Mining Causal Association Rules

Jiuyong Li*, Thuc Duy Le*, Lin Liu*, Jixue Liu*, Zhou Jin†‡ and Bingyu Sun†
*School of Information Technology and Mathematical Sciences
University of South Australia, Mawson Lakes, SA 5095, Australia
†Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China
‡Department of Automation, University of Science and Technology, Hefei 230026, China

***Abstract*—Discovering causal relationships is the ultimate goal of many scientific explorations. Causal relationships can be identified with controlled experiments, but such experiments are often very expensive and sometimes impossible to conduct. On the other hand, the collection of observational data has increased dramatically in recent decades. Therefore it is desirable to find causal relationships from the data directly. Significant progress has been made in the field of discovering causal relationships using the Causal Bayesian Network (CBN) theory. The applications of CBNs, however, are greatly limited due to the high computational complexity. In another direction, association rule mining has been shown to be an efficient data mining means for relationship discovery. However, although causal relationships imply associations, the reverse does not always hold. In this paper we study how to use an efficient association mining approach to discover potential causal rules in observational data. We make use of the idea of retrospective cohort studies, a widely used approach in medical and social research, to detect causal association rules. In comparison with the constraint-based methods within the CBN paradigm, the proposed approach is faster and is capable of finding a cause consisting of combined variables.**

***Keywords*—causal discovery; association rules; cohort study; odds ratio**

I. Introduction

The widely collected data in various areas has enabled us to discover relationships among different variables in observational data. Associations are the mostly studied relationships and they have broad applications [14]. For example, the association of dietary habits and a certain disease identified in a large survey may be used to infer a potential cause of the disease. Basket data analysis [1] may find the associated items purchased by customers, which may help with the sales of supermarkets. However, associations do not necessarily mean causality. For instance, buying two products together (e.g. formula and nappies) does not indicate that buying one is the cause of buying the other.

Causal relationships do not only indicate that the variables are related (associated) in general, more importantly they show how the variations of one variable cause the changes of the other variable. Therefore causality is more useful for prediction and reasoning.

A commonly accepted definition of causal relationships is based on the manipulation of variable values in controlled experiments. With two binary variables A and B, if the value of A is manipulated to change, the value of B changes too, then we say that A is a cause of B. However, it is often impossible to discover causality based on the manipulation, as it is an action to force a variable to take a value, e.g. forcing a non-drinker to drink, which is different from observing someone (who has chosen to drink) drinking. There are legal, ethical and practical limitations that prevent us from manipulating a population in many studies. For example, it is unethical to ask a non-drinker to drink for the sake of studying the effects of alcohol consumption.

Then the question is how to identify causal relationships in observational data.

Significant progress has been made in the area of graphical causal modelling where causal relationships are represented with Bayesian networks [24] or similar probabilistic graphical models. The discussions of probabilistic causality started long time ago [27], [28], [12], [31]. In early 1990s, Pearl [25] and other researchers began to explore the causal semantics of a Bayesian network, a directed acyclic graph (DAG) representing the conditional independence of a set of variables. A large number of approaches have then been proposed for discovering such causal structures from data, e.g. in [23], [15], [16], [33], [22]. Though Bayesian network learning is a sound approach for causal relationship discovery, the computational costs for leaning Bayesian networks is very high [8] and the methods only handle low dimensional data sets.

Some constraint based approaches do not search for a complete Bayesian network, so they can be more efficient for casual relationship discovery. Several such algorithms have shown promising results [9], [29], [21], [26], [3]. Based on observational data, these methods determine conditional independence of variables and learn local causal structures. As these approaches are only capable of discovering the causal relationships represented with some fixed structures in a DAG, e.g. CCC [9], CCU [29] and the Y structures [21], they do not identify causal relationships that cannot be represented with these structures. Additionally these methods are not for identifying combined cause factors. However, in practice, it is possible that when individual variables each does not cause the changes of the values of a response variable, the combination of two or more variables may do.

Association rule mining [1] has proven to be an effective and versatile means for discovering relationships in data [14]. We are interested in taking the advantage of association rule mining for causality discoveries.

However, statistically reliable associations do not indicate causal relationships although causality is mostly observed as associations in data. Therefore in this paper we propose *causal association rules*. Before discussing how to discover causal association rules, we use the following example to show the motivation of our research.

Example 1. Suppose that we have generated an association

rule: "Gender = m" $\rightarrow$ "Salary = low" from a data set. The statistics of the data set is summarised as the following.

	Salary = l	Salary = h
Gender = m	185	120
Gender = f	65	60

The ratio of low salary earners to high salary earners in the male group is 1.54:1 while the ratio in the female group is 1.08:1. In other words, the odds for a male worker receiving a low salary is 1.54 and the odds for a female worker receiving a low salary is 1.08. The odds ratio of male and female groups receiving low salaries is 1.43. An odds ratio equaling to 1 indicates that an event has an equal probability to occur in both groups. An odds ratio deviating from 1 indicates that the probabilities of an event are unequal in two groups. The odds ratio of 1.43 indicates a positive association between "Gender = m" and "Salary = low".

Is this association valid? Let us do further analysis by stratifying the samples by the Education attribute. Assume that the statistics of the stratified data sets are as the following.

	Salary = l	Salary = h
Gender = m & College = y	5	20
Gender = f & College = y	15	40

and

	Salary = l	Salary = h
Gender = m & College = n	180	100
Gender = f & College = n	50	20

The above two tables indicate a negative association between "Gender = m" and "Salary = low" because the odds ratio in the College education group is 0.67 and odds ratio in the non-College education group is 0.72. Both contradict the association rule "Gender = m" $\rightarrow$ "Salary = low".

We obtain two conflicting results here. This means that an association may be volatile in a sub data set or a super data set. This is a phenomenon of the famous Simpson Paradox [24], indicating that associations may not imply causal relationships.

Causal relationships have to be analysed by taking all the variables under consideration into account [24], [30]. Such analysis is normally very difficult when the number of variables is large. On the other hand, as mentioned earlier association rule mining has been shown to be an efficient method for exploring relationships in large data sets. The question is whether we can use association rule mining to discover causal rules.

In this paper, we propose an approach which "manipulates" observational data (instead of manipulating the populations as in controlled experiments), in a similar way as a retrospective cohort study does [11], [10], and we design the test to identify causal association rules based on the manipulated data.

The contributions of this work are listed in the following.

Firstly, we integrate association rule mining with a traditional observational study method, cohort study [11], for statistically sound and computational efficient causal discovery. Cohort studies have been used for hypothesis tests in statistics for a long time, mostly with human interactions [10], but they have not been applied to large-scale data exploration for causal relationship discovery. A large number of spurious association rules is a major hurdle for association rule applications and a lot of work has been done to tackle the problem [35], [34], [32], [18]. However, causality is the ultimate goal for most applications but a causal measure has not been used for removing spurious association rules. The framework for combining cohort studies with association rule mining makes it possible to automatically generate causal hypotheses and test the hypotheses in a discovery process with large data sets.

Secondly, the proposed algorithm is capable of finding causes consisting of combined variables, which are impossible to be detected by a method in causal Bayesian network scheme [23], [16], [33], [9], [29], [3]. Note that a Bayesian network approach is possible to find two or more individual causes of an effect, such as rain causes wet road and sprinkler causes wet road too. The combined cause that we study in this paper is the interaction of two or more variables. Each individual variable is not a cause (even is not associated with the response variable), but their combination is a cause. A challenge is the exponential increase of the number of combined variables. The anti-monotone properties developed in association rule mining can handle this problem well.

II. Problem definitions

A. Notations and definitions

Let us consider a data set, D, for a set of binary variables $(X_1, X_2, \ldots, X_m, Z)$, where $X_1, X_2, \ldots, X_m$ are *predictive* variables and Z is a *response* variable. Values of Z are of user's interest, such as having a disease or being normal. Note that using a binary data set does not lose the generality of a data set that contains attributes of multiple discrete values. For example, a multi-valued data set (Gender, Age, ...) is equivalent to a binary data set (Male, Female, 0-19, 20-39, 40-69, ...). Both the Male and Female variables are kept in this case since this allows us to have combined variables that involve them separately, for example, (Female, 40-59, Diabetes) and (Male, 40-59, Smoking). A binary data set makes the conceptual discussions in the paper easier.

Let P be a *combined* variable consisting of multiple variables $X_1, \ldots, X_n$ where $n \geq 1$, and $P = 1$ when $(X_1 = 1, \ldots, X_n = 1)$ and $P = 0$ otherwise. A rule is in the form of $(P = 1) \rightarrow (Z = 1)$, or $p \rightarrow z$ where z stands for $Z = 1$ and p for $P = 1$. Our ultimate goal is to final out whether $p \rightarrow z$ is a causal rule.

With our approach, we first consider the association between P and Z since an association is necessary for a causal relationship. Odds ratio is a widely used measure for associations in retrospective studies [11], and we define the odds ratio of a rule as follows.

Definition 1 (Odds ratio of a rule) *The contingency table of a rule, $p \rightarrow z$, is listed as the following, where* $\mathrm{supp}(x)$ *indicates the support of pattern X, the count of value x in the given data set, D.*

	$z(Z=1)$	$\neg z(Z=0)$
$p(P=1)$	$\text{supp}(pz)$	$\text{supp}(p\neg z)$
$\neg p(P=0)$	$\text{supp}(\neg pz)$	$\text{supp}(\neg p\neg z)$

We have $\text{supp}(p) = \text{supp}(pz) + \text{supp}(p\neg z)$, $\text{supp}(z) = \text{supp}(pz) + \text{supp}(\neg pz)$, *and* $\text{supp}(pz) + \text{supp}(p\neg z) + \text{supp}(\neg pz) + \text{supp}(\neg p\neg z) = n$, *where* n *is the number of records in the data set.*

The odds ratio of the rule $p \to z$ *on data set* D *is defined as the following.*

$$\text{oddsratio}_{\text{D}}(p \to z) = \frac{\text{supp}(pz) * \text{supp}(\neg p\neg z)}{\text{supp}(p\neg z) * \text{supp}(\neg pz)}$$

This is the ratio of the odds of value z occurring in group $P = 1$ to the odds of value z occurring in group $P = 0$, so an odds ratio of 1 means that z has an equal chance to occur in both groups, and an odds ratio deviating from 1 indicates an association (positive or negative) between Z and P.

Definition 2 (Association rule) *Using the notations in Definition 1, the support of a rule* $p \to z$ *is defined as* $\text{supp}(p \to z) = \text{supp}(pz)$. *Given a data set D, let min_supp and min_oratio be the minimum support and odds ratio respectively,* $p \to z$ *is an association rule if* $\text{supp}(p \to z) > min\_supp$ *and* $\text{oddsratio}_{\text{D}}(p \to z) > min\_oratio$, *and LHS*$(p \to z) = p$ *and RHS*$(p \to z) = z$.

Traditional association rules are defined by support and confidence [1]. An association rule in the support and confidence scheme may not show a real association between the LHS and the RHS of a rule [7]. In the above definition, an association rule indicates an association between its LHS and the RHS since a high odds ratio indicates a real association. In our algorithm, the minimum odds ratio will be replaced by a significance test that $\text{oddsratio}_{\text{D}}(p \to z) > 1$. (See Section III-B for details.)

In the definition, we consider z as the RHS of a rule. An association rule that has $\neg z$ ($Z = 0$) as its RHS can be defined in the same way. These association rules ($p \to z$ and $p \to \neg z$) are class association rules [20] where the confidence ($\text{prob}(z|p)$) is replaced by the odds ratio.

We note that the distribution of the values of the response variable can be skewed and a uniform minimum support may lead to too many rules for the frequent values and few rules for the infrequent values. In the implementation, we use the local support that is relative to the frequency of a value in the response variable, i.e. $lsupp(p \to z) = \frac{\text{supp}(pz)}{supp(z)}$. The local support is a ratio and can be set the same, say 5%, for rules that have z or $\neg z$ as the RHS.

B. Causal association rules

As Example 1 shows, associations may not indicate causal relationships, therefore our idea is to conduct a retrospective cohort study to detect true causal relationships from identified association rules.

1) Cohort study: In medical and social research, when randomised controlled trials are practically impossible, a cohort study is often used to infer risk factors [11], [10]. A cohort study is a type of observational studies. It follows two groups of individuals, called cohorts, who share common characteristics but differ with respect to a certain factor of interest, to determine how the factor causes an outcome. It normally gives highly generalisable results. There are two types of cohort studies: prospective and retrospective cohort studies. In a perspective cohort study, researchers follow cohorts over time to observe their development of a certain outcome. In the retrospective study, researchers look back at events that already occurred.

A retrospective cohort study selects individuals who have exposed and have not exposed to a suspected risk factor but are alike within many other aspects. For example, middle aged male labours who have been smoking and who have not been smoking for a certain time period are selected for studying the effect of smoking on lung cancer. Here smoking is the risk factor or *exposure variable*. Middle aged males indicate the common characteristics shared by the cohorts. A significant difference in the value of the outcome or response variable (say having lung cancer or not) of the two cohorts indicates a possible causal relationship between the exposure variable and the response variable.

In the rest of the paper, with a binary exposure variable, we call the cohort where the exposure variable takes value 1 the *exposure group*, the cohort where the exposure variable takes value 0 the *non-exposure group*, and the set of variables determining the common characteristics of the two groups the *controlled variable set*.

From the above description, the core requirement for a cohort study is that the distribution of controlled variable set of the two groups should be same (or very similar). For example, in a cohort study to test whether gender is a cause of salary difference, the exposure variable is gender. The controlled variable set consists of variables: Education, Profession, Experience and Location. From a given data set, we will need to select samples for the exposure and non-exposure groups so that the two groups have the same distribution regarding the controlled variables. Then if there is a significant difference in salary between the two groups, we can conclude that gender is a cause of salary difference.

2) Identifying causal association rules: Given an association rule as a hypothesis that its LHS causes its RHS. The variable for the LHS is an exposure variable and the variable for the RHS is the response variable. Let all other variables be included the controlled variable set initially. We will discuss how to refine this controlled variable set in the next section.

Given a data set D, for an exposure variable, we use the following process to select samples for the exposure and non-exposure groups (while the RHS outcome is blinded). We firstly pick up a record t_i containing the LHS factor ($P = 1$), and then pick up another record t_j of which $P = 0$, and both t_i and t_j have the same values for all the controlled variables. Then t_i is added to the exposure group, t_j is added to the non-exposure group, and both are removed from the original data set. This process repeats until there are no matched pairs can be found. As a result, the distributions of the controlled variables

in the exposure and non-exposure groups are identical.

We formulate the above discussions as the following.

Definition 3 (Matched record pair) *Given an association rule $p \rightarrow z$ and a set of controlled variables C, a pair of records match if one contains value p, the other does not, and both have the same value for C.*

For example, assume that $C = (A, B, D)$ is the controlled variable set for association rule $p \rightarrow z$, then records $(P = 1, A = 1, B = 0, D = 1)$ and $(P = 0, A = 1, B = 0, D = 1)$ form a matched pair.

Definition 4 (Fair data set for a rule) *Given an association rule $p \rightarrow z$ that has been identified from a data set D and a set of controlled variables C, the fair data set D_f for the rule is the maximum sub data set of D that contains only matched record pairs from D.*

In the above definition, the requirement of the maximum sub data set of D is for the best utilisation of the data set. However, even with this requirement, the number of records in a fair data set may not be sufficiently large, because in order to obtain statistically significant results based on a fair data set, a minimum number of matched pairs is required to be contained in the set. More details on this will be provided in Section III-B2.

Example 2. Given an association rule $a \rightarrow z$ identified using the following data set, and the controlled variable set $C = (M, F, H, B, P)$, where M stands for Male, F for Female, H for High school graduate, B for University graduate, and P for postgraduate.

ID	A	M	F	H	B	P	Z
1	1	0	1	0	0	1	1
2	1	0	1	0	1	0	1
3	1	1	0	1	0	0	0
4	1	1	0	0	0	1	1
5	0	0	1	0	0	1	0
6	0	0	1	0	1	0	0
7	0	1	0	1	0	0	0
8	0	1	0	1	0	0	1

Records (#1, #5), (#2, #6) and (#3, #7) form three matched pairs. A fair data set for $a \rightarrow z$ includes records (#1, #2, #3 #5, #6, #7).

Matches in a data set are not unique. A record possibly matches more than one record, but we only choose one. For example, (#3, #7) and (#3, #8) both are matched pairs (in terms of record #3). When there are two or more possible matches, a matched pair is selected randomly without knowing z. In the experiments, we show that such random selection will cause variance in the results (different causal rules validated in different runs), but the variance is very small in a large set (one rule difference in three runs). Even in a small data set, more than 80% rules are consistent over different runs. We pick frequently supported rules in multiple runs to reduce the variance.

Since with a fair data set for a rule the exposure and non-exposure groups are identical except for the value of the exposure variable, if there is a significant difference in the values of the response value between the two groups, it is reasonable to assumes that the difference of the outcome is caused by the difference of the values of the exposure variable.

Now, we discuss how to detect the statistical difference of the values of the response variable between the exposure and non-exposure groups, which will provide us the method for testing whether an association rule is a causal rule or not.

When the values of the response variable are taken into consideration, there are four combinations for a matched pair: both records containing z, neither containing z, record $(P = 1)$ containing z and record $(P = 0)$ not; record $(P = 0)$ containing z and record $(P = 1)$ not. The counts of the four different types of matched pairs in the fair data set for rule $p \rightarrow z$ is listed as the following.

	$P = 0$	
$P = 1$	z	$\neg z$
z	n_{11}	n_{12}
$\neg z$	n_{21}	n_{22}

In this table n_{11} is the number of matched pairs containing z in both the exposure and non-exposure group; n_{12} the number of matched pairs containing z in the exposure group and $\neg z$ in the non-exposure group; n_{21} the number of matched pairs containing $\neg z$ in the exposure group and z in the non-exposure group; and n_{22} the number of matched pairs containing $\neg z$ in both the exposure and the non-exposure group. In Example 2, $n_{11} = 1$, $n_{21} = 1$, $n_{12} = 0$, and $n_{22} = 0$. In our experiments, we replace zero count by 1 to avoid infinite odds ratios.

Using the above notation, we can have the following definition:

Definition 5 (Odds ratio of a rule on its fair data set) *The odds ratio of an association rule $p \rightarrow z$ on its fair data set D_f is:*

$$\text{oddsratio}_{D_f}(p \rightarrow z) = \frac{n_{12}}{n_{21}}$$

This leads to the definition of a causal association rule:

Definition 6 (Causal association rule) *An association rule $(p \rightarrow z)$ indicates a causal relationship between P and Z (the variables for its LHS and RHS) and thus is called a causal association rule, if its odds ratio on its fair data set, $\text{oddsratio}_{D_f}(p \rightarrow z)$, is significantly greater than 1.*

We will discuss how to test if $\text{oddsratio}_{D_f}(p \rightarrow z)$ is significantly greater than 1 in Section III-B2.

Based on Definition 6, testing if an association rule is a causal rule becomes the problem of finding the fair data set for the rule. A fair data set simulates the controlled environment for testing the causal hypothesis represented by an association rule. When the odds ratio of an association rule on its fair data set is significantly greater than 1, it means that a change of the response variable is resulted from the change of the exposure

variable since in the fair data set, all controlled variables are balanced between the two groups.

3) Selecting controlled variable set: Let X represent the set of all predictive variables, and as before P is the exposure variable and C is a set of controlled variables. Initially, let $C = X \backslash P$.

The set of controlled variables determines the size of a fair data set. If the size of the controlled variable set is large, the chance of finding a non-empty fair data set is small. Therefore we need to find a proper controlled variable set, without compromising the quality of the causal discovery. In the following we discuss how to obtain such a controlled variable set.

Definition 7 (Relevant and irrelevant variables) *If a variable is associated with the response variable, it is relevant. Otherwise, it is irrelevant.*

We do not control irrelevant variables, hence $C = X \backslash (P, I)$ where I stands for a set of irrelevant variables. The major purpose for controlling is to eliminate the effects of other possible causal factors on the response variable. Other variables that are randomised with respect to the value of the response variable can be considered as noises and need not to be controlled. With Example 1, when we test the association rule "Gender = m" $\rightarrow$ "Salary = low" for finding a causal relationship, we should control variables like education, location, profession and working experience. However, we do not control variables like blood type and eye colour, since they are irrelevant to salary.

Definition 8 (Exclusive variables) *Variables P and Q are mutually exclusive if* $\text{supp}(pq) \leq \epsilon$ *or* $\text{supp}(\neg pq) \leq \epsilon$ *where ϵ is a small integer.*

We do not control an exclusive variable of the exposure variable P, i.e. we let $C = X \backslash (P, I, Q)$ where Q stands for a set of exclusive variables of P. Let us have an operational explanation firstly. Assume that $\epsilon = 0$, if Q is in the controlled variable set, the non-exposure group or the exposure group will be empty since there is no record containing $(P = 0, Q = 1)$ or $(P = 1, Q = 1)$. In other words, we are unable to do a cohort study due to the exclusiveness of variables P and Q.

One reason Q being exclusive to P is the data set constraint. An example is when P is for high school graduate and Q is for university graduate. They both belong to the same domain in a relational data set. An individual has only one highest education qualification. Therefore, P and Q are exclusive. Semantically, Q should not be a controlled variable of P.

Another reason for Q being exclusive to P is the negative (or positive) association of P and Q. This is a complicated issue in causality discovery. Let us explore possible causal relationships between P, Q, and Z (response variable) when we observe that both P and Q are relevant to Z, and P and Q are associated. Assume that Z is not a cause of other variables and there are no unobserved variables. The causal relationships between P, Q and Z can be (1) $Q \rightarrow P \rightarrow Z$; (2) $P \rightarrow Q \rightarrow Z$; (3) $(P \rightarrow Z) \wedge (Q \rightarrow Z)$; (4) $(P \rightarrow Q) \wedge (P \rightarrow Z)$; and (5) $(Q \rightarrow P) \wedge (Q \rightarrow Z)$. In Case (1) P is a direct cause of Z. In Case (2) P is an indirect cause of Z intermediated by Q. In Case (3) both P and Q are direct causes of Z. In Case (4) P causes both P and Z. In Case (5) Q causes both P and Z. If we do not control Q, it will lead to false positives like in Cases (2) and (5). However, if we control Q, it will lead to false negatives like in Cases (1), (3) and (4).

Leaving an exclusive variable uncontrolled may lead to false discoveries, and controlling an exclusive variable will lead to the miss of true discoveries. We trade some false discoveries for more true discoveries. This is a limitation of the proposed method. However, in the case that P and Q are associated and they both are associated with Z, P and Q are potential confounding variables. Other causality discovery methods make an assumption that there are not confounding variables [9], [29]. When this assumption is violated, they produce false discoveries too.

The combination of multiple irrelevant variables can be relevant. However, we do not consider combined variables in controlled variable set. There will be many combined relevant variables and the support of combined variables are normally small. When they are included in the controlled variable set, the chance for having non-empty exposure and non-exposure groups is very small.

III. Algorithm

In this section we present the algorithm (Algorithm 1) for causal association rule mining. The algorithm integrates association rule mining with causal rule detection based on fair data sets, as introduced in the previous section. In the following, we firstly discuss the two anti-monotone properties for efficient causal association rule identification, and then we introduce the details of detecting causal rules from identified association rules.

A. Anti-monotone properties

Anti-monotone properties are the core for efficient association rule mining. For example a well known anti-monotone property is that a superset of an infrequent pattern is infrequent, and infrequent patterns are pruned before they are generated (called forward pruning). We firstly discuss the anti-monotone properties we will apply for causal association rule discovery.

In the following discussions, we say that rule $px \rightarrow z$ is more specific than rule $p \rightarrow z$, or $p \rightarrow z$ is more general than $px \rightarrow z$. Furthermore, we use $\text{cov}(p)$ to represent the set of records in D containing value p, and we call $\text{cov}(p)$ the covering set of p. A rule is *redundant* if it is implied by one of its more general rules. For example, if the more general rule is a causal rule, the rule is a causal rule. If the more general rule is not, the rule is not. A k-pattern is a pattern containing k values.

Observation 1 (Anti-monotone property 1) *All more specific rules of a causal rule are redundant.*

Proof: This observation is based on the persistence property of a real causal relationship. Persistence means that a causal relationship holds in any condition. When a rule is

specified, additional conditions are added to the LHS of the rule, and the conditions do not change the causal relationship. The more specific rules are implied by the general rule, and hence are redundant. ■

For example, if rule "college graduate → high salary" holds, then we know that both male college graduates and female college graduates enjoy high salaries. It is therefore redundant to have the rules "male college graduate → high salary" and "female college graduate → high salary".

Observation 2 (Anti-monotone property 2) *If* $\text{supp}(px) = \text{supp}(p)$*, rule* $px \rightarrow z$ *and all more specific rules of* $px \rightarrow z$ *are redundant.*

Proof: If $\text{supp}(px) = \text{supp}(p)$, then $\text{cov}(px) = cov(p)$. In other words, both rules $p \rightarrow z$ and $px \rightarrow z$ cover the same set of records. There will be the same fair data set for both rules. Therefore, if $p \rightarrow z$ is a causal rule, so it $px \rightarrow z$. If $p \rightarrow z$ is not a causal rule, nor is $px \rightarrow z$. Rule $px \rightarrow z$ is redundant.

Let rule $pxy \rightarrow z$ be a more specific rule of rule $px \rightarrow z$. If $\text{supp}(px) = \text{supp}(p)$, then $\text{supp}(pxy) = \text{supp}(py)$. Using the same reasoning above, we conclude that rule $pxy \rightarrow z$ is redundant with respect to rule $px \rightarrow z$. ■

Since there are two anti-monotone properties in addition to the anti-monotone property of support, it is efficient to use a level wise algorithm like Apriori [2]. Both anti-monotone properties 1 and 2 can be used in the same way as the anti-monotone property of support. We make use of a prefix tree structure for rule generation and pruning as in [6]. Due to page limit, we omit the discussions of association rule generation part including forward pruning by Observations 1 and 2.

Algorithm 1 Causal Association Rule discovery (CAR)
Input: Data set D, the minimal local support δ, the maximum length of rules k_0, and a z value for significance test
Output: A set of causal association rules

let causal association rule set $R_C = \emptyset$
add 1-pattern to a prefix tree T as the 1-st level nodes
count support of the 1-st level nodes with and without z
remove nodes whose local support is no more than δ
Let X be the set of attributes containing frequent 1-patterns
find the set of irrelevant attributes I
let $k = 1$
while $k \leq k_0$ **do**
  generate association rules at the k-th level of T
  for each generated rule r_i **do**
    find exclusive variables E of $LHS(r_i)$
    let controlled variable set $C = X \backslash (I, E, LHS(r_i))$
    find the fair data set for r_i
    if $\text{oddsratio}_{D_f}(r_i) > 1$ significantly **then**
      move r_i to R_C
      remove $LHS(r_i)$ from the k-th level of T
    end if
  end for
  $k = k + 1$
  generate k-th level nodes of T
  count the support of the k-th level nodes with and without z
  remove nodes whose local support is no more than δ
  remove nodes of patterns whose supports are the same as those of their sub-patterns respectively
end while
output R_C

B. Detecting causal rules

This process involves three steps, as discussed below.

1) Determining controlled variables: We firstly determine the set of irrelevant variables, each of which is not associated with the response variable. For a variable Y, its association with the response variable Z can be determined by the odds ratio of $y \rightarrow z$.

Let ω be the odds ratio of the rule $p \rightarrow z$ on the given data set D, i.e. $\text{oddsratio}_{\text{D}}(p \rightarrow z) = \omega$ The confidence interval of ω is defined as $\exp(\ln \omega \pm z' \sqrt{\frac{1}{\text{supp}(pz)} + \frac{1}{\text{supp}(p\neg z)} + \frac{1}{\text{supp}(\neg pz)} + \frac{1}{\text{supp}(\neg p\neg z)}}) = [\omega_-, \omega_+]$, where z' is a standard normal deviate corresponding to the desired level of confidence ($z' = 1.96$ for 95% confidence). ω_- and ω_+ are the lower and upper bounds respectively of an odds ratio at a confidence level. If $\omega_- > 1$, the odds ratio is significantly higher than 1, hence P and Z are associated. Equivalently, $p \rightarrow z$ is an association rule. Therefore, we do not use the minimum odds ratio in the algorithm.

Another advantage of the above process is that it is automatically adaptive to the size of a data set. For a large data set, the confidence interval of an odds ratio is small and hence a small odds ratio can be significantly higher than 1. For a small data set, the confidence interval of an odds ratio is large and hence a large odds ratio is needed to be significantly higher than 1.

The above identified irrelevant variables are excluded from the controlled variable set, so are attributes with infrequent non-zero values.

Secondly we identify the exclusive variables of an exposure variable, say P, according to Definition 8 where ϵ is set to the same value as the minimum local support. We exclude the identified exclusive variables from the controlled variable set.

The remaining variables then form the controlled variable set. The controlled variable set can be viewed as multiple patterns in association rule mining. For example, if male, female, college and postgraduate form the controlled variable set, the set includes the patterns {(male, college), (male, postgraduate), (female, college), (female, postgraduate)}.

2) Creating fair data set: We select the samples from the given data set D to get the fair data set for rule $p \rightarrow z$, following the procedure listed in Function 1. We firstly find the covering set of c. Then the covering set of c is split into two subsets: one containing value p, denoted by D_{cp}, and the other containing value $\neg p$ (or $P = 0$), denoted by $D_{c\neg p}$. Assume that $|D_{cp}| \leq |D_{c\neg p}|$ (if not, we swap the order in the following description). For each record in D_{cp}, find an identical record in $D_{c\neg p}$ in terms of the controlled variable set. If there are more

Function 1 Sample a fair data set for rule $p \rightarrow z$
Input: Data set D, rule $p \rightarrow z$, and controlled variable set C
Output: a fair data set for rule $p \rightarrow z$, D_f

find the covering set of $c(C = 1)$, D_c
split D_c into D_{cp} and $D_{c\neg p}$ // D_{cp} contains value p and $D_{c\neg p}$ does not
let $D_f = \emptyset$
for each record t_i in D_{cp} // assuming $|D_{cp}| \leq |D_{c\neg p}|$. If not, swap D_{cp} and $D_{c\neg p}$. **do**
  for each record t_j in $D_{c\neg p}$ **do**
    if t_i and t_j have the identical value of C **then**
      move t_i and t_j to D_f
    end if
  end for
end for
output D_f

than one such records, choose one randomly. Add the pair of records to the fair data set. If there is no identical record in $D_{c\neg p}$, move to the next record.

3) Testing causal rules: To check if an association rule is a causal association rule, we use the following means to test the significance of the odds ratio of a rule on its fair data set. Let $\text{oddsratio}_{D_p}(p \rightarrow z) = \omega'$ in the fair data set, the confidence interval of the odds ratio is defined as $\exp(\ln \omega' \pm z' \sqrt{\frac{1}{n_{12}} + \frac{1}{n_{21}}}) = [\omega'_-, \omega'_+]$ where z' is a standard normal deviate corresponding to the desired level of confidence ($z' = 1.96$ for 95% confidence) and ω'_- is the lower bound of $\text{oddsratio}_{\text{D}_f}(p \rightarrow z)$ in the confidence level. If $\omega'_- > 1$, the odds ratio is significantly higher than 1, then we conclude that P is a cause of Z.

IV. Experiments

A. Data and parameters

A number of frequently used public data sets [4] are employed in our experiments, to test the effectiveness and efficiency of the proposed causal rule discovery method. A summary of the data sets is given in Table I. All variables have the values of 1 or 0, indicating the presence or absence of an attribute value correspondingly.

Hypothyroid and Sick are two medical data sets included in the Thyroid Disease folder of the UCI repository [4] and they are discretised by MLC++ discretise utility [17]. The Adult data set is an extraction of the USA census database in 1994.

We also use a large data set, Census Income, to test the scalability of our method with the size of data. We obtain $250K$ records by combining the training and test data sets. Continuous attributes have been removed. The Harvard Lung Cancer data set is used to test the scalability of our method with the number of attributes.

The Harvard Lung Cancer data set is a microarray data from [5]. The original data set contains 11657 genes. The top 89 genes used in our experiments are obtained by feature selection with the information gain ratio implementation of Weka [13]. The gene expression values are discretised by using the medium as the cut point, and the values are categorised as "up" and "down".

TABLE I: A brief description of data sets used in experiments

Name	#Records	#Variables	Distributions
Hypothyroid	3163	51	4.8% & 95.2%
Sick	2800	58	6.1% & 93.9%
Adult	48842	99	23.9% & 76.1%
Census income	250000	495	6.2% & 93.8%
Harvard	156	178	89.1 % & 10.9%

TABLE II: Comparison of the numbers of association rules (AR), non-redundant rules (NRR), optimal rules (OR) and causal association rules (CAR)

	#AR	#NRR	#OR	#CAR
Sick	33771	19506	2611	18
Hypothyroid	16939	7690	2422	22
Adult	3285	3017	1748	49
Census Income	77780	46796	13687	41

In the experiments, the class attributes in the original data sets are set as response variables, and values such as high/low incomes and sick/normal are used as the RHS of a rule. The default minimum local support is set to 0.05. We set the local support to 0.01 for the Adult set in the comparison of its results with those of the CCC [9] and CCU [29] methods. The minimum local support for the Harvard Lung Cancer data set is set to 0.35 since the size of the data set is small.

B. Causal association rules vs. association rules

An association does not mean a causal relationship. In fact, the majority of association rules are not causal rules. We compare causal association rules with various types of association rules in Table II. The number of causal association rules is significantly smaller than the number of other types of rules, including association rules [2], non-redundant rules [36], and optimal rules [19]. All these rules satisfy the same minimum local support. Associations are measured by the odds ratio with the same significance test as discussed in Section III-B. The maximum length of the rules is 4.

The number of causal rules is very small. They may not be enough for classification since not every record in the data is covered by a causal association rule. However, they are reliable relationships since each causal rule is tested by the cohort study in data.

Most discovered causal rules (99%) are short and include one or two variables. This makes the rules easily interpreted, and also supports the application of association rule mining to solve real world problems, where only short rules are considered and used.

C. Causal association rules vs. constraint causal structures

We compare the causal association rules discovered using our method (called CAR in the following) with the causal relationships discovered by the constraint based methods, CCC [9] and CCU [29].

The most widely used methods for causal discovery are based on graphical causal modeling [23], [15], [24]. Based

TABLE III: Comparison of the numbers of causal association rules, CCC rules, and CCU rules

	CAR	CCC	CCU
Adult	49	53	46
Sick	18	13	3

on causal sufficiency and causal faithfulness assumptions, an edge of a Bayesian network is interpreted as a causal relationship [24]. However, learning Bayesian networks is computational challenging. Constraint based approaches have been proposed to learn causal structures directly without learning a Bayesian network. CCC [9] and CCU [29] are two efficient implementations. Both methods learn triplex structures involving three variables with certain dependency and independency relationships among them, and infer causal relationships from the structures. Both assume that there are not hidden and confounding variables in data sets.

The numbers of rules (relationships) discovered by CAR, CCC and CCU are listed in Table III. CCC and CCU are constrained to the Salary and Sick attributes only. When a statistical significance test is involved, 95% confidence level is used. Since there are small variations between causal association rules discovered in different runs, due to the random selection of matching pairs when a record has multiple matches, in the experiments, we generated causal rules three times and chose the rules occurring twice in the three runs.

When only looking at the number of rules produced from the Adult data set, they are very similar. However, when we look into the rules, they are quite different. We list the most similar and dissimilar rule groups on the Adult data from the three methods in Table IV.

Rules discovered by CAR, CCC and CCU are similar for the variables related to the attributes Education and Workclass. They are the major factors affecting incomes. We see that people with higher education have a better chance for a high salary, such as, doctorate, masters, bachelors, and professional school (prof-School). In contrast, people with lower education more likely receive a low salary, for example high school graduate (HS-grad) and lower. The effects of Workclass on salary are also intuitively right. Rules discovered by the three methods on the variables are consistent.

Rules discovered by CAR, CCC and CCU are dissimilar in variables related to the attributes Occupation and Native-country. There are 12 rules discovered by CAR for variables related to the attribute Occupation, only one rule is discovered by CCC and CCU. CCC and CCU have missed some very reasonable causal factors for high/low salary. For example, "exec-managerial" and "prof-specialty" for high salary, and "handlers-cleansers" and "adm-clerical" for low salary are reasonable causal rules in the Occupation attribute, but they have been missed by CCC and CCU. On the other hand, there are 22 rules for variables related to the attribute Native Country discovered by CCC, 17 rules by CCU and only 2 rules by CAR. Intuitively, Native Country is not a factor for high/low salary. This shows that the CAR method is able to discover reasonable causal rules.

Another advantage of CAR is that it produces causal rules of combined variables while CCC and CCA do not. Rules discovered in the Sick data set with variables related to the Age attribute is shown in Table V. It is clear that old age is a factor for disease and young age is a factor for negative, and the three methods have the similar discoveries for the causal relationship. However, for the age in between, the findings are different by the three methods. CCC and CCA do not find any rules related to the age in between. CAR considers the combined variables and provides some insights for the age in between cases. CAR is able to provide more insights in data than CCC and CCU.

We should be aware that CAR, CCC and CCU do their work under certain assumptions, which are most likely violated. For example CCC and CCU assume that there are no hidden variables and no confounding variables. CAR assumes that all other causal factors are controlled. They will be violated in real world applications and we should be aware of their limitations.

TABLE IV: The most similar and dissimilar causal rule groups discovered by CAR and CCC and CCU in the Adult data set

Causal rules	CAR	CCC	CCU
Education=doctorate $\rightarrow > 50$K	√	√	√
Education=masters $\rightarrow > 50$K	√	√	√
Education=bachelors $\rightarrow > 50$K	√	√	√
Education=prof-School $\rightarrow > 50$K	√	√	√
Education=some-college $\rightarrow \leq 50$K	√	√	√
Education=HS-grad $\rightarrow \leq 50$K	√	√	√
Education=12th $\rightarrow \leq 50$K	√	√	√
Education=11th $\rightarrow \leq 50$K	√	√	√
Education=10th $\rightarrow \leq 50$K	√	√	√
Education=9th $\rightarrow \leq 50$K	√	√	√
Education=7-8th $\rightarrow \leq 50$K	√	√	√
Education=5-6th $\rightarrow \leq 50$K	√	√	√
Education=1-4th $\rightarrow \leq 50$K		√	√
Education=preschool $\rightarrow \leq 50$K		√	
Occupation=exec-managerial $\rightarrow > 50$K	√		
Occupation=prof-specialty $\rightarrow > 50$K	√		
Occupation=protective serv $\rightarrow > 50$K	√		
Occupation=tech-support $\rightarrow > 50$K	√	√	√
Occupation=sales $\rightarrow > 50$K	√		
Occupation=handlers-cleaners $\rightarrow \leq 50$K	√		
Occupation=machine-op-inspct $\rightarrow \leq 50$K	√		
Occupation=adm-clerical $\rightarrow \leq 50$K	√		
Occupation=other-service $\rightarrow \leq 50$K	√		
Occupation=farming-fishing $\rightarrow \leq 50$K	√		
Occupation=transport-moving $\rightarrow \leq 50$K	√		
Occupation=craft-repair $\rightarrow \leq 50$K	√		
Workclass=sel-emp-inc $\rightarrow > 50$K	√	√	
Workclass=sel-emp-not-inc $\rightarrow > 50$K	√	√	√
Workclass=federal-gov $\rightarrow > 50$K	√	√	√
Workclass=state-gov $\rightarrow > 50$K	√		
Workclass=local-gov $\rightarrow > 50$K	√	√	√
Workclass=private $\rightarrow \leq 50$K	√	√	√
Native Country=USA > 50K	√	√	√
Native Country=various countries	1	22	17

TABLE V: Causal rules discovered by CAR, CCC and CCU related to the Age attribute in the Sick data set

Causal rules	CAR	CCC	CCU
Age$>$71.5 $\rightarrow$ disease	√	√	
Age $\leq$ 43.5 $\rightarrow$ negative	√	√	√
Age$\in$(43.5,71.5] & T4U $<$ 0.895 $\rightarrow$ disease	√		
Age$\in$(43.5,71.5] & On Thyroxine = false & TSH measured = true $\rightarrow$ disease	√		

TABLE VI: The numbers of causal rules of different runs and the frequent causal rules in two results.

fair data set	1	2	3	frequent
Adult	49	48	49	49
Sick	17	21	19	18

D. Stability

The selection of the fair data set is subject to selection bias. Normally, the data distribution is skewed for two conditions: exposure and non-exposure. Usually there are significantly more exposed cases than non-exposed cases. When we choose pairs of matched records to form a fair data set, we pick up one record from the exposure group and find a matched record from the non-exposure group. In this process, the values of the response variable are blinded. When there are more than one matched record to choose from, we randomly choose one. It is possible that the value distribution of the response variable in a fair data set is affected by the random selection. This will cause misses or false discoveries of causal rules. This situation is the same as the real world sample process, which is subjected to sampling bias.

To reduce the impact of selection bias, we run the method on a data set multiple times and select consistent rules in multiple causal rule sets as the final causal rules. The variance is not big and the casual discovery is quite stable. The numbers of causal rules from different runs and the rules supported by two causal rule sets are listed in Table VI. On a large data set, such as the Adult data set, the change of rules between different runs is very small. Only 1 rule changes in three runs. Even in a small data set, such as the Sick data set, 80% rules are consistent over three runs.

E. Efficiency

To compare the time efficiency of different methods, we modify the CCC and CCU algorithms to find causal relationships with the response variable. The original CCU and CCC do not assume a fixed response variable. There is no minimum support pruning for CCC and CCU. For a fair comparison, we use the same number of variables after the support pruning as the input for CCC and CCU in the following experiments. For our method, we constrain the length of rules as 1, which is comparable with CCC and CCU, and the length of rules as 2 for combined causal rules, denoted as CAR1 and CAR2 respectively. CAR1 and CAR2 were implemented in Java and CCC and CCU were implemented in Matlab. The comparisons were carried out using the same desktop computer.

The comparison of the execution time of CAR1, CAR2, CCC and CCU with increasing data size is shown in Figure 1 (the top diagram). CAR1 is faster than CCC and CCU consistently. CAR2 is similar to CCC and CCU in the small data size end and slower than CCC and CCU in the large data size end. The additional time taken by CAR2 is resulted from forming fair data sets. When increasing the data set size, more rules are accepted as association rules since the confidence interval of an odds ratio becomes narrow (see Section III-B2 for explanation). The increased number of association rules adds additional time for testing causal rules.

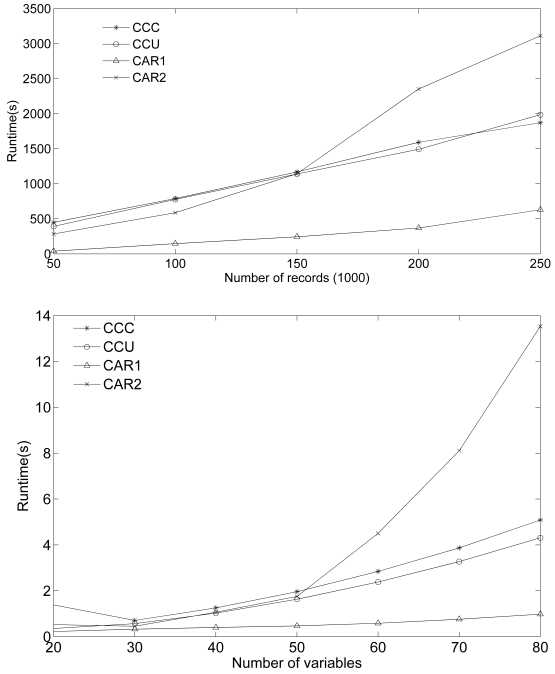

Fig. 1: Scalability with data size and the number of attributes

The comparison of the execution time of CAR1, CAR2, CCC and CCU with increasing number of variables is shown in Figure 1 (the bottom diagram). The relative speeds of different methods are similar to the results above. For CAR1 and CAR2, the number of association rules increases significantly with the increase of the number of variables. The increased number of association rules adds additional time for testing causal rules.

To summarise, for discovering causal rules of single (cause) variables, our method is faster than CCC and CCU. CCC and CCU do not find combined causal factors, and our method does with very competitive efficiency.

V. Conclusion

In this paper, we have proposed a method to find causal association rules from observational data, by integrating association rule mining with retrospective cohort study. We have shown that the method is faster than two efficient constraint based causal relationship discovery methods. Our method is capable of finding causes consisting of combined variables, which are not possible to be uncovered by the other existing methods in the causal Bayesian network scheme. The proposed method is an efficient alternative for causal discovery to the causal Bayesian network approach. It is promising for causal discovery in large and high dimensional data sets. In the proposed method, the selection of controlled variable set is a key for discovering quality causal rules. The validation of the controlled variable set in real world applications will ensure the quality of causal rules discovered. The causal association mining method and constraint based causal discovery methods approach the problem of casual discovery from different directions. They each have their own strengths and limitations. Our future work will study how they complement each other

and explore integrated methods for efficient and quality causal relationship discovery.

REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in SIGMOD, 1993, pp. 207–216.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in Advances in Knowledge Discovery and Data Mining, 307–328, 1996.

[3] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical evaluation," JMLR 11,pp. 171–234, 2010.

[4] A. Asuncion and D. Newman, UCI machine learning repository. http://archive.ics.uci.edu/ml/, 2007.

[5] A. Bhattacharjee, and et al, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," in Proceedings of the National Academy of Sciences 98, pp. 13790–13795, 2001.

[6] C. Borgelt, "Efficient implementations of Apriori and Eclat," in IEEE ICDM Workshop on Frequent Item Set Mining Implementations , 2003, pp. 24–32.

[7] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," in SIGMOD, 1997, pp. 265–276.

[8] D. M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard", JMLR 5, 1287–1330, 2004.

[9] G. F. Cooper, "A simple constraint-based algorithm for efficiently mining observational databases for causal relationships," Data Mining and Knowledge Discovery 1, pp. 203–224, 1997.

[10] A. M. Euser, C. Zoccali, K. Jager, and F. W. Dekker, "Cohort studies: prospective versus retrospective," Nephron Clinical Practice 113, pp. 214–217, 2009.

[11] J. L. Fleiss, B. Levin, and M. C. Paik, Statistical Methods for Rates and Proportions, 3rd ed. Wiley, 2003.

[12] I. Good, "A theory of causality," British Journal for the Philosophy of Science 9, pp. 307–310, 1959.

[13] Q. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," SIGKDD Explorations 11, 2009.

[14] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[15] D. Heckerman, "A Bayesian approach to learning causal networks," in UAI, 1995, pp. 285–295.

[16] D. Heckerman, "Bayesian networks for data mining", Data Mining and Knowledge Discovery 1, pp. 79–119, 1997.

[17] R. Kohavi, D. Sommerfield, and J. Dougherty, "Data mining using MLC++: A machine learning library in C++," in Tools with Artificial Intelligence, IEEE Computer Society Press, pp. 234–245, 1996.

[18] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich, "On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid," European Journal of Operational Research, 184(2), pp. 610–626, 2008.

[19] J. Li, "On optimal rule discovery," IEEE Transactions on Knowledge and Data Engineering 18(4), pp. 460 – 471, 2006.

[20] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in KDD, 1998, pp. 27–31.

[21] S. Mani, G. F. Cooper, and P. Spirtes, "A theoretical study of y structures for causal discovery," in UAI, 2006, AUAI Press.

[22] S. Nadkarni and P. P. Shenoy, "A Bayesian network approach to making inferences in causal maps," European Journal of Operational Research 128(3), pp. 479–498, 2001.

[23] J. Pearl, "From Bayesian network to causal networks," in Bayesian Networks and Probabilistic Reasoning, pp. 1–31, 1994.

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

[25] J. Pearl, and T. S. Verma, "A theory of inferred causation," in Knowledge Representation and Reasoning: Proceedings of the Second International Conference, 1991, pp. 441–452.

[26] J. P. Pellet, "Using Markov blankets for causal structure learning," JMLR 9, pp. 1295–1342, 2008.

[27] H. Reichenbach, The principle of causality and the possibility of its empirical confirmation. Routledge and Kegan Paul, London, 1923.

[28] H. Reichenbach, The direction of time. University of California Press, Berkeley and Los Angeles, 1956.

[29] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable techniques for mining causal structures," Data Mining and Knowledge Discovery 4, pp. 163–192, 2000.

[30] P. Spirtes, C. Glymour, and R. Scheines, Causation, Prediction, and Search, second ed. The MIT Press, Cambridge, MA, USA, 2000.

[31] P. Suppes, A probabilistic theory of causality. North-Holland, Amsterdam, 1970.

[32] P. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," Information Systems 29(4), pp. 293–313, 2004.

[33] M. R. Waldmann and L. Martignon, "A Bayesian network model of causal learning," in Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, 1998, pp. 1102–1107.

[34] G. I. Webb, "Discovering significant patterns," Machine Learning 71, pp. 1–31, 2008.

[35] G. I. Webb, "Layered critical values: a powerful direct-adjustment approach to discovering significant patterns," Machine Learning 71, pp. 307–323, 2008.

[36] M. J. Zaki, "Mining non-redundant association rules," Data Mining and Knowledge Discovery 9, pp. 223–248, 2004.