B}(T))$
5  if $X \not\perp T \mid \mathbf{B}(T)$ then
6          $\mathbf{B}(T) \leftarrow \mathbf{B}(T) \cup \{X\}$
7  end if
```

[Tsamardinos and Aliferis, 2003] suggest that any information-theoretic distance measure can be used in step 4 as the heuristic function. They recommend the use of mutual information as a good heuristic. Unfortunately, using mutual information or even the $\mathcal{G}^2$ statistic can lead to premature termination of these algorithms causing them to not discover the correct Markov blanket. To see why, we first note that mutual information is directly proportional to $\mathcal{G}^2$ (refer

	T, X	T, Y
Mutual information	5.3	6.1
Degrees of freedom	$(2-1) \times (2-1) = 1$	$(2-1) \times (4-1) = 3$
p-value of test	0.0213	0.1068

Table 5.2 Example showing why the use of mutual information alone can sometimes be misleading.

to Appendix A for a proof). We also note that the $\mathcal{G}^2$ statistic has an associated ν parameter that denotes the degrees of freedom. The test statistic and the degrees of freedom are both used to determine if the statistic is above the cut-off value. The problem with using only $\mathcal{G}^2$ (or mutual information) as a measure of association is that a higher value of the test statistic does not necessarily mean that the variable under consideration is more correlated with the target — variables with large number of values are disproportionately weighted (favored) in the mutual information calculation. Consider the following example. Assume that T is a binary variable for which we wish to determine the blanket; X and Y are other variables in the domain $\mathcal{U}$, taking on 2 and 4 values respectively. Further assume that the mutual information between T, X and between T, Y is as given in Table 5.2.

We see that $\mathcal{G}^2(T, X)$ is smaller than $\mathcal{G}^2(T, Y)$. However, since the degrees of freedom of the two test statics are different, the use of $\mathcal{G}^2$ can be misleading. Indeed, this becomes very clear if we look at the p-value of the tests of conditional independence. The p-value of a statistical test is the probability of obtaining a value greater than the test statistic under the null hypothesis *i.e.* it is the area under the chi-squared curve that lies to the right of the statistic. The smaller the area, the more significant is the departure from the null hypothesis. Using the values given in Table 5.2, the p-value of the independence test between T and X is lower than that of the test between T and Y. Therefore, if we assume a 95% confidence level ($\alpha = 0.05$) X and T are determined dependent whereas Y and T are determined independent. Thus, the use of $\mathcal{G}^2$ or mutual information would cause IAMB or Inter-IAMB to only consider Y and terminate before X is considered. Thus, mutual information is not an accurate indicator of the strength of association between variables. Hence, we suggest the use of ($1 - p$-value) as a true measure of association.

5.2 Experimental Results

In order to empirically evaluate the performance of Fast-IAMB with the other Markov blanket discovery algorithms, we ran a number of experiments the results of which are described in this section. The experiments were conducted on both synthetic and real-world data sets. These data sets are listed in Table 5.3.

	Data set	No. of features	No. of instances
synthetic	Alarm20k	37	20,000
real-world	Ling-spam	201	2,893
	Adult	9	45,222

Table 5.3 List of data sets used in our experiments.

The confidence level of each independence test was set to 95% ($\alpha = 0.05$). The CT-support parameters p and s were set to 10% and 30 instances/cell respectively. For each experiment, we report the number of conditional independence tests conducted to discover the blanket, the time taken, and a distance measure that indicates the "fitness" of the discovered blanket. The latter is the average (over all variables not in the blanket) of the expected KL-divergence between the PMF[1] of T given its blanket $\mathbf{B}(T)$ and the variable X (that is outside the blanket), and the PMF of T given just its blanket. In mathematical terms, this measure is:

$$\delta(T \mid \mathbf{B}(T)) = \frac{1}{|\mathcal{U} - \{T\} - \mathbf{B}(T)|} \sum_{X \in \, \mathcal{U} - \{T\} - \mathbf{B}(T)} \text{Expected-KL-Div}(\Pr(T \mid \mathbf{B}(T), X) \parallel \Pr(T \mid \mathbf{B}(T))) \tag{5.1}$$

where

$$\text{Expected-KL-Div}(\Pr(T \mid \mathbf{B}(T), X) \parallel \Pr(T \mid \mathbf{B}(T))) = \sum_{\mathbf{b}, x} \Pr(\mathbf{B}(T) = \mathbf{b}, X = x) \cdot D(\Pr(T \mid \mathbf{B}(T) = \mathbf{b}, X = x) \parallel \Pr(T \mid \mathbf{B}(T) = \mathbf{b}))$$

Intuitively, the above measure tells us how close $\mathbf{B}(T)$ is to being a Markov blanket for variable T. If $\mathbf{B}(T)$ is indeed a blanket for T, then the distance $D(\Pr(T \mid \mathbf{B}(T) = \mathbf{b}, X = x) \parallel \Pr(T \mid \mathbf{B}(T) = \mathbf{b}))$ would be zero because T and X would be independent given $\mathbf{B}(T)$. If it is

[1] Probability mass function.

an approximate blanket, then we can expect this measure to be close to zero. This measure is similar to the one proposed in [Koller and Sahami, 1996].

5.2.1 The Alarm data set

The Alarm data set is a synthetic data set that was generated by sampling the Alarm Bayesian network using logic sampling. The Alarm network was created by medical experts for monitoring patients in intensive care. It consists of 37 variables with the arity of each ranging between two and four. The network was sampled to generate a data set of size 20,000 instances.

Figure 5.2 shows a performance comparison of Fast-IAMB, IAMB and Inter-IAMB on the Alarm data set containing 20,000 instances. The Y axis denotes the number of conditional independence tests executed by the algorithms for each of the 37 variables on the X axis. The variables along the X-axis are sorted on the number of independence tests of Fast-IAMB for ease of comparison. From the graphs it can be seen that in almost all of the cases Fast-IAMB converges to the blanket using fewer conditional independence tests than IAMB or Inter-IAMB. The top figure in Figure 5.2 shows the graphs obtained by assuming dependence in the presence of insufficient data. The bottom figure shows the graphs obtained by assuming independence when there is insufficient data to reliably perform a statistical test. For both these sets of graphs, the parameter k was set to 10. Figure 5.3 shows the actual running times of the algorithms on each of the 37 variables in the data set. As expected, the running time is a linear function of the number of conditional independence tests performed by the algorithms. Figure 5.4 shows the graphs of the "fitness" of the Markov blankets recovered by these algorithms. In addition to being faster, Fast-IAMB is also able to correctly identify the Markov blankets for most of the variables. In fact, its reliability is about the same as that of IAMB or Inter-IAMB on the Alarm data set. GS is the faster algorithm in all cases. However, Figure 5.4 shows the reliability of GS is not as good as that of the other algorithms.

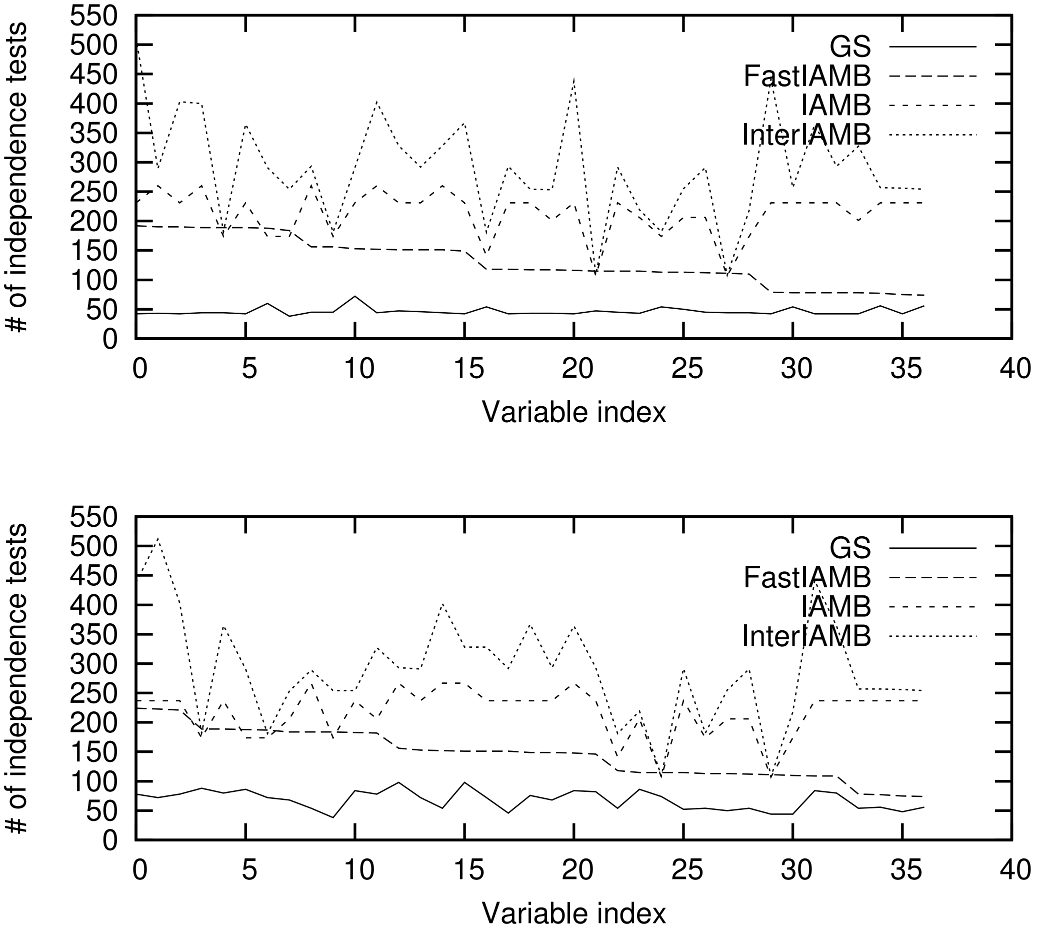

Figure 5.2 Alarm data set with 20k instances: Graphs showing no. of conditional independence tests performed on each variable; assuming dependence (top graph) or assuming independence (bottom graph).

5.2.2 The Ling-spam data set

The Ling-spam corpus contains discussion email messages on a linguistics mailing list along with the spam emails received by that list. The data set was built by indexing a corpus of 2891 messages on 201 different attributes (in this case, the attributes are words extracted from the message bodies). There are 2412 non-spam messages and 481 spam ones.

Figure 5.5 shows a performance comparison of Fast-IAMB, IAMB and Inter-IAMB on the Ling-spam data set. The Y axis denotes the number of conditional independence tests

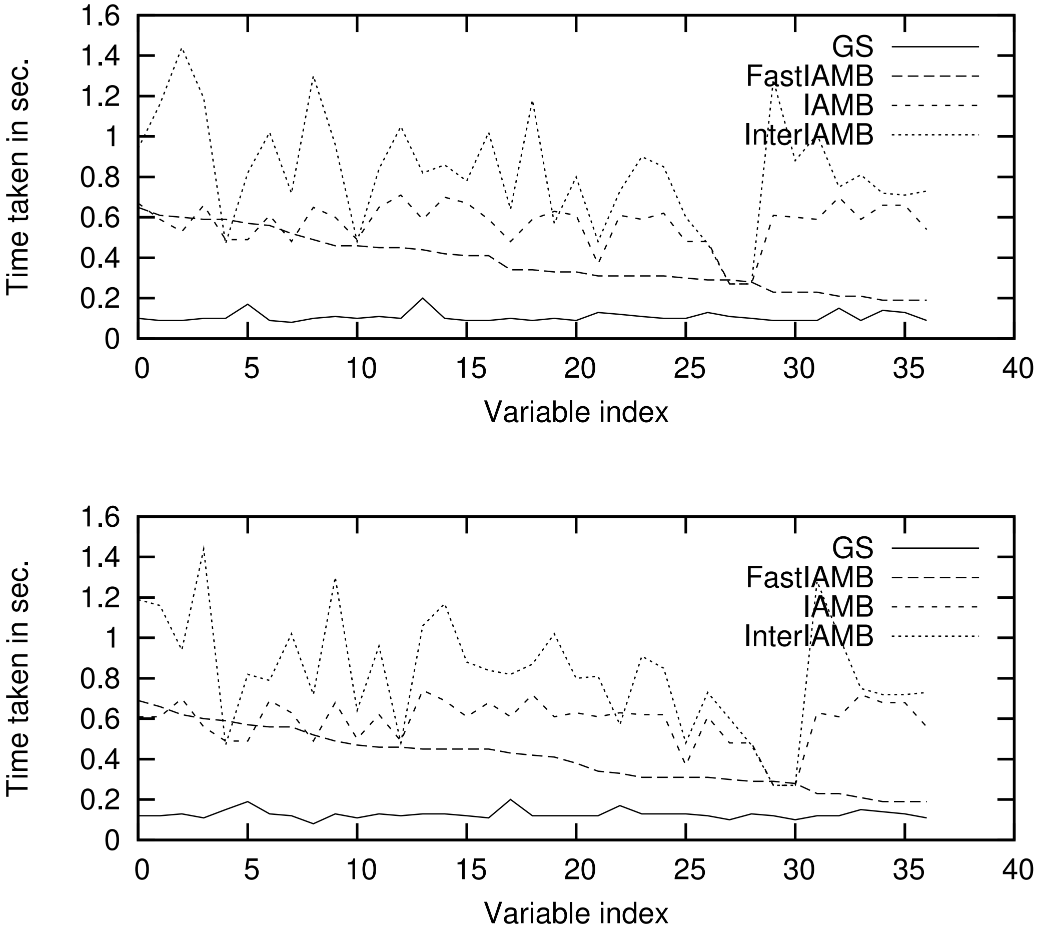

Figure 5.3 Alarm data set with 20k instances: Graphs showing time taken to discover the Markov blankets; assuming dependence (top graph) or assuming independence (bottom graph).

executed by the algorithms for each of the 201 variables on the X axis. Again, the results are sorted on the number of independence tests of Fast-IAMB for ease of comparison. As can be seen from these graphs, in some cases Fast-IAMB outperforms IAMB and Inter-IAMB. In others, Fast-IAMB performs slower than the other two. From our analysis of the algorithm, the main reason for the large number of tests performed by Fast-IAMB in these cases seems to be the limited size of the Ling-spam data set as compared to the large number of attributes over which it is defined. The small size sometimes causes the conditional independence tests to yield a set of independence statements that cannot be captured by graphical models like

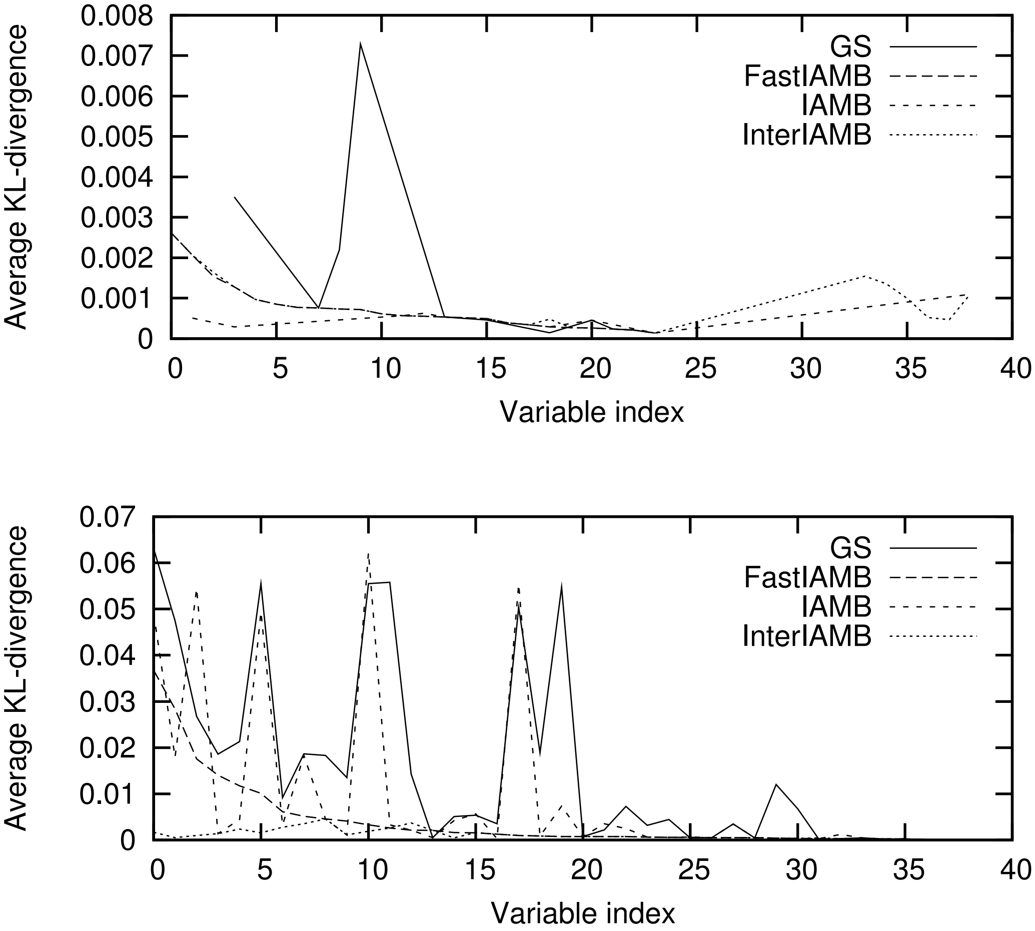

Figure 5.4 Alarm data set with 20k instances: Graphs showing “fitness” of the recovered blankets using the measure described above; assuming dependence (top graph) or assuming independence (bottom graph).

Bayesian networks.

For example, consider the following set of dependence/independence statements:

- $T \not\perp X \mid \mathbf{S}$
- $T \not\perp Y \mid \mathbf{S} \cup \{X\}$
- $T \perp X \mid \mathbf{S} \cup \{Y\}$
- $T \perp Y \mid \mathbf{S}$

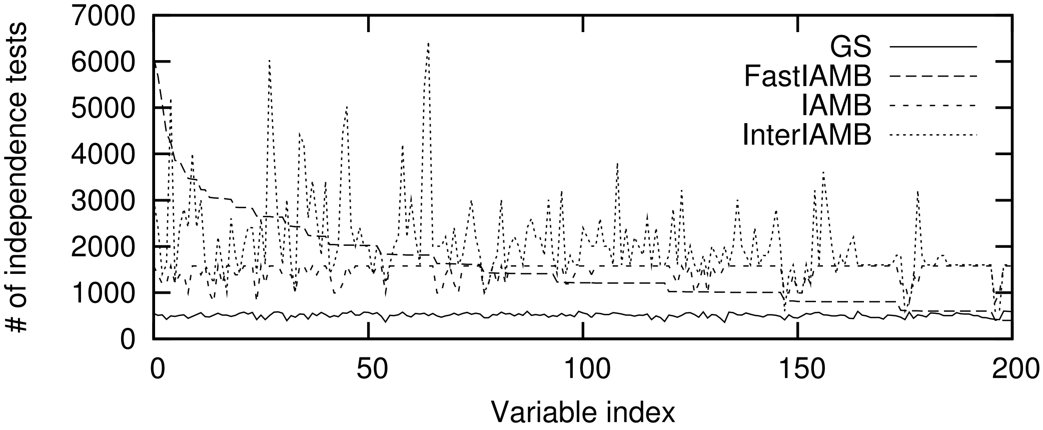

Figure 5.5 Ling-spam data set: Graphs showing no. of independence tests performed. When data was scarce, variables were assumed to be independent.

This set of statements cannot be captured by any Bayesian network that is faithful. In situations like these, there is also the possibility of algorithms like Inter-IAMB and Fast-IAMB to go into an infinite loop when trying to determine a Markov blanket. The example just presented is one such case. X is initially added to the blanket of T, followed by the addition of Y. X and Y are then removed in that order from the blanket. The Inter-IAMB and Fast-IAMB algorithms would then proceed to add X (since it is now dependent), and the cycle continues. As such, modifications should be made to these algorithms to appropriately handle these cases. A simple modification is to store the blanket at each iteration of the algorithm and halt whenever the current blanket matches a previously encountered blanket.

In addition, the assumption of the uniqueness of a Markov blanket is sometimes violated when data is scarce, leading to the existence of multiple sets of variables that can shield the target from the rest of the variables in the universe. This can possibly cause different blanket discovery algorithms to output different blankets for the same target variable.

Figure 5.6 shows the time taken by Fast-IAMB, IAMB and Inter-IAMB to discover the Markov blankets for each variable in the data set. Figure 5.7 shows the "fitness" of the discovered blankets. Table 5.4 gives the total times for each algorithm (summed for all variables).

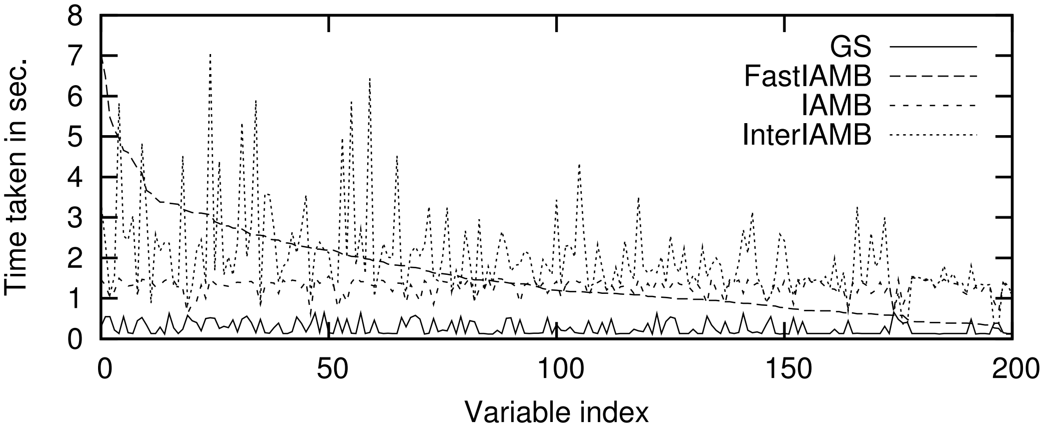

Figure 5.6 Ling-spam data set: Graphs showing time taken to discover the blankets.

Algorithm	Total time (sec.)
GS	53.12
Fast-IAMB	322.92
IAMB	256.69
Inter-IAMB	405.29

Table 5.4 Ling-spam data set: Total time taken by each algorithm.

5.2.3 The Adult data set

The data set contains demographic information about individuals gathered form the Census Bureau database. The original data set had 16 attributes in total, 9 of which were discrete and 7 continuous. We only used the 9 discrete attributes for our experiment. The original data set also contained missing values; after removing instances with missing attributes the resulting data set contained 45,222 instances. The attributes used were:

1. Working class

2. Education

3. Martial status

4. Occupation

5. Relationship

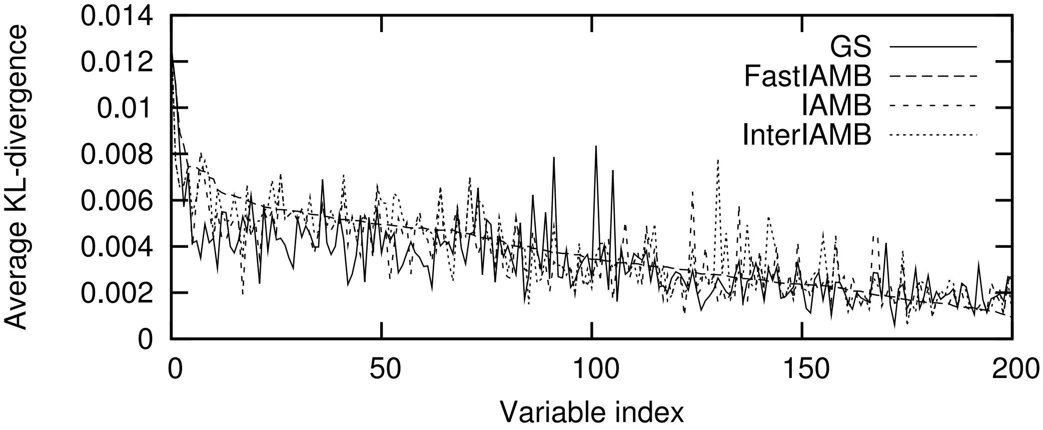

Figure 5.7 Ling-spam data set: Graphs showing "fitness" of discovered blankets.

6. Race

7. Sex

8. Native country

9. Income

The experiments we conducted assumed independence when data was scarce to perform a reliable statistical test and the parameter k was set to 10. Figure 5.8 shows how Fast-IAMB compares to the other two algorithms in terms of the number of independence tests conducted. Figure 5.9 shows the time taken to discover the Markov blankets for each of the 9 variables in the data set. It is clearly evident from these two figures that Fast-IAMB quickly converges to the blanket when compared to the other two algorithms. Figure 5.10 shows that the reliability of Fast-IAMB is comparable to that of IAMB and Inter-IAMB.

5.3 Conclusions and Future Research

In this thesis we address the important role of Markov blankets discovery, which is useful for a number of tasks including feature selection and Bayesian network induction. We reviewed past work on Markov blanket discovery and analyzed the advantages and disadvantages of each

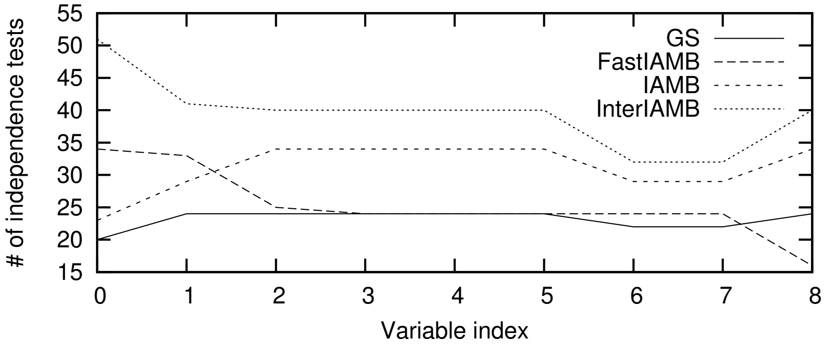

Figure 5.8 Adult data set: Graphs showing number of independence tests performed to discover the blankets. When data was scarce, variables were assumed to be independent.

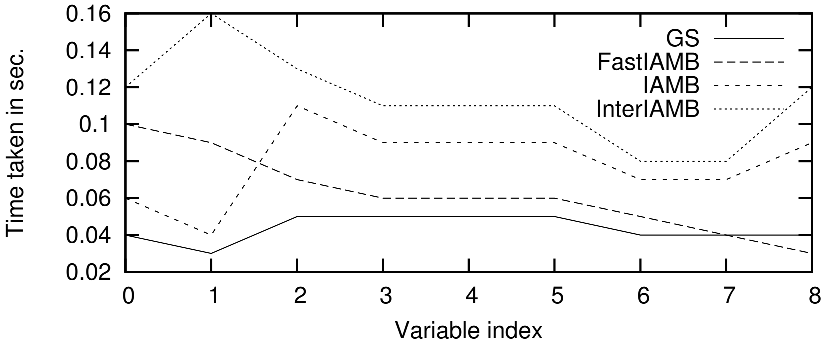

Figure 5.9 Adult data set: Graphs showing time taken to discover the blankets.

algorithm. Our main contribution is a novel algorithm called Fast-IAMB for the induction of Markov blankets that employs a fast heuristic to quickly converge to the Markov blanket. Our experiments indicate that our algorithm is faster than IAMB and its variants since it conducts fewer independence tests. In addition, the algorithm correctly identifies the blankets under assumptions. Its reliability is comparable to IAMB and Inter-IAMB in the majority of the cases.

An interesting direction for further research is to study the use of Markov blankets for association rule discovery. Association rules have been an active topic of recent research in the field of Data Mining. It is conceivable that knowledge of Markov blankets could be useful in

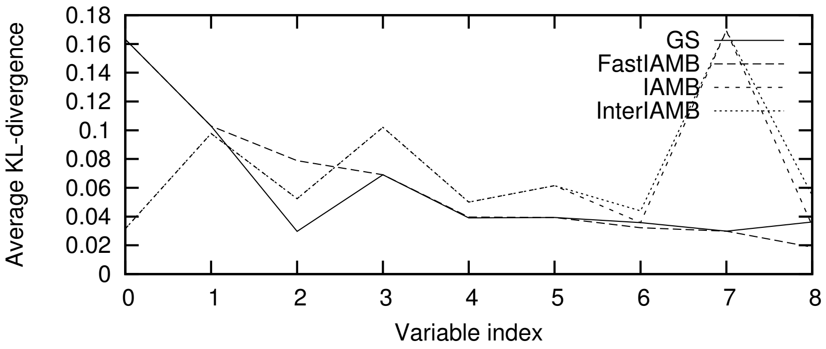

Figure 5.10 Adult data set: Graphs showing “fitness” of discovered blankets.

the pruning of uninteresting association rules. Another potential use of Markov blankets is the development of efficient algorithms to discover conditional correlation rules [Brin et al., 1997] in large data sets.

An important concern in any blanket discovery algorithm is the fact that data can be scarce and conditional independence tests can be unreliable. Gene expression analysis is a prime example of such a field where the size of the data set (the number of instances in it) is extremely small compared to the number of attributes. An interesting research direction is the development of algorithms that can recover approximate Markov blankets in those cases. One possible solution to this is the assumption that a known type of distribution or model generated the data set and then try to recover the Markov blanket.

In addition, the assumption of faithfulness of the underlying Bayesian network is somewhat restrictive. As we have seen in the previous section there can be a set of independence statements that cannot be captured by any single Bayesian network. Hence, another direction of future research could study the the possibility of recovering Markov blankets of variables in arbitrary probability distributions and not just distributions that can be represented by graphical models such as Bayesian networks.

APPENDIX A. Mutual Information is proportional to $\mathcal{G}^2$

In this appendix we will prove the following two claims:

Claim 1 $\frac{1}{2N}\mathcal{G}^2(X,Y) = MI(X,Y)$

Claim 2 $\frac{1}{2N}\mathcal{G}^2(X,Y \mid \mathbf{Z}) = MI(X,Y \mid \mathbf{Z})$

where MI stands for Mutual Information, $\mathcal{G}^2$ is the likelihood ratio χ^2, N is the size of the data set and $\{X,Y\} \cup \mathbf{Z} \subseteq \mathcal{U}$.

We start by giving the following definitions.

$$\begin{aligned}
MI(X,Y) &\equiv D\left(\widehat{\Pr}(x,y) \parallel \widehat{\Pr}(x)\cdot\widehat{\Pr}(y)\right)\\
MI(X,Y \mid \mathbf{Z}) &\equiv \sum_{\mathbf{z}} \widehat{\Pr}(\mathbf{z})\cdot D\left(\widehat{\Pr}(x,y \mid \mathbf{z}) \parallel \widehat{\Pr}(x \mid \mathbf{z})\cdot\widehat{\Pr}(y \mid \mathbf{z})\right)\\
\mathcal{G}^2(X,Y) &\equiv 2\sum_{x,y} O(x,y)\cdot\ln\left(\frac{O(x,y)}{E(x,y)}\right)\\
\mathcal{G}^2(X,Y \mid \mathbf{Z}) &\equiv \sum_{\mathbf{z}} \mathcal{G}^2(X,Y \mid \mathbf{z})
\end{aligned}$$

where we abbreviate $X = x$ by x for any variable or set of variables X. We shall now prove the above two claims.

Proof of claim 1

$$\begin{aligned}
\frac{1}{2N}\mathcal{G}^2(X,Y) &= \frac{1}{2N}\left[2\sum_{x,y} O(x,y)\cdot\ln\left(\frac{O(x,y)}{E(x,y)}\right)\right]\\
&= \sum_{x,y}\widehat{\Pr}(x,y)\cdot\ln\left(\frac{O(x,y)}{N\cdot\widehat{\Pr}(x)\cdot\widehat{\Pr}(y)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{x,y} \widehat{\Pr}(x,y) \cdot \ln\left(\frac{\widehat{\Pr}(x,y)}{\widehat{\Pr}(x)\widehat{\Pr}(y)}\right) \\
&= D\left(\widehat{\Pr}(x,y) \parallel \widehat{\Pr}(x) \cdot \widehat{\Pr}(y)\right) \\
&= MI(X,Y) \quad \square
\end{aligned}
$$

Proof of claim 2

The proof of claim 2 is much simplified by using the proof of claim 1 to note that

$$\frac{1}{2N_{\mathbf{z}}}\mathcal{G}^2(X,Y\,|\,\mathbf{z}) = MI(X,Y\,|\,\mathbf{z})$$

where $N_{\mathbf{z}}$ denotes the number of instances in the data set that have $\mathbf{Z} = \mathbf{z}$. We now prove the claim.

$$
\begin{aligned}
\frac{1}{2N}\mathcal{G}^2(X,Y\,|\,\mathbf{Z}) &= \frac{1}{2N}\sum_{\mathbf{z}} \mathcal{G}^2(X,Y\,|\,\mathbf{z}) \\
&= \frac{1}{2N}\sum_{\mathbf{z}} 2N_{\mathbf{z}} \cdot MI(X,Y\,|\,\mathbf{z}) \\
&= \sum_{\mathbf{z}} \widehat{\Pr}(\mathbf{z}) \cdot MI(X,Y\,|\,\mathbf{z}) \\
&= \sum_{\mathbf{z}} \widehat{\Pr}(\mathbf{z}) \cdot D\left(\widehat{\Pr}(x,y\,|\,\mathbf{z}) \parallel \widehat{\Pr}(x\,|\,\mathbf{z}) \cdot \widehat{\Pr}(y\,|\,\mathbf{z})\right) \\
&= MI(X,Y\,|\,\mathbf{Z}) \quad \square
\end{aligned}
$$

Bibliography

[Agresti, 1990] Agresti, A. (1990). *Categorical Data Analysis.* John Wiley and Sons Ltd.

[Brin et al., 1997] Brin, S., Motwani, R., and Silverstein, C. (1997). Beyond market baskets: Generalizing association rules to correlations. In Peckham, J., editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 265–276. ACM Press.

[Caruana and Freitag, 1994] Caruana, R. and Freitag, D. (1994). Greedy attribute selection. In *International Conference on Machine Learning*, pages 28–36.

[Dash and Liu, 2000] Dash, M. and Liu, H. (2000). Feature selection for clustering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 110–121.

[Everitt, 1977] Everitt, B. (1977). *The Analysis of Contingency Tables.* Chapman and Hall Ltd.

[Friedman and Goldszmidt, 1998] Friedman, N. and Goldszmidt, M. (1998). Aaai-98 tutorial: Learning bayesian networks from data.

[Heckerman, 1995] Heckerman, D. (1995). A Tutorial on Learning Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research.

[John et al., 1994] John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, pages 121–129. Journal version in AIJ, available at http://citeseer.nj.nec.com/13663.html.

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

[Ku and Kullback, 1974] Ku, H. H. and Kullback, S. (1974). Loglinear models in contingency table analysis. *The American Statistician*, 28(4):115–122.

[Langley, 1994] Langley, P. (1994). Selection of relevant features in machine learning. In *AAAI Fall Symposium on Relevance*, pages 140–144.

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

[Margaritis and Thrun, 1999b] Margaritis, D. and Thrun, S. (1999b). Bayesian network induction via local neighborhoods. Technical Report CMU-CS-99-134, Carnegie Mellon University.

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

[Singh and Provan, 1996] Singh, M. and Provan, G. M. (1996). Efficient learning of selective bayesian network classifiers. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 453–461. Morgan Kaufmann.

[Tsamardinos and Aliferis, 2003] Tsamardinos, I. and Aliferis, C. (2003). Towards principled feature selection: Relevancy, filters, and wrappers. In *Ninth International Workshop on Artificial Intelligence and Statistics*.

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

[Upton, 1978] Upton, G. J. G. (1978). *The Analysis of Cross-tabulated Data.* John Wiley and Sons Ltd.

[Williams, 1976] Williams, K. (1976). The failure of pearson's goodness of fit statistic. *The Statistician*, 25(1):49.

[Yao and Tritchler, 1993] Yao, Q. and Tritchler, D. (1993). An exact analysis of conditional independence in several 2×2 contingency tables. *Biometrics*, 49(1):233–236.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my adviser Dr. Dimitris Margaritis for all his help. His intellectual support, patience and cheerfulness in addressing my questions have been very encouraging and have helped me complete this research.

I would also like to thank my committee members Dr. Vasant Honavar and Dr. Tapabrata Maiti for their helpful discussions and comments. My thanks are also due to the Department of Computer Science and the Graduate College at ISU for providing financial support to complete my degree work.