Integrating Bayesian Networks and Simpson's Paradox in Data Mining

Alex A. Freitas
Ken McGarry
Elon Correa

ABSTRACT. This paper proposes to integrate two very different kinds of methods for data mining, namely the construction of Bayesian networks from data and the detection of occurrences of Simpson's paradox. The former aims at discovering potentially causal knowledge in the data, whilst the latter aims at detecting surprising patterns in the data. By integrating these two kinds of methods we can hopefully discover patterns which are more likely to be useful to the user, a challenging data mining goal which is under-explored in the literature. The proposed integration method involves two approaches. The first approach uses the detection of occurrences of Simpson's paradox as a preprocessing for a more effective construction of Bayesian networks; whilst the second approach uses the construction of a Bayesian network from data as a preprocessing for the detection of occurrences of Simpson's paradox.

1 Introduction

Data mining consists of the (semi-)automatic extraction of interesting patterns from real-world data-sets. Data mining is usually considered the core step in a broader process called knowledge discovery, which includes several steps related to preprocessing of the data to be mined, the data mining step, and other steps related to the post-processing of the discovered patterns. A well-known and informative definition of knowledge discovery is as follows [Fayyad et al., 1996]:

> "Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."

Although this definition is often quoted in the literature, in general it has not been taken very seriously by the data mining and knowledge discovery research community. This claim is supported by the fact that the vast majority of the data mining literature focuses on discovering patterns that are valid - or accurate - ignoring the other aforementioned pattern-quality criteria. Unfortunately, the focus on the maximization of predictive accuracy often hinders the discovery of surprising, novel patterns, which is the kind of pattern that tends to be more interesting and more useful to the user [Silberchatz and Tuzhilin, 1996].

One possible explanation for this focus on accuracy in the literature seems to be that discovering novel, surprising patterns is in general a lot harder than discovering accurate patterns. A couple of examples can illustrate this point, as follows.

[Brin et al., 1997] found, in a Census data set, several rules which were very accurate but were also useless, because they represented obvious patterns in the data, such as "five-year olds don't work", "unemployed residents don't earn income from work" and "men don't give birth". [Tsumoto, 2000] found 29,050 rules, out of which only 220 (less than 1% of them) were considered interesting or unexpected by the user.

These two works are examples of the fact that high accuracy is not a sufficient condition for the interestingness (novelty or surprisingness) or usefulness of a pattern. In addition, although high accuracy is clearly a very desirable property of a discovered pattern, high accuracy is not always a necessary condition for the usefulness or interestingness of a pattern. For instance, [Wong and Leung, 2000] found rules with just 40-60% confidence that were considered, by senior medical doctors, novel and more accurate than the knowledge of some junior doctors.

This paper focuses on Bayesian networks, an increasingly popular data mining technique. In terms of the aforementioned pattern-quality criteria, methods for constructing Bayesian networks from data tend to discover patterns that satisfy the criteria of good accuracy (due to the solid mathematical basis of probability theory) and good comprehensibility (due to the graphical representation of Bayesian networks). However, methods for constructing Bayesian networks are not designed to discover surprising patterns. Hence, it is quite possible that a certain Bayesian network constructed from data be accurate and comprehensible to the user, yet not very interesting, because it is only representing well-known correlations in the data, without representing any novel, surprising pattern to the user. The goal of this paper is to discuss how to remedy this situation, by integrating methods for constructing Bayesian network from data with a method for discovering surprising patterns from data, based on the detection of Simpson's paradox.

The remainder of this paper is organized as follows. Section 2 presents an overview of methods for constructing Bayesian networks from data. Section 3 presents an overview of methods for the discovery of interesting patterns and Simpson's paradox. Section 4 describes the proposed method for integrating the two aforementioned kinds of methods, and Section 5 presents the conclusions.

2 An Overview of Methods for Constructing Bayesian Networks from Data

A Bayesian network is essentially a directed acyclic graph (DAG) where each node represents a random variable (an attribute in the context of data mining) and an edge pointing from node X_i to node X_j means that the value of variable X_j is directly dependent on the value of the variable X_i. Assuming discrete variables (which is the focus of this paper), the strength of the dependence between two variables X_i and X_j connected by an edge is quantified by the conditional probability of the child variable X_j given the parent variable X_i.

A very simple example of a Bayesian network is illustrated in Figure 1.1, showing hypothetical relationships between four variables: the amount of unhealthy food eaten by a person, whether or not certain genetic factors are present in a person, the level of bad cholesterol of a person and the probability of a person having a heart attack. The hypothetical network in Figure 1.1 basically indicates that the probability of heart attack is directly dependent only on the level of the cholesterol of a person; whilst the latter variable is directly dependent on the amount of unhealthy food and genetic factors associated with that person. The conditional probabilities associated with the strengths of dependence between the variables are not shown, for the sake of simplicity.

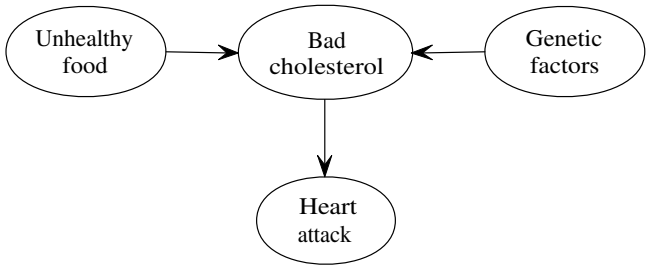

Figure 1.1. A very simple, hypothetical example of a Bayesian network involving 4 variables.

A Bayesian network summarizes the (in)dependencies in the data set being mined in the sense that the joint probability distribution of the set of attributes in the data set being mined - denoted by $p(X)$ - is factorized into the following expression:

$$p(x) = \prod_{i=1}^{M} p(X_i|Par(X_i)), \tag{1.1}$$

where $Par(X_i)$ is the set of variables that are parent nodes of the X_i node in the DAG representing the Bayesian network and M is the number of variables (attributes) in the data set being mined.

There are two major kinds of methods for constructing a Bayesian network from data, namely methods based on conditional independence tests and methods based on a search guided by a scoring function [Blanco, 2005], [Korb and Nicholson, 2004]. The former is based on the assumption that the results of a statistical independence test matches the true independence relationships in the data. A popular method following this approach is the PC algorithm [Spirtes et al., 1993]. This algorithm uses the concept of the order of a conditional independence, which means the number of variables on which the independence between two variables is conditioned [Shipley, 2000]. Hence, a zero-order conditional independence means an independence between two variables without conditioning in any other variable, a first-order conditional independence means an independence between two variables conditioning on just one other variable, and so on. This algorithm starts with the complete undirected graph. Then it reduces the graph by removing edges with zero-order conditional independence. Next it reduces the graph again by removing first-order conditional independencies, and so on. The main problem with this kind of method is that it is very computationally expensive and it does not scale up well to data sets with a large number of attributes, because, for each pair of variables candidate to have an independence relationship, it may have to test the conditional independence for all possible order sizes.

As a result, in the last few years methods based on a different approach, viz. search guided by a scoring function, have significantly grown in popularity. Methods following this approach can be classified according to different criteria, such as the kind of search engine that they use, the kind of scoring function that they use, etc. In our brief review of these methods we focus on the search engine only, which is more relevant for an understanding of the discussion in Section 4. A more comprehensive discussion about methods based on a search guided by a scoring function can be found e.g., in [Blanco, 2005], [Korb and Nicholson, 2004].

Concerning their search engine, methods following this approach can be broadly classified into sequential or population-based methods. Sequential methods work by considering a single candidate solution (a candidate DAG) at a time. They iteratively modify the current candidate solution, trying to improve it as assessed by a given scoring function, until a stopping criterion is satisfied.

The most popular sequential method for constructing Bayesian networks seems to be the B algorithm, which is essentially a hill-climbing, greedy method. This algorithm starts with an empty DAG and at each iteration it adds, to the current candidate solution, the edge that maximizes the value of the scoring function.

Another popular method is the K2 algorithm, which is also essentially a hill-climbing, greedy method. Unlike the B algorithm, K2 requires that the variables be ordered and it requires, as a user-defined parameter, the maximum number of parents of each variable in the DAG to be constructed. Hence, K2 performs a more restricted search than the B algorithm.

Population-based methods work by considering a population of candidate solutions at a time. They iteratively use information from the current population of candidate solutions to create new candidate solutions, again guided by a scoring function, until a stopping criterion is satisfied. In general population-based methods perform a global search in the search space, reducing the risk of getting stuck in local optima in the search space - which often happens with greedy, hill-climbing methods such as the B and K2 algorithms.

In addition, population-based methods normally are stochastic (non-deterministic) methods, whereas sequential methods can be either stochastic or deterministic. (Both the B algorithm and K2 are deterministic methods.)

A major kind of population-based method is evolutionary algorithms (EAs), and there are several methods designed for estimating the joint probability distribution $p(X)$ from data in EAs [Larranaga and Lozano, 2002]. These methods vary from the simplest ones - in which $p(X)$ is simply factorized as the product of independent univariate marginal distributions - to the most complex ones - which can construct an arbitrarily complex Bayesian network.

Finally, we mention in passing that another important kind of method for constructing Bayesian networks from data consists of using Markov Chain Monte Carlo sampling [Husmeier, 2003], [Korb and Nicholson, 2004]. However, this kind of method is not discussed here because it is not relevant to the proposed method for integrating Bayesian network construction and Simpson's paradox detection, to be discussed in Section 4.

3 On the Discovery of Interesting (Novel or Surprising) Patterns and Simpson's Paradox

There are two basic approaches to discover novel or surprising (unexpected) patterns in the context of data mining, namely the user-driven (or "subjective") approach and the data-driven (or "objective") approach [Silberchatz and Tuzhilin, 1996], [Freitas, 2006]. In essence, the user-driven approach is based on using the domain knowledge, beliefs or preferences of the user; whilst the data-driven approach is based on statistical properties of the patterns. Hence, the data-driven approach is more generic, independent of the application domain. This makes it easier to use this approach, avoiding difficult issues associated with the manual acquisition of the user's background knowledge and its transformation into a computational form suitable for a data mining algorithm. On the other hand, the user-driven approach tends to be more effective at discovering truly novel or surprising knowledge to the user, since it explicitly takes into account the user's background knowledge.

To illustrate these approaches, let us mention one simple example of each of them. The two following examples will be based on the knowledge representation of IF-THEN rules, i.e., rules of the form:

IF (a-set-of-conditions-on-some-attributes-is-true)
THEN (predict-a-certain-value-for-another-attribute)

Although this knowledge representation is quite different from Bayesian networks (the focus of this paper), IF-THEN rules are used in the following examples because the vast majority of works on the discovery of interesting patterns have focused on this kind of representation.

An example of user-driven method for discovering interesting patterns is the use of user-defined general impressions [Liu et al., 1997], [Romao et al., 2004]. In this case the user specifies general impressions in the form of IF-THEN rules, such as "IF (job_contract_length = long_term) AND (salary = high) THEN (credit = good)". Note that this is a general impression because its conditions are not precisely defined. By contrast, the data mining algorithm is supposed to produce rules with well-defined conditions, such as "job_contract_length > 4 years" or "salary > £50K". Once such rules are produced by the data mining algorithm, the system can match the rules with the general impressions, in order to find surprising rules. In particular, if a rule and a general impression have similar antecedents ("IF part") but different consequents ("THEN part"), the rule can be considered surprising, in the sense of contradicting a user's belief (general impression). For instance, the rule "IF (job_contract_length > 4 years) AND (salary >

£50k) AND (Mortgage = yes) THEN (credit = bad)" would be considered surprising with respect to the aforementioned general impression.

One kind of data-driven method consists of using a data-driven measure of rule interestingness, which assigns a numerical degree of interestingness to a rule based on some kind of statistical property of the rule. A classic example of this approach is the data-driven rule interestingness measure proposed by [Piatetsky-Shapiro, 1991], defined as Interest = $|A \cap C|$ - ($|A| \times |C|$) / N, where $|A \cap C|$ is the number of data instances (database records) satisfying both the rule antecedent A and the rule consequent C, $|A|$ and $|C|$ are the number of data instances satisfying the rule antecedent A and rule consequent C respectively, and N is the total number of data instances in the data set being mined. Hence, Interest is a measure of the deviation from statistical independence between A and C. Note that it measures the symmetric correlation between A and C, and not an asymmetric implication, i.e., Interest has the same value for the two "opposite" rules: IF A THEN C, IF C THEN A. There are more than 50 data-driven measures of rule quality that have been called rule "interestingness" measures in the literature. A review of these measures is out of the scope of this paper - the interested reader is refereed to [Hilderman and Hamilton, 2001], [Tan et al., 2002] - but it is important to point out that recent results have questioned the effectiveness of such data-driven rule interestingness measures [Ohsaki et al., 2004], [Carvalho et al., 2005]. These recent results support the intuitive argument that it is difficult to use a purely data-driven approach for discovering patterns that are truly novel or surprising to the user.

There is, however, another kind of data-driven approach for discovering surprising patterns which is not based on statistical properties of rules, but rather based on the idea of detecting occurrences of Simpson's paradox. This is the approach followed in this paper, and although it is mainly a data-driven approach - since occurrences of the paradox are extracted from the data without the need for background knowledge specified by the user - it is explicitly designed for discovering surprising patterns to users, based on the fact that instances of Simpson's paradox tend to be very surprising to users in general - almost by definition, due to the nature of the "paradox". Hence, this approach tries to combine the best of both worlds (data-driven and user-driven measures of interestingness) [McGarry, 2005].

An occurrence of Simpson's paradox can be described as follows [Pearl, 2000]. Let the event C be the apparent "cause" of an event E, the "effect". Simpson's paradox occurs if the event C increases the probability of the event E in a given population Pop and, at the same time, decreases the probability of event E in every sub-population of Pop. Let F and $\neg F$ denote two opposite values of a confounding variable, representing complementary

properties describing two sub-populations of Pop. Then, mathematically Simpson's paradox occurs if the following 3 inequalities hold for a given data set:

$$P(E|C) > P(E|\neg C), \tag{1.2}$$

$$P(E|C,F) < P(E|\neg C,F), \tag{1.3}$$

$$P(E|C,\neg F) < P(E|\neg C,\neg F), \tag{1.4}$$

where $P(X|Y)$ denotes the conditional probability of X given Y.

To illustrate these concepts, consider the hypothetical example involving Tables 1.1 and 1.2 [Pearl, 2000]. Table 1.1 shows the number of patients who recovered (E) or not ($\neg E$), given that they received a drug (C) or no drug ($\neg C$), as well as the corresponding recovery rate. Table 1.2 shows the data considering the sub-populations of males and females separately.

Table 1.1. Recovery rates for the entire population.

Combined		Recovery (E)	($\neg E$)	Total	Recovery rate
Drug	(C)	20	20	40	50%
	($\neg C$)	16	24	40	40%
	Total	36	44	80	

Table 1.2. Recovery rates for the sub-populations of males and females separately.

Males		Recovery (E)	($\neg E$)	Total	Recovery rate
Drug	(C)	18	12	30	60%
	($\neg C$)	7	3	10	70%
	Total	25	15	40	
Females		Recovery (E)	($\neg E$)	Total	Recovery rate
Drug	(C)	2	8	10	20%
	($\neg C$)	9	21	30	30%
	Total	11	29	40	

Observing Table 1.1 only we would conclude that receiving the drug *improves* the recovery rate. However, when we observe the data partitioned

by sub-population, as shown in Table 1.2, we observe a reversal of the effect of receiving the drug, because *in each sub-population* - i.e., in both the male and the female sub-populations - receiving the drug *reduces* the recovery rate.

This kind of reversal of the effect of the apparent "cause" seems paradoxical (under a causal interpretation) and tends to be very surprising to users. However, there is actually an explanation for the paradox. In the examples of Tables 1.1 and 1.2, the drug seems beneficial overall (in the entire population) due to a combination of two factors, namely males have higher recovery rates than females - both in the case of people who receive the drug and in the case of people who do not receive the drug - and more males receive the drug.

It is important to note that, strictly speaking, Simpson's paradox is not really a paradox in the context of probability theory, because it does not involve a real contradiction, just an apparent contradiction that is eliminated by providing an explanation for the occurrence of the paradox, as just discussed in the case of the paradox occurrence shown in Tables 1.1 and 1.2.

A second important and related point is that Simpson's paradox is well-known by statisticians in general. Despite these two points, it should be noted that the apparent contradiction associated with Simpson's paradox looks very puzzling (actually, it *looks like* a real contradiction) to most users, and, as pointed out by [Fabris and Freitas, 2006], in practice Simpson's paradox is usually very surprising to *users*, who typically have no formal statistical training. Hence, the discovery of Simpson's paradox instances is a valid approach for discovering surprising patterns from data, since the goal of data mining is to discover patterns that are surprising to users, rather than to statisticians or data analysts.

A number of occurrences of Simpson's paradox in real-world data are reported in [Fabris and Freitas, 1999], [Fabris and Freitas, 2006], [Kohavi, 2005]. The two works by Fabris & Freitas also describe algorithms that systematically search for occurrences of Simpson's paradox in data.

An analysis of the computational complexity of an algorithm for detecting occurrences of Simpson's paradox in data was presented in [Fabris and Freitas, 2006]. In summary, that analysis showed that the computational time taken by the algorithm: (a) grows linearly with respect to the number of data instances (records) in the data; (b) in the best (worst) case, it has a linear (cubic) growth with respect to the number of categorical (or discrete) attributes in the data - the algorithm ignores continuous (real-valued) attributes; (c) grows linearly with respect to the number of values per categorical attribute.

4 Integrating Bayesian Networks and Simpson's Paradox

So far the construction of Bayesian networks from data and the detection of Simpson's paradox in data have been considered as independent tasks in the data mining literature. Each of these two kinds of methods ignores the other, and they have been proposed to solve very different data mining problems.

This paper proposes to integrate these two kinds of methods. The basic idea of this integration is to combine the advantages of both kinds of methods, as follows. On one hand, Bayesian networks provide a graphical, easy-to-interpret representation of the structure of important relationships in the data, and their causal interpretation is potentially more useful for intelligent decision making than other knowledge representations that do not even attempt to represent causal knowledge. On the other hand, algorithms for detecting Simpson's paradox potentially discover very surprising patterns to the user. Hence, a synergistic combination of these two kinds of methods improves our chances of discovering patterns that are both potentially useful and surprising to the user, satisfying the two hardest-to-satisfy criteria mentioned in Fayyad et al.'s definition of knowledge discovery - quoted in the beginning of the Introduction.

In this paper we assume "causal sufficiency", i.e., all relevant variables involved in the underlying causal process being modeled are present in the data, so that there are no latent variables.

The remainder of this section is organized as follows. Section 4.1 discusses some limitations of Bayesian network construction algorithms, which served as the foundational ideas for the design of the proposed method for integrating Bayesian network construction algorithms and algorithms for detecting occurrences of Simpson's paradox. Section 4.2 discusses the proposed method itself.

4.1 The Framework for the Proposed Method

First of all, in general Bayesian networks are Independence-maps (I-maps) of the true probability distribution [Pearl, 1988], [Korb and Nicholson, 2004]. This means that - if the Bayesian network was correctly constructed (see below) - every independence between variables represented in the network corresponds to an actual independence in the true probability distribution, but the converse is not true, i.e., dependencies between variables represented in the network are not guaranteed to correspond to actual dependencies in the true probability distribution.

Another limitation of conventional Bayesian network learning algorithms is that such algorithms learn only up to the Markov equivalence class of

Bayesian networks [Neapolitan, 2003]. All DAGs that are members of that class are an I-map of the data, but only one of them is the truly causal DAG.

In addition to the just-mentioned "theoretical limitations" of Bayesian networks, there are two other "practical limitations" of Bayesian network construction methods. First, the problem of learning the optimal topology of a Bayesian network is NP-hard [Chickering et al., 1994], [Blanco, 2005], and the size of the search space grows exponentially with the number of variables in the data set being mined. More precisely, the number of DAGs that can be generated from a set of M variables - denoted $NumDAG(M)$ - is given by the following recursive equation [Correa, 2005]:

$$NumDAG(M) = \sum_{i=1}^{M} (-1)^{i-1} C_{M,i} 2^{i(M-1)} NumDAG(M-1), \qquad (1.5)$$

where the recursion stopping criterion is given by $NumDAG(0) = 1$ and $C_{M,i}$ is the number of combinations of i elements that can be taken from M elements.

Hence, when mining data sets with a large number of attributes - which are commonplace in the area of data mining - we usually have to accept that it will not be possible to construct the optimal network within a certain reasonably-constrained amount of time. This justifies the use of heuristic methods for constructing a good network (rather than the ideal, optimal network) within the time constraints determined by the application domain or the user, and explains the popularity of the heuristic methods reviewed in Section 2, despite the fact that they usually discover suboptimal solutions.

The second practical problem is that, even if there was a computational method that could construct the optimal Bayesian network within an acceptable amount of time in the target application domain, in practice the actual discovery of the optimal Bayesian network for the target application domain would still depend on to what extent the following assumption is satisfied: the probability distribution of the observed data (*which is just a sample of the underlying population*) is the same as the probability distribution of the population [Shipley, 2000]. In practice, the data usually has some sampling variation and/or it is noisy, making it even more difficult for a computational method to discover the optimal Bayesian network. In addition, even if there is no sample variation or noise in the observed data, there can be cases where not all independencies in the data are mirrored in the structure of the Bayesian network because, in the underlying "causal process" that produced the data, two causal paths exactly cancel each other out, thus making the learning of the Bayesian network with classical techniques impossible.

4.2 The Proposed Method

A major type of spurious dependence in a Bayesian network is the apparent dependence between events C and E (apparent cause and effect) when in reality this dependence is due to a confounding event F [Pearl, 2000]. This kind of spurious dependence can potentially be discovered by detecting occurrences of Simpson's paradox, as follows. (This assumes that the confounding event is observed in the data, of course - recall that we are assuming there are no latent variables.)

An occurrence of Simpson's paradox involving a triple of events C, E, F - whose meanings were explained in Section 3 - is evidence that we should not always believe that C is a cause of E. This is the basic idea of the method proposed here. To implement this idea we propose two approaches for integrating an algorithm for detecting Simpson's paradox (using algorithms described in [Fabris and Freitas, 2006]) and algorithms for constructing Bayesian networks (reviewed in Section 2). It should be noted that both approaches involve a "loosely-coupled" integration between the two aforementioned kinds of algorithm, in the sense that in each of these approaches first one of the algorithms is run as usual, and the results of that run are passed to the second algorithm, which is somewhat modified to take advantage of those results. The development of a more "tightly-coupled" integration is left for future research.

The first approach, here called "paradox detection before Bayesian network construction", consists of running a Simpson's paradox detection algorithm as a kind of "preprocessing step" for the Bayesian network construction algorithm, producing a list of occurrences of the paradox found in the data. This list of paradox occurrences can then be used to modify the Bayesian network construction algorithms' procedures for generating candidate networks. More precisely, consider a potential dependence of the form $C \rightarrow E$ (where C and E are thought to be cause and effect events). If C and E are associated in an occurrence of Simpson's paradox where F is a confounding event, this indicates that the effect E can be caused by the confounding event F, rather than by the apparent cause C, which suggests that the dependence $C \rightarrow E$ might be an apparent one. So, the Bayesian network construction algorithms could be modified to include, in the network being constructed, not only the edge $C \rightarrow E$, but also the edge $F \rightarrow E$. This would avoid that the constructed network have just that former edge, and not the latter, which would miss the (potentially causal) effect of F on E.

The second approach for integrating an algorithm for detecting Simpson's paradox and algorithms for constructing Bayesian networks is here called "paradox detection after Bayesian network construction". This approach

consists of using the result of a previously constructed Bayesian network to prune the search space for the Simpson's paradox detection algorithm. More precisely, the Simpson's paradox detection algorithm will focus its search on the pairs of variables for which there is a direct dependence represented by an edge from the potential cause C to the effect E in the Bayesian network. For each such pair of variables, the paradox detection method will try to find an occurrence of the paradox involving those two variables - by trying to find a third variable that acts as a confounding variable F between those two variables. If an occurrence of the paradox is found involving the two variables and a third confounding variable, intuitively this occurrence of the paradox is likely to be particularly surprising to the user. The detection of this occurrence of Simpson's paradox would be particularly interesting if there is no edge in the network pointing from the confounding F to the effect E, because in this case the paradox detection method would have detected a pattern not present in the Bayesian network.

At this point it should be noted that the proposed integration method (in both approaches) has a natural limitation. In particular, it is possible that the data to be mined does not contain any occurrence of Simpson's paradox. If this is the case, then the proposed method will have a limited usefulness. However, even in this case the application of the method can be considered to some extent useful, because, if no occurrence of Simpson's paradox was detected in the data, we would have an increased degree of confidence that the dependencies represented in the network are true (rather than spurious) dependencies, since the candidate dependencies represented in the network would have passed an additional test - i.e., no confounding variable related to the dependence was detected. This additional test complements (and not replaces) conventional methods for evaluating Bayesian networks.

4.3 Preliminary Computational Results

In this section we report preliminary computational results for the proposed method - in the approach of "paradox detection after Bayesian network construction" - in the Congressional Voting data set. This is a well-known public domain data set often used in machine learning research, available from the UCI Machine Learning Repository[1]. Each record (example) contains the votes of a United States Congressperson with respect to 16 key questions - each vote is represented by a binary attribute. In addition, each record is assigned to one out of two classes: democrat or republican. This data set is typically used for evaluating a classification algorithm, where the goal is to predict the Class (party affiliation) of a Congressperson based on

[1]University of California at Irvine, UCI Machine Learning Repository, World Wide Web address: http://www.ics.uci.edu/~mlearn/MLRepository.html

her/his votes with respect to the 16 questions (attributes). In the context of this paper, however, we are interested in constructing a Bayesian network from this data set, detecting occurrences of Simpson's paradox in the data set, and then integrate the results of these two kinds of data mining techniques.

To construct this Bayesian network (BN) we used a search procedure and a scoring metric. The scoring metric evaluates the goodness-of-fit of a candidate BN structure to the data. The search procedure generates alternative structures and selects the best one based on the scoring metric. We use a greedy search algorithm to generate BN structures. As a rule, the greedy search algorithm starts with an empty network. At each step, it then adds the edge, considering all possible pairs of nodes, that most increases the scoring metric of the current network structure. The search terminates when none of the possible edge additions improve the score of the BN. To reduce the search space of networks, only candidate networks in which each node has at most k inward edges (parents) are considered - k is a parameter determined by the user. For the experiments reported in this paper $k = 5$.

The scoring metric assigns a score to each candidate BN. Its purpose is to measure how well that BN describes the given data set. In this work we use the $K2$ scoring metric [Cooper and Herskovits, 1991; Heckerman, 1995] because its requirements exactly match our assumptions about the data set: (1) that the process that generated the database can be accurately modeled as a Bayesian network; (2) that given a Bayesian network model data instances (records) occur independently; and (3) that there are no latent variables.

The Voting data set has numerous missing values. To cope with this problem we used the following approach. When computing a given probability referring to a set of attributes X, a data instance (record) was ignored, i.e., it was not counted for probability-computation purposes, if the data set instance had a missing value for any of the attributes in the set X. This approach to cope with missing values was also used in [Fabris and Freitas, 1999], where, in the computation of probabilities associated with an occurrence of Simpson's paradox involving variables C, E and F, a data instance was ignored if it had a missing value for any of those three variables. (The results reported in [Fabris and Freitas, 1999] will be used later in this paper when analyzing the results of the constructed Bayesian network.)

Once the Bayesian network for the Voting data set has been constructed, we can ask two related questions: (a) Is there any occurrence of Simpson's paradox in this data set? (b) If the answer to (a) is "yes", is any of the paradox occurrences referring to a certain relationship between a triple of variables (C, E and F in the notation of Section 3) which is not

a relationship observed in the constructed Bayesian network? In order to answer these questions we can use some computational results about the detection of Simpson's paradox reported in [Fabris and Freitas, 1999]. In particular, that work reports 4 occurrences of Simpson's paradox in the Voting data set, so that the answer to the above question (a) is clearly "yes". The answer to question (b) is more elaborate, as follows. In all the 4 reported occurrences of the paradox, the "effect" (E) variable is the "Class" attribute. Hence, we looked at the constructed Bayesian network to identify which attributes are the parents of the Class attribute in that network. The parent attributes are "*El-Salvador-aid*", "*Anti-satellite-test-ban*", "*Aid-to-Nicaraguan-Contras*", "*MX-missile*", "*Immigration*". Out of these attributes, just one, Anti-satellite-test-ban, occurs as variable C (potential cause) in a paradox reported in [Fabris and Freitas, 1999]. Hence, our analysis here focuses on this occurrence of the paradox, as shown in Tables 1.3 and 1.4.

Looking at Table 1.3, with data combined for the entire population, it seems that Congress Members voting "yes" to Anti-satellite-test-ban are much more likely to be democrats than Congress Members voting "no" to the same question. However, looking at Table 1.4, with data partitioned into two sub-populations based on the kind of vote ("yes" or "no") to the Physician-fee-freeze question, there is a reversal of the relationship shown in Table 1.3. In Table 1.4 Congress Members voting "yes" to Anti-satellite-test-ban are *less* likely to be democrats than Congress Members voting "no" to the same question, in both sub-populations - i.e., for both values "yes" and "no" of the attribute Physician-fee-freeze.

Note that in the paradox occurrence reported in Tables 1.3 and 1.4 the potential cause variable (C) is "Anti-satellite-test-ban", the effect variable (E) is Class (party affiliation which can be Democrat or Republican), and the confounding variable (F) is "Physician-fee-freeze". Now, let us focus on two alternative causal models of the relationships between these 3 variables, as shown in Figure 1.2.

In the causal model of Figure 1.2(a) Physician-fee-freeze affects both Anti-satellite-test-ban and Class. This suggests that Table 1.4 (the sub-population-specific table) better represents the causal process underlying the data. By contrast, in the causal model of Figure 1.2(b) it is Anti-satellite-test-ban that affects both Physician-fee-freeze and Class, which suggests that Table 1.3 (the entire-population table) better represents the causal process underlying the data. This is because in Figure 1.2(b) Physician-fee-freeze is in the middle of the causal path from Anti-satellite-test-ban to Class, so we should not condition on Physician-fee-freeze when determining the effect of Anti-satellite-test-ban on Class. For an analogous and more

detailed discussion of these points in the context of the artificial data presented in Tables 1.3 and 1.4, [see Pearl, 2000, Section 6.1][2].

Table 1.3. Democrat rates for the entire population.

		Democrat			
Anti-sat-ban		(E)	$(\neg E)$	Total	Democrat rate
Vote	Yes (C)	197	39	236	83.5%
	No $(\neg C)$	55	122	177	31.1%
	Total	252	161	413	

Table 1.4. Democrat rates for sub-populations based on different values of Physician-fee-freeze.

Phys-fee-free = yes (F)					
		Democrat			
Anti-sat-ban		(E)	$(\neg E)$	Total	Democrat rate
Vote	Yes (C)	2	37	39	5.1%
	No $(\neg C)$	12	122	134	9.0%
	Total	14	159	173	
Phys-fee-free = no $(\neg F)$					
		Democrat			
Anti-sat-ban		(E)	$(\neg E)$	Total	Democrat rate
Vote	Yes (C)	195	2	197	99.0%
	No $(\neg C)$	43	0	43	100%
	Total	238	2	240	

Note that neither the causal model in Figure 1.2(a) nor the causal model in Figure 1.2(b) are represented in the Bayesian network constructed from the data. Actually, the subset of the network containing the variables Anti-satellite-test-ban, Class and Physician-fee-freeze, as well as the edge connecting these variables in the network, is shown in figure 1.3. The network in that figure suggests that Class affects Physician-fee-freeze, whilst Figures

[2]Actually, it should be noted that the causal models in Figure 1.2(a) and Figure 1.2(b) are analogous to the causal models in Figure 6.2(a) and 6.2(b) in [Pearl, 2000]. By "analogous" we mean that, once the variables Anti-satellite-test-ban, Class and Physician-fee-freeze of our Figure 1.2 are mapped into the variables C, E, F of Pearl's Figure 6.2, the network of our Figure 1.2(a) has the same structure (the same directed edges) as the network of Pearl's Figure 6.2(a), and the network of our Figure 1.2(b) has the same structure as the network of Pearl's Figure 6.2(b).

1.2(a) and 1.2(b) suggest that it is actually the Physician-fee-freeze that affects the Class.

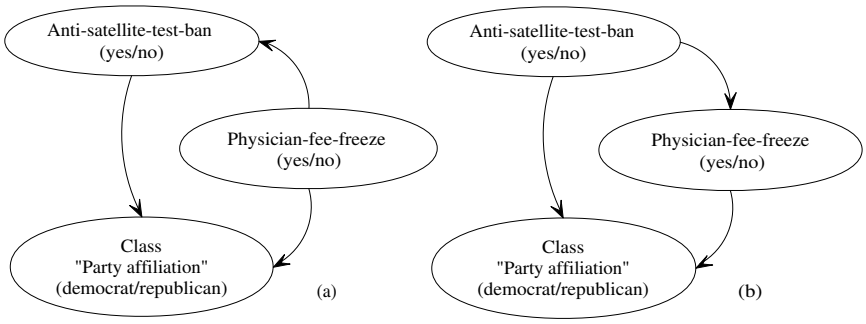

Figure 1.2. Two alternative causal models for the data in Tables 1.3 and 1.4.

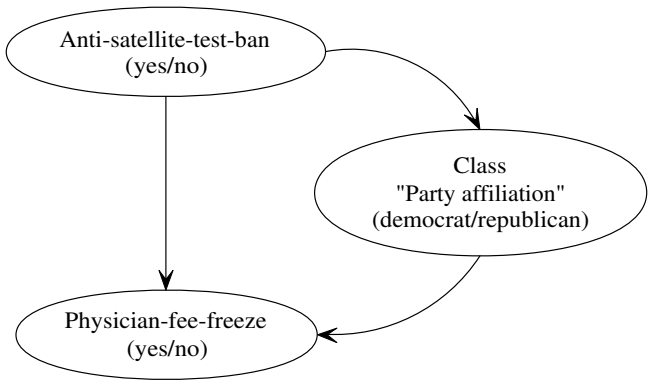

Figure 1.3. Relationship among the variables *Physician-fee-freeze*, *Anti-satellite-test-ban* and *Class* (Party affiliation) as observed in the Bayesian network constructed for the Congressional Voting data set.

It is beyond the scope of this paper to decide which causal model - Figure 1.2(a), 1.2(b) or 1.3 - better represents the causal process underlying the data. Such a decision is left to an expert in the meaning of the variables and American politics, who would make the role of the user in data mining. In particular, if it is feasible, it would be advisable to perform randomized controlled experiments to study more precisely the causal dependence between variables. Note, however, that such controlled experiments are not always feasible [Shipley, 2000]. It should be recalled that data mining is mainly used as a decision *support* technology, rather than as a decision

making technology - the actual decision is made by human beings using their expert background knowledge about the application domain.

The main point of the computational results reported in this section was just to show a kind of "proof of existence" of a situation where the detection of Simpson's paradox can discover a (potentially causal) pattern in the data that is not represented in a Bayesian network constructed from the data. Such proof of existence constitutes, of course, a very preliminary result. Much more extensive experiments will be necessary to evaluate to what extent the proposed method for integrating Simpson's paradox detection and Bayesian network construction is really useful in practice, when mining real-world data sets where a user expert in the data and the application domain will analyze the discovered patterns.

5 Conclusions

This paper proposed a method for integrating two very different kinds of algorithms, namely algorithms for constructing Bayesian networks from data and algorithms for detecting occurrences of Simpson's paradox in data. The basic idea of this integration is to combine the advantages (from a data mining perspective) of both kinds of algorithms, as follows. First, the causal interpretation of Bayesian networks is potentially more useful for intelligent decision making than other knowledge representations used in data mining - which typically do not even attempt to represent causal knowledge. Second, algorithms for detecting Simpson's paradox potentially discover very surprising patterns to the user, almost by definition - due to the nature of the "paradox".

Hence, intuitively the proposed method improves our chances of (but of course does not guarantee) discovering patterns that are both potentially useful and surprising to the user, satisfying two very demanding criteria to evaluate the quality of the patterns discovered by a data mining algorithm - an area of research clearly under-explored in the data mining literature. Since only a preliminary computational result was reported in this paper, much more extensive computational experiments and analyses of the results by users, in several real-world data sets, are required in the future.

BIBLIOGRAPHY

[Blanco, 2005] Blanco, R. (2005). *Learning Bayesian networks from data with factorization and classification purposes. Applications in Biomedicine.* PhD thesis, Dept. of Computer Science and Artificial Intelligence, University of the Basque Country, Spain.

[Brin et al., 1997] Brin, S., Motwani, R., Ullman, J., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD-97).* AAAI Press.

[Carvalho et al., 2005] Carvalho, D., Freitas, A., and Ebecken, N. (2005). Evaluating the correlation between objective rule interestingness measures and real human interest. In *Proc. 9th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD-2005)*, Lecture Notes in Artificial Intelligence 3721, pages 453–461. Springer.

[Chickering et al., 1994] Chickering, D., Geiger, D., and Heckerman, D. (1994). Learning bayesian networks is NP-hard. Technical Report MSR-TR-94-17, Microsoft Research Technical Report.

[Cooper and Herskovits, 1991] Cooper, G. F. and Herskovits, E. (1991). A Bayesian method for induction of probabilistic networks from data. Technical Report SMI-91-01, University of Pittsburgh, Pittsburgh, PA, USA.

[Correa, 2005] Correa, E. S. (2005). *Model complexity and convergence pressure in estimation of distribution algorithms*. PhD thesis, Faculty of Engineering and Physical Sciences, School of Computer Science, University of Manchester, Manchester, United Kingdom.

[Fabris and Freitas, 1999] Fabris, C. and Freitas, A. (1999). Discovering surprising patterns by detecting instances of Simpson's paradox. In *Research and Development in Intelligent Systems XVI*, pages 148–160. Springer.

[Fabris and Freitas, 2006] Fabris, C. and Freitas, A. (2006). Discovering surprising instances of Simpson's paradox in hierarchical multi-dimensional data. In *Int. J. on Data Warehousing & Mining, 2(1)*, pages 26–48.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining, 1-34*. AAAI Press.

[Freitas, 2006] Freitas, A. (2006). Are we really discovering "interesting" knowledge from data? *Expert Update Magazine*. Specialist Group on Artificial Intelligence - British Computer Society, in press.

[Heckerman, 1995] Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Technical Report MSR-TR-94-09, Microsoft Research, Redmond, WA, USA.

[Hilderman and Hamilton, 2001] Hilderman, R. and Hamilton, H. (2001). *Knowledge Discovery and Measures of Interest*. Kluwer.

[Husmeier, 2003] Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics, 19(17)*, pages 2271–2282.

[Kohavi, 2005] Kohavi, R. (2005). Focusing the mining beacon: lessons and challenges from the world of e-commerce. Invited talk at PKDD-2005. www.kohavi.com, Visited on Jan. 2006.

[Korb and Nicholson, 2004] Korb, K. and Nicholson, A. (2004). *Bayesian Artificial Intelligence*. Chapman.

[Larranaga and Lozano, 2002] Larranaga, P. and Lozano, J. (2002). *Estimation of Distribution Algorithms: a new tool for evolutionary computation*. Kluwer.

[Liu et al., 1997] Liu, B., Hsu, W., and Chen, S. (1997). Using general impressions to analyze discovered classification rules. In *Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD-97)*, pages 31–36. AAAI Press.

[McGarry, 2005] McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *Knowledge Engineering Review*, 20(1):39–61.

[Neapolitan, 2003] Neapolitan, R. E. (2003). *Learning Bayesian networks*. Prentice Hall, first edition.

[Ohsaki et al., 2004] Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., and Yamaguchi, T. (2004). Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In *Proc. 8th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD-2004)*, pages 362–373. Springer.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

[Pearl, 2000] Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge University Press.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. pages 229–248. AAAI/MIT Press.

[Romao et al., 2004] Romao, W., Freitas, A., and Gimenes, I. (2004). Discovering interesting knowledge from a science & technology database with a genetic algorithm. In *Applied Soft Computing 4*, pages 121–137.

[Shipley, 2000] Shipley, B. (2000). *Cause and Correlation in Biology: a user's guide to path analysis, structural equations and causal inference*. Cambridge University Press.

[Silberchatz and Tuzhilin, 1996] Silberchatz, S. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. In *IEEE Trans. Knowledge and Data Engineering, 8(6)*.

[Spirtes et al., 1993] Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction and Search*. Springer-Verlag.

[Tan et al., 2002] Tan, P.-N., Kumar, V., and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD-2002)*, pages 32–41. ACM Press.

[Tsumoto, 2000] Tsumoto, S. (2000). Clinical knowledge discovery in hospital information systems: two case studies. In *Proc. 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000)*, Lecture Notes in Artificial Intelligence 1910, pages 652–656. Springer.

[Wong and Leung, 2000] Wong, M. and Leung, K. (2000). *Data mining using grammar based genetic programming and applications*. Kluwer.

Alex A. Freitas
Computing Laboratory, University of Kent, Canterbury, CT2 7NF, UK.
Email: A.A.Freitas@kent.ac.uk

Ken McGarry
School of Computing & Technology, University of Sunderland, Sunderland SR6 ODD, UK.
Email: ken.mcgarry@sunderland.ac.uk

Elon Correa
Computing Laboratory, University of Kent, Canterbury, CT2 7NF, UK.
Email: E.S.Correa@kent.ac.uk