From Observational Studies to Causal Rule Mining

JIUYONG LI, University of South Australia
THUC DUY LE, University of South Australia
LIN LIU, University of South Australia
JIXUE LIU, University of South Australia
ZHOU JIN, University of Science and Technology China
BINGYU SUN, Chinese Academy of Sciences
SAISAI MA, University of South Australia

Randomised controlled trials (RCTs) are the most effective approach to causal discovery, but in many circumstances it is impossible to conduct RCTs. Therefore observational studies based on passively observed data are widely accepted as an alternative to RCTs. However, in observational studies, prior knowledge is required to generate the hypotheses about the cause-effect relationships to be tested, hence they can only be applied to problems with available domain knowledge and a handful of variables. In practice, many data sets are of high dimensionality, which leaves observational studies out of the opportunities for causal discovery from such a wealth of data sources. In another direction, many efficient data mining methods have been developed to identify associations among variables in large data sets. The problem is, causal relationships imply associations, but the reverse is not always true. However we can see the synergy between the two paradigms here. Specifically, association rule mining can be used to deal with the high-dimensionality problem while observational studies can be utilised to eliminate non-causal associations. In this paper we propose the concept of causal rules (CRs) and develop an algorithm for mining CRs in large data sets. We use the idea of retrospective cohort studies to detect CRs based on the results of association rule mining. Experiments with both synthetic and real world data sets have demonstrated the effectiveness and efficiency of CR mining. In comparison with the commonly used causal discovery methods, the proposed approach in general is faster and has better or competitive performance in finding correct or sensible causes. It is also capable of finding a cause consisting of multiple variables, a feature that other causal discovery methods do not possess.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data Mining

General Terms: Algorithms

Additional Key Words and Phrases: causal discovery, association rule, cohort study, odds ratio

ACM Reference Format:
Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and Saisai Ma, 2015. From Observational Studies to Causal Rule Mining. *ACM Trans. Intell. Syst. Technol.* 00, 00, Article 00 (2015), 27 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Causal discovery aims to infer the cause-effect relationships between variables. Such relationships imply the mechanism of outcome variables taking their values and how

A preliminary version of this work was published in the Proceedings of 2013 IEEE 13th International Conference on Data Mining Workshops, the First IEEE ICDM Workshop on Causal Discovery 2013 (CD2013), pp. 114-123, Dallas, Texas, USA, December 7-10, 2013.
Authors' addresses: J. Li, T. D. Le, L. Liu J. Liu and S. Ma, School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA, 5095, Australia; Z. Jin, Department of Automation, University of Science and Technology, Hefei 230026, China; B. Sun, Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China.

Publication title/2157-6904/2015/MonthOfPublication - ArticleNumber
http://dx.doi.org/10.1145/2746410

the change of cause variables would lead to the change of the outcome variables [Spirtes 2010]. In other words, causality provides the basis for explaining how things have happened and for predicting how the outcomes would be when their causes have changed. Therefore apart from being a fundamental philosophical topic, causality has been studied and utilised in almost all disciplines, e.g. medicine, epidemiology, biology, economics, physics, social science, as a basic and effective tool for explanation, prediction and decision making [Guyon et al. 2010; Kleinberg and Hripcsak 2011]. Some specific examples include the applications in medicine for developing new treatments or drugs for a disease; and in economics for forecasting the results of a particular financial policy and in turn to assist decision and/or policy making.

Randomised controlled trials (RCTs) are recognised as the gold standard for testing the effects of interventions [Shadish et al. 2002; Stolberg et al. 2004]. However, it is also widely acknowledged that in many cases it is impossible to conduct RCTs due to cost and/or ethical concerns. For example, to find out the causal effect of alcohol consumption on heart diseases, it will be unethical to require an experiment participant to drink. Sometimes it is totally forbidden to manipulate a possible cause factor, for example, in a life-threatening situation.

Under these circumstances, observational studies [Rosenbaum 2010; Concato et al. 2000] are considered as the best alternatives to RCTs, and it has been shown that well-designed observational studies can achieve comparative results as RCTs [Concato et al. 2000]. As suggested by the name, observational studies are based on passively observed data, and they do not require manipulation of an exposure (i.e. a potential cause factor). There are two main types of observational studies for causal discovery, cohort studies and case-control studies [Song and Chung 2010; Blackmore and Cummings 2004; Euser et al. 2009]. In a cohort study, based on the status of being exposed to a potential cause factor (e.g. certain radiation), an exposure group of subjects and a non-exposure or control group of subjects are selected, and then followed to observe the occurrence of the outcome (e.g. cancer). In a case-control study, subjects are selected based on the status of the outcome, i.e. the case group consisting of subjects with the outcome and a control group of subjects without the outcome are identified, and then their status of exposure to the potential cause factor is examined. In both types of studies, the effect of an exposure on the outcome is determined by comparing the difference between the exposed/case group and control group. In order to achieve convincing result, an observational study must try to replicate a RCT as much as possible, i.e. the covariate distributions of the two contrasting groups should be as close as possible.

Although observational studies provide an effective approach to causal discovery, they work in the fashion of hypothesis testing, that is, at the commencement of a study, a cause-effect relationship needs to be hypothesised. Then data are collected or retrieved from databases for testing the hypothesis. This requires the prior knowledge or anticipation of the exposures and outcomes, which may not always be available, especially when the number of variables under study is large and the purpose is to explore possible cause-effect relationships, instead of validating an assumed causal relationship. For example, in the study of gene regulation, we may have a clear idea of the possible genetic diseases (outcomes), but which genes could be the possible genetic causes of the diseases may not be known at all. Given the huge number of genes (tens of thousands), it is infeasible to test each gene to find the causes. Therefore to exploit the wealth information in observational data using the well-established methodology of observational studies, we firstly need some efficient ways to generate the hypotheses with high confidence.

Another challenge with observational studies (as well as RCTs) is that even with domain knowledge, it is difficult to foresee a combined cause. For example, multiple

genes may work together to cause a disease, which is normally hard to identify with domain knowledge only.

This is where we can take the advantage of the outcome of data mining research. In the last two decades, huge efforts have been made on association rule mining [Agrawal et al. 1993] and many efficient algorithms have been developed to discover association rules from large data sets [Han and Kamber 2005]. An association rule represents interesting associations among variables, for example, $pizza \rightarrow garlic\ bread$; $\{strong\ wind,\ high\ temperature\} \rightarrow falling\ trees$. Although statistical associations do not necessarily mean causality (for instance, buying garlic bread and pizza together does not indicate that buying one is the cause of buying the other. Mostly likely this is a consequence of a meal deal), it is commonly accepted that associations are necessary for causality.

Our idea is thus to utilise the synergy of observational studies and association rule mining to develop an efficient method for *automated* discovery of causal relationships in large data sets. We firstly use association rule mining to find out the hypothesised cause-effect relationships (represented as association rules) regarding an outcome. Then for each of the hypotheses, we conduct an observational study (e.g. a cohort study) to test if the exposure is a real cause, i.e. to identify if the association rule is a *causal rule*.

As the LHS (left-hand-side or antecedent) of an association rule can comprise multiple attributes, a favourable consequence of using association rule mining here is that it can generate hypothesised causal relationships with compound exposure, such as the rule shown above $\{strong\ wind,\ high\ temperature\} \rightarrow falling\ trees$. In this case, we consider the two attributes as one variable/exposure in our observational studies, hence the validity of the combined cause can be tested.

In the rest of the paper, we will present the definition of causal rules and our approach to identifying CRs (Section 3), the algorithm for mining CRs (Section 4), and the experiment results demonstrating the effectiveness and efficiency of the algorithm (Section 5). Before the presentation, in Section 2, we firstly outline the related work and show the contribution of this paper.

2. RELATED WORK AND CONTRIBUTION

Observational studies [Rosenbaum 2010; Concato et al. 2000] have had a very long history, and there has been a great deal of research on observational studies, by both statisticians and practitioners in medicine and other application areas. The main focus of the research is on how to design good observational studies, including selection of subjects or records, methods for identifying exposed and non-exposed groups to replicate RCTs as closely as possible, and the ways for analysing the data. However, as far as we know, there is little work done on using observational studies for automated causal discovery in large, especially high-dimensional data.

In the field of computer science, causal discovery from observational data has attracted enormous research efforts in the past three decades. Currently Bayesian network techniques are at the core of the methodologies for causal discovery in computer science [Spirtes 2010]. Bayesian networks provide a graphical representation of conditional independence among a set of variables. Under certain causal assumptions, a directed edge between two nodes (variables) in a Bayesian network represents a causal relationship between the two variables [Spirtes 2010; Spirtes et al. 2001]. Over the years, many algorithms have been developed for learning Bayesian networks from data [Neapolitan 2003; Spirtes 2010]. However, up to now it is only feasible to learn a Bayesian network with dozens of variables, or hundreds if the network is sparse [Spirtes 2010]. Therefore, in practice it is infeasible to identify causal relationships using Bayesian network based approaches in most cases.

Indeed the difficulties faced by these causal discovery approaches originate from their goal, i.e. to discover a complete causal model of the domain under consideration. Such a model indicates all pairwise causal relationships among the variables. This, unfortunately, is essentially impossible to achieve when the domain contains a large number of variables. It has been shown that in general learning a Bayesian network is NP-hard [Chickering et al. 2004].

Some constraint based approaches do not search for a complete Bayesian network, so they can be more efficient for causal relationship discovery. Several such algorithms have shown promising results [Cooper 1997; Silverstein et al. 2000; Mani et al. 2006; Pellet 2008; Aliferis et al. 2010]. Based on observational data, these methods determine conditional independence of variables and learn local causal structures. However some of the methods are only capable of discovering the causal relationships represented with some fixed structures, e.g. CCC [Cooper 1997], CCU [Silverstein et al. 2000] and the Y structures [Mani et al. 2006], and they do not identify causal relationships that cannot be represented with these structures. The complexity of other methods for learning a partial Bayesian network in general is still exponential to the number of variables, unless accuracy and/or completeness are traded with efficiency [Aliferis et al. 2010].

Our method tackles the problem of causal discovery from a different perspective. It integrates two well-established methodologies in two different fields for relationship discoveries. The main contribution of this paper is to propose a statistically sound and computational efficient causal discovery method for causal relationship exploration. Cohort studies have been widely accepted for identifying causal links in health, medical and social studies, so the use of cohort studies to uncover causal relationships is methodologically sound. In this paper, the theoretical validity of the proposed method has also been justified by its connection with a well-known causal inference framework – the potential outcome model [Pearl 2000; Morgan and Winship 2007]. Our goal is to automate causal relationship discovery in data, making it possible to explore causal relationships in both large and high dimensional data sets.

Our work also contributes to the area of association rule mining. Association rule mining is a main data mining technique and has many applications in various fields, but a major obstacle of association rule mining is that it produces too many rules and many of them are uninteresting since they represent random associations in a data set [Webb 2008; 2009; Tan et al. 2004; Lenca et al. 2008]. Cohort studies enable us to filter out a large proportion of such uninteresting rules and keep the most interesting ones for a broad range of applications since discovering causal relationships is the goal of the majority of applications.

This paper is an extension of our preliminary work in [Li et al. 2013], with three major developments: (1) A more explicit presentation of the motivation, goal and contribution of the research in the newly written Sections 1 and 2; (2) A new section (Section 3.5) for justifying the validity of the CR framework; (3) A new set of experiments with a total of 13 synthetic data sets for evaluating the performance and scalability of the proposed method , and new experiments on investigating the effect of different matching methods (see Section 5).

3. CAUSAL RULES

3.1. Notations

Let D be a data set for a set of binary variables $(X_1, X_2, \ldots, X_m, Z)$, where $X_1, X_2, \ldots, X_m$ are *predictor* variables and Z is a *response* variable. Values of Z are of user's interest, e.g. having a disease or being normal. Considering a binary data set makes the conceptual discussions in the paper easier, and it does not lose the gener-

ality of a data set that contains attributes of multiple discrete values. For example, a multi-valued data set for the variables (Gender, Age, ...) is equivalent to a binary data set for the variables (Male, Female, 0-19, 20-39, 40-69, ...). In this paper, both the Male and Female variables are kept to allow us to have combined variables that involve them separately, for example, (Female, 40-59, Diabetes) and (Male, 40-59, Smoking).

P is a *combined* variable if it consists of multiple variables $X_1, \ldots, X_n$ where $n \geq 2$, 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 $P = 1$. p is also called a *k-pattern* where k is the length of P (the number of component variables of P). Our ultimate goal is to find out whether $p \rightarrow z$ is a causal rule.

3.2. Association rules

With our approach, we first consider the association between P and Z since an association is normally necessary for a causal relationship.

Odds ratio is a widely used measure for associations in retrospective studies [Fleiss et al. 2003], and we define the odds ratio of a rule as follows.

Definition 3.1 (*Odds ratio of a rule*). Given the following contingency table of a rule, $p \rightarrow z$,

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

where $\mathrm{supp}(x)$ indicates the support of pattern X, the count of value x in the given data set, D, and we have $\mathrm{supp}(p) = \mathrm{supp}(pz) + \mathrm{supp}(p\neg z)$, $\mathrm{supp}(z) = \mathrm{supp}(pz) + \mathrm{supp}(\neg pz)$, and $\mathrm{supp}(pz) + \mathrm{supp}(p\neg z) + \mathrm{supp}(\neg pz) + \mathrm{supp}(\neg p\neg z) = n$, where n is the number of records in the data set, then the odds ratio of the rule $p \rightarrow z$ on D is defined as:

$$\mathrm{oddsratio}_{\mathrm{D}}(p \rightarrow z) = \frac{\mathrm{supp}(pz) * \mathrm{supp}(\neg p\neg z)}{\mathrm{supp}(p\neg z) * \mathrm{supp}(\neg pz)} \tag{1}$$

From the definition, the odds ratio of a rule 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 3.2 (*Association rule*). Using the notations in Definition 3.1, the support of a rule $p \rightarrow z$ is defined as $\mathrm{supp}(p \rightarrow z) = \mathrm{supp}(pz)$. Given a data set D, let min_supp and min_oratio be the minimum support and odds ratio respectively, $p \rightarrow z$ is an association rule if $\mathrm{supp}(p \rightarrow z) > min\_supp$ and $\mathrm{oddsratio}_{\mathrm{D}}(p \rightarrow z) > min\_oratio$, and LHS($p \rightarrow z$) = p and RHS($p \rightarrow z$) = z.

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 ($p \rightarrow z$ and $p \rightarrow \neg z$) are class association rules [Liu et al. 1998] where the confidence ($\mathrm{prob}(z|p)$) is replaced by the odds ratio. Furthermore, only positive association between a predictor variable and the response variable is considered in the above definition as in most cased in practice, we are concerned about the occurrence of the predictor (i.e. $P = 1$) leading to the occurrence of the response (i.e. $Z = 1$).

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.

Traditional association rules are defined by support and confidence [Agrawal et al. 1993]. An association rule in the support and confidence scheme may not show a real association between the LHS and RHS of a rule [Brin et al. 1997]. Therefore in the above definition, we use odds ratio as the indicator of association. The minimum odds ratio in the definition may be replaced by a significance test on $\text{oddsratio}_{\text{D}}(p \to z) > 1$ to ensure that an association rule indicates a significant association between the LHS and RHS of the rule.

The test of significant association is determined as the following.

Let ω be the odds ratio of the rule $p \to z$ on the given data set D, i.e. $\text{oddsratio}_{\text{D}}(p \to z) = \omega$. The confidence interval of ω, $[\omega_-, \omega_+]$, is defined as [Fleiss et al. 2003]:

$$\omega_- = \exp(\ln \omega - z' \sqrt{\frac{1}{\text{supp}(pz)} + \frac{1}{\text{supp}(p\neg z)} + \frac{1}{\text{supp}(\neg pz)} + \frac{1}{\text{supp}(\neg p\neg z)}})$$

and

$$\omega_+ = \exp(\ln \omega + z' \sqrt{\frac{1}{\text{supp}(pz)} + \frac{1}{\text{supp}(p\neg z)} + \frac{1}{\text{supp}(\neg pz)} + \frac{1}{\text{supp}(\neg p\neg z)}})$$

where z' is the critical value corresponding to a 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, $p \to z$ is an association rule.

An important 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.

However, statistically reliable associations do not always indicate causal relationships although causality is mostly observed as associations in data, which can be illustrated by the following example.

Example 3.3. Suppose that we have generated an association rule: "Gender = m" $\to$ "Salary = low" from a data set with the following statistics:

	Salary = low	Salary = $high$
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, which is greater than 1. Therefore as described previously, this odds ratio 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:

	Salary = low	Salary = $high$
Gender = m & College = y	5	20
Gender = f & College = y	15	40

and

	Salary = low	Salary = $high$
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 [Pearl 2000], indicating that associations may not imply causal relationships.

Therefore our idea is to conduct a retrospective cohort study to detect true causal relationships from identified association rules.

3.3. Cohort study

As discussed in Section 1, when randomised controlled trials are practically impossible, observational studies are often used as the alternative approach to finding out the possible cause-effect relationships. A major type of observational studies is cohort studies, which can be conducted in either of the two ways, prospective and retrospective [Euser et al. 2009; Fleiss et al. 2003]. In a perspective cohort study, researchers follow cohorts over time to observe their development of a certain outcome. In a retrospective study, researchers look back at events that already occurred. In a data mining setting, as the data we have are historical records, we adopt the idea of a retrospective cohort study in this paper.

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 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*, and "middle aged" and "males" indicate the common characteristics shared by the two cohorts. A significant difference in the value of the outcome or response variable (lung cancer) 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 *control variable set*.

From the above description, the core requirement for a cohort study is to obtain the matched exposure and non-exposure groups such that the distribution of control variable set of the two groups are the 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 and the control 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 control 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.

In the following, we will define causal rules using the idea of retrospective cohort studies.

3.4. Causal rule definition

Given an association rule as a hypothesis that the LHS of the rule causes its RHS. The variable of the LHS is an exposure variable and the variable of the RHS is the response variable. Let all other variables be included in the control variable set initially. We will discuss how to refine this control variable set in Section 4.2.

3.4.1. Fair data sets. 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 response 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 "matched" values for all the control 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 no more matched pairs can be found. As a result, the distributions of the control variables in the exposure and non-exposure groups are identical or similar to each other.

We formulate the above discussions as the following definition.

Definition 3.4 (*Matched record pair*). Given an association rule $p \rightarrow z$ and a set of control variables C, a pair of records match if one contains value p, the other does not, and both have the matched values for C according to certain similarity measure.

The simplest matching is the exact matching, in which we require a pair of records have exactly the same values for control variables. For example, assume that $C = (A, B, E)$ is the control variable set for association rule $p \rightarrow z$, then records $(P = 1, A = 1, B = 0, E = 1)$ and $(P = 0, A = 1, B = 0, E = 1)$ form a matched pair. Many other similarity measures can be used for finding matched pairs of records, e.g. Euclidean distance, Jaccard distance [Han and Kamber 2005], Mahalanobis distance and propensity score [Stuart 2010], each having its own merit and disadvantages. As this paper is focused on developing and evaluating the idea of integrating association rule mining and cohort studies for causal discovery, we do not conduct extensive investigation on the different matching methods, and in our experiments, we use the exact matching and compare it with Jaccard distance matching.

Definition 3.5. [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 control 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.

Example 3.6. Given an association rule $a \rightarrow z$ identified using the following data set, and the control variable set $C = (M, F, H, U, P)$, where M stands for Male, F for Female, H for High school graduate, U for Undergraduate, and P for Postgraduate.

ID	A	M	F	H	U	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

With exact matching, 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).

In the above definition, the requirement of the maximum sub data set of D is for the best utilisation of the data set.

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

Since with a fair data set for a rule the exposure and non-exposure groups are identical or similar 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 assume that the difference of the outcome is caused by the difference of the values of the exposure variable.

Next, 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.

3.4.2. Causal rules. When the values of the response variable are taken into consideration, there are four possibilities 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 $p \rightarrow z$ can be represented 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 groups; 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 non-exposure groups. In Example 3.6, $n_{11} = 0$, $n_{12} = 2$, $n_{21} = 0$, and $n_{22} = 1$.

Using the above notation, we can have the following definition [Fleiss et al. 2003]:

Definition 3.7 (*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:

$$\text{oddsratio}_{D_f}(p \rightarrow z) = \frac{n_{12}}{n_{21}} \tag{2}$$

In our experiments, we replace zero count by 1 to avoid infinite odds ratios.

The above definition leads to the definition of a causal rule:

Definition 3.8 (*Causal 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 rule, if its odds ratio on its fair data set, $\text{oddsratio}_{D_f}(p \rightarrow z) > min\_oratio$, where min_oratio is the minimum odds ratio.

Alternatively, to check if an association rule is a causal rule, we can use the significance test on the odds ratio of the rule on its fair data set with matched pairs. Let $\text{oddsratio}_{D_f}(p \to z) = \omega'$ in the fair data set, the confidence interval of the odds ratio for matched pairs is defined as [Fleiss et al. 2003]:

$$\omega_- = \exp(\ln \omega - z' \sqrt{\frac{1}{n_{12}} + \frac{1}{n_{21}}})$$

and

$$\omega_+ = \exp(\ln \omega + z' \sqrt{\frac{1}{n_{12}} + \frac{1}{n_{21}}})$$

where z' is the critical value corresponding to a desired level of confidence ($z' = 1.96$ for 95% confidence) and ω'_- is the lower bound of $\text{oddsratio}_{D_f}(p \to z)$ 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.

Based on Definition 3.8, 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. We provide further justifications in the following section.

3.5. Justifications for the definition of causal rules

The potential outcome or counterfactual model [Pearl 2000; Morgan and Winship 2007] is a major framework for causal inference and it is widely used in social science, health and medical research. In this section, we will demonstrate that the causal rules defined over a fair data set is consistent with the causal relationships modelled under the potential outcome framework.

In the potential outcome model, each individual i in a population has two potential outcomes with respect to a treatment: when taking the treatment ($T_i = 1$), the potential outcome is Z_i^1; and when not taking the treatment ($T_i = 0$), the potential outcome is Z_i^0, where Z_i^1 and Z_i^0 are random variables taking values in $\{0, 1\}$. $Z_i^j = 1$ ($j \in \{0, 1\}$) stands for an outcome of interest, such as a recovery.

In practice, we are only able to observe one potential outcome (Z_i^1 or Z_i^0) since an individual can only be placed in either the treatment group ($T_i = 1$) or the control group ($T_i = 0$), and the other potential outcome will need to be estimated. For example, if we know that Jack did not take Panadol (i.e. $T_i = 0$ considering Panadol is the treatment), and now he gets a high temperature (i.e. $Z_i^0 = 1$ assuming high temperature is an outcome), the question that we are asking is, what the outcome would be if Jack had taken Panadol, i.e. we want to know the potential outcome Z_i^1. So the potential outcome model is also called counterfactual model.

Let us assume that we have both Z_i^1 and Z_i^0 of an individual i. With the potential outcome model, the causal effect of the treatment on i is defined as:

$$\delta_i = Z_i^1 - Z_i^0 \tag{3}$$

We often aggregate the causal effects on individuals in the population (or samples) and obtain the average causal effect as the following, where $E[.]$ is the expectation of a random variable.

$$E[\delta_i] = E[Z_i^1] - E[Z_i^0] \tag{4}$$

In the above equation i is kept as in other work on the counterfactual framework, to indicate individual level heterogeneity of potential outcomes and causal effects [Morgan and Winship 2007].

To link the above discussion to our definition of causal rules, treatment T and Z_i^j ($j \in \{0,1\}$) are the exposure variable P and the response variable Z respectively in the causal rule definition. In the following, we keep using the notation of the potential outcome framework.

Since we are only able to observe one of the two potential outcomes for each individual i, the causal effect in equation (4) cannot be estimated from any data set directly. However, it can be estimated under a perfect stratification of the data [Morgan and Winship 2007], where for a stratum samples within treatment and control groups are collectively indistinguishable from each other on the values of the stratifying variables and the samples are only different on the observed treatment status. Furthermore the outcome status of a sample is purely random. In this case, we can assume that:

$$E[Z_i^1 \mid T_i = 0, D_{ps}] = E[Z_i^1 \mid T_i = 1, D_{ps}] \tag{5}$$

$$E[Z_i^0 \mid T_i = 1, D_{ps}] = E[Z_i^0 \mid T_i = 0, D_{ps}] \tag{6}$$

where S represents that the data set is perfectly stratified using the stratifying variables.

The above equations indicate that the potential outcome of an individual taking a treatment (in fact she/he has not) can be estimated by the 'real' outcome of the matched individual who has taken the treatment. Similarly, the potential outcome of an individual not taking a treatment (in fact she/he has taken) can be estimated by the 'real' outcome of the matched individual who has not taken the treatment.

Samples in a fair data set in fact are perfectly stratified, as samples in the exposure and non-exposure groups have the same distribution in terms of the values of control variables, and the value of the response variable of a sample in the exposure or in the non-exposure group is random. Therefore according to equations (5) and (6), for a fair data set D_f, we have:

$$E[Z_i^1 \mid T_i = 0, D_f] = E[Z_i^1 \mid T_i = 1, D_f] \tag{7}$$

$$E[Z_i^0 \mid T_i = 1, D_f] = E[Z_i^0 \mid T_i = 0, D_f] \tag{8}$$

Let us now show how to estimate the causal effect, $E[\delta_i]$ with a fair data set. In a fair data set, the number of individuals being treated is the same as the number of individuals not being treated. Therefore the average causal effect can be represented as the following:

$$E[\delta_i]_{D_f} = \frac{1}{2}(E[Z_i^1 \mid T_i{=}1, D_f] - E[Z_i^0 \mid T_i{=}1, D_f]) + \frac{1}{2}(E[Z_i^1 \mid T_i{=}0, D_f] - E[Z_i^0 \mid T_i{=}0, D_f]) \tag{9}$$

In the above formula, based on equations (7) and (8), we substitute $E[Z_i^0 \mid T_i = 0, D_f]$ and $E[Z_i^1 \mid T_i = 1, D_f]$ for $E[Z_i^0 \mid T_i = 1, D_f]$ and $E[Z_i^1 \mid T_i = 0, D_f]$ respectively. As a result, the average causal effect in the fair data set is estimated as the following:

$$E[\delta_i]_{D_f} = E[Z_i^1 \mid T_i = 1, D_f] - E[Z_i^0 \mid T_i = 0, D_f] \tag{10}$$

where both outcomes are observable. So when there is no sample bias, we can remove the superscripts and subscripts and obtain the average causal effect of the samples (or a population) as the following:

$$\Delta = E[Z \mid T = 1, D_f] - E[Z \mid T = 0, D_f] \tag{11}$$

This formula suggests that following the potential outcome model, the causal effect is the difference of the outcomes in the treatment (exposure) group and the control (non-exposure) group in a fair data set. In our definition of a causal rule, we also use the difference of outcomes in different groups to identify causal rules, except that we use the odds ratio to represent the difference as a cohort study does instead of the above arithmetic difference. Therefore, the definition of a causal rule over a fair data set is correct, in the sense that it is consistent with the approach under the potential outcome framework.

4. ALGORITHM

In this section we present the algorithm (Algorithm 1) for causal rule mining (called CR-CS in the rest of this paper). The algorithm integrates association rule mining with causal relationship test based on cohort studies. In the following, we firstly discuss two anti-monotone properties for efficient generation of candidate causal rules, and we then discuss the selection of control variables for building a fair data set. Finally, we introduce the details of detecting causal rules from the candidate causal rules.

ALGORITHM 1: Causal Rule mining with Cohort Study (CR-CS)

Input: Data set D with the response variable Z, the minimal local support δ, the maximum length of rules k_0, and the minimum odds ratio α.
Output: A set of causal rules

let causal rule set $R_C = \emptyset$
add 1-patterns to a prefix tree T (see Section 4.3) as the 1st level nodes
count support of the 1st level nodes with and without response z
remove nodes whose local support is no more than δ // Support pruning
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 control variable set $C = X \backslash (I, E, LHS(r_i))$
    create a fair data set for r_i // Function 1
    if $\text{oddsratio}_{D_f}(r_i) > \alpha$ **then**
      move r_i to R_C
      remove $LHS(r_i)$ from the k-th level of T // Observation 1
    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 response z
  remove nodes whose local support is no more than δ // Support pruning
  remove nodes of patterns whose supports are the same as those of their sub-patterns respectively // Observation 2
end while
output R_C

4.1. Anti-monotone properties

Anti-monotone properties are at the core for efficient association rule mining. For example a well known anti-monotone property is that a super set 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 that we will apply to candidate causal rule pruning.

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 $\mathrm{cov}(p)$ to represent the set of records in D containing value p, and we call $\mathrm{cov}(p)$ the covering set of p. A rule is *redundant* if it is implied by one of its more general rules.

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. This implies that when a rule is specified, although additional conditions are added to the LHS of the rule, the conditions do not change the causal relationship. Therefore for the purpose of discovering causal rules/relationships, more specific candidate causal 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* $\mathrm{supp}(px) = \mathrm{supp}(p)$*, rule* $px \rightarrow z$ *and all more specific rules of* $px \rightarrow z$ *are redundant.*

PROOF. If $\mathrm{supp}(px) = \mathrm{supp}(p)$, then $\mathrm{cov}(px) = cov(p)$. In other words, both $p \rightarrow z$ 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 is $px \rightarrow z$. If $p \rightarrow z$ is not a causal rule, nor is $px \rightarrow z$. Hence rule $px \rightarrow z$ is redundant.

Let rule $pxy \rightarrow z$ be a more specific rule of rule $px \rightarrow z$. If $\mathrm{supp}(px) = \mathrm{supp}(p)$, then $\mathrm{supp}(pxy) = \mathrm{supp}(py)$. 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 [Agrawal et al. 1996]. Both anti-monotone properties 1 and 2 can be used in the same way as the anti-monotone property of support.

4.2. Control variables

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

Let X represent the set of all predictor variables, and as before P is the exposure variable and C is a set of control variables. Initially, let $C = X \backslash P$.

Definition 4.1 (*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 random with respect to the value of the response variable can be considered as noises and need not to be controlled. With Example 3.3, when we test the association rule "Gender = m" → "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.

The combination of multiple irrelevant variables can be relevant. However, we do not consider combined variables in the control variable set. There will be many combined relevant variables and the support of combined variables are normally small. Therefore when they are included in the control variable set, it is very likely to have empty exposure or non-exposure groups.

Definition 4.2 (*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. Because if an exclusive variable is controlled, the exposure group or the non-exposure group may be empty, thus we are unable to do a cohort study. Let us take $\epsilon = 0$ as an example. When $\text{supp}(pq) = 0$, we will have samples with $(P = 1, Q = 0)$, $(P = 0, Q = 1)$ and $(P = 0, Q = 0)$, but not $(P = 1, Q = 1)$. In this case, for a record in the non-exposure group with $(P = 0, Q = 1)$, no match can be found in the exposure group with $(P = 1, Q = 1)$. When $\text{supp}(\neg pq) = 0$, we will have samples with $(P = 1, Q = 1)$, $(P = 1, Q = 0)$ and $(P = 0, Q = 0)$, but not $(P = 0, Q = 1)$, then for a record in the exposure group with $(P = 1, Q = 1)$, it is impossible to find a match with $(P = 0, Q = 1)$.

A main type of exclusive variables are those caused by database constraints. For example, P represents the highest qualification being high school and Q represents the highest qualification being university degree. As they both belong to the same domain in a relational data set, and an individual has only one highest qualification, P and Q are mutually exclusive ($\text{supp}(pq) = 0$). In this case, it is not necessary to control Q as it does not affect the finding about whether P is a cause of the response variable.

Another type of exclusive variables are redundant attributes. Let us assume that two variables have the identical values but different names. e.g. P and Q. They are mutually exclusive since $\text{supp}(\neg pq) \leq \epsilon$. We do not need to test both separately to see if they are causes of the response variable since one test is enough. However, if we include Q in the control variable set, we will not be able to test P since the fair data set is empty.

Exclusive variables can be confounding variables, for example $P = thunder$ and $Q = storm$ may be mutually exclusive in a data set since $\text{supp}(\neg pq) \leq \epsilon$. Let us assume that they jointly cause the response. If we control Q, P will not be tested as a cause. When we remove Q from the control variable set, we will be able to find P as a cause. It is not difficult to find out that Q is a confounder of P in post processing since they are strongly associated.

4.3. Candidate causal rule generation

This algorithm makes use of branch and bound search similar to Apriori [Agrawal et al. 1996] for association rule mining. The algorithm employs support pruning plus the two pruning criteria (Observations 1 and 2) presented in Section 4.1, and therefore searches much smaller search space than Apriori. The algorithm is based on a prefix tree structure for candidate generation, storage and counting. The prefix tree structure has been shown to support efficient implementation for branch and bound search [Borgelt 2003].

A prefix tree is an ordered tree to store ordered sets (see Figure 1 for an example). In our algorithm, each node stores a set of nonzero variable values (or a potential LHS of a rule). We assume that nonzero variable values are coded and ordered, and this is to prevent generating duplicate candidate causal rules. A node stores the prefix set of the

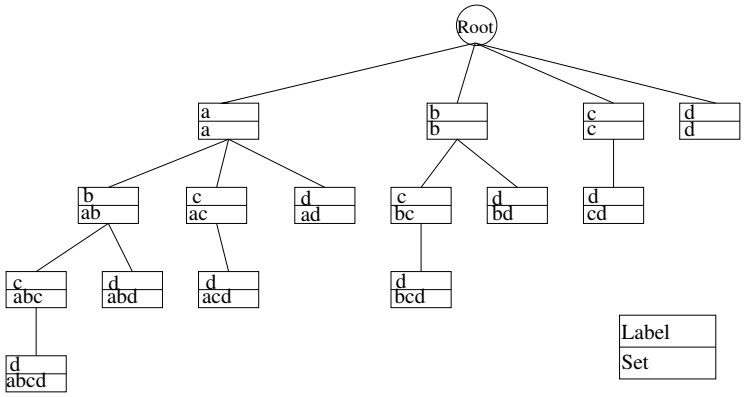

Fig. 1: An example prefix tree. a, b, c, d stand for nonzero values of variables A, B, C and D.

sets stored in its child nodes, and a child node is labeled by the different value between its stored set and the set of its parent. The root of the prefix tree stores an empty set.

The set of labels along the path from the root to a node is the set of values stored in the node. This index makes the search, insert, and delete a node easy. This property also makes the counting of supports of value sets efficient too. If a set stored in a parent node is not in a record, all sets stored in its child nodes will not be in the record too. This property avoids many unnecessary search in the counting process. In the counting process, each node store two counts, for example, $\mathrm{supp}(pz)$ and $\mathrm{supp}(p\neg z)$. As a result, the contingency table of $p \to z$ is determined.

All super sets with the same prefix stored under a parent node storing the prefix. When the parent node is removed, so are all super sets. This suits the forward pruning for candidate causal rule generation very well. For more efficient pruning, the backtrack links to other parent nodes in the prefix tree are also added. For example, in Figure 1 node abc links to node bc, and such link will facilitate the pruning by Anti-monotone property 2.

Referring to Algorithm 1, the code involved in the candidate generation includes Lines 2-4, 9, and 20-23. They are self-explanatory based on the discussions in this and the previous subsection.

4.4. Causal rule detection

This process involves three steps, as discussed below.

4.4.1. Determining control 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 $y \to z$. The above identified irrelevant variables are excluded from the control variable set. These are implemented by Lines 6 and 12.

Secondly we identify the exclusive variables of an exposure variable, say P, according to Definition 4.2 where ϵ is set to the same value as the minimum local support. We exclude the identified exclusive variables from the control variable set. These are implemented by Lines 11 and 12.

The remaining variables then form the control variable set. The control variable set can be viewed as a set of patterns in association rule mining. For example, if male, female, college and postgraduate form the control variable set, the set includes the

Function 1 Create a fair data set for rule $p \to z$
Input: Data set D, rule $p \to z$, and control variable set C
Output: a fair data set for rule $p \to z$, D_f
1: find the covering set of $c(C = 1)$, D_c
2: split D_c into D_{cp} and $D_{c\neg p}$ // D_{cp} contains value p and $D_{c\neg p}$ does not
3: let $D_f = \emptyset$
4: **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**
5: **for** each record t_j in $D_{c\neg p}$ **do**
6: **if** t_i and t_j are matched w.r.t the values of C **then**
7: move t_i and t_j to D_f
8: **end if**
9: **end for**
10: **end for**
11: output D_f

patterns {(male, college), (male, postgraduate), (female, college), (female, postgraduate)}.

4.4.2. Creating fair data set. We select the samples from the given data set D to get the fair data set for rule $p \to z$, 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 a matched record in $D_{c\neg p}$ with respect to the control variable set. We have implemented exact matching and the matching using Jaccard distance. If there are more than one matched records, choose one randomly. Add the pair of records to the fair data set. If there is no matched record in $D_{c\neg p}$, move to the next record.

4.4.3. Testing causal rules. To check if an association rule $p \to z$ is a causal rule, we firstly follow Definition 3.7 to calculate the odds ratio of the rule on its fair data set created in the previous step. Then according to Definition 3.8, if the odds ratio is greater than the given minimum odds ratio, we can say that $p \to z$ is a causal rule. This has been implemented by Line 14 in Algorithm 1. Alternatively, we can use the method introduced in Section 3.4.2 to test the significance of the odds ratio of the rule on its the fair data set. If the odds ratio is significantly higher than 1 for a given confidence level, then we conclude that P is a cause of Z.

5. EXPERIMENTS

5.1. Data sets and parameters

To evaluate CR-CS, the proposed causal rule mining algorithm, twenty four synthetic data sets and eight frequently used public data sets were employed in the experiments. A summary of the data sets is given in Table I. The number of variables in the table refers to the number of predictor variables in a data set. All predictor variables and the response variable are binary variables, with values of 1 or 0 indicating the presence or absence of an attribute correspondingly. The class variable in each of the eight public data sets is set as the response variable in our experiments. The distributions refer to the percentages of the two different values of response variables in the data sets. For the synthetic data sets, the ground truth column represents the number of true single causes and known combined causes each consisting of two predictor variables.

The first fifteen synthetic data sets in Table I were used to evaluate the performance of CR-CS in finding single causal rules in comparison with the Bayesian network based

Table I: A summary of data sets used in experiments

Name	#Records	#Variables	Distributions	Ground Truth
V20-2K	2000	19	41.9% & 58.1%	7
V20-5K	5000	19	41.9% & 58.1%	7
V20-10K	10000	19	41.9% & 58.1%	7
V40-2K	2000	39	37.6% & 62.4%	7
V40-5K	5000	39	37.6% & 62.4%	7
V40-10K	10000	39	37.6% & 62.4%	7
V60-2K	2000	59	52.5% & 47.5%	7
V60-5K	5000	59	52.5% & 47.5%	7
V60-10K	10000	59	52.5% & 47.5%	7
V80-2K	2000	79	50.6% & 49.4%	8
V80-5K	5000	79	50.6% & 49.4%	8
V80-10K	10000	79	50.6% & 49.4%	8
V100-2K	2000	99	48.1% & 51.9%	6
V100-5K	5000	99	48.1% & 51.9%	6
V100-10K	10000	99	48.1% & 51.9%	6
V200-10K	10000	199	19.8% & 80.2%	20
V400-10K	10000	399	19.8% & 80.2%	40
V600-10K	10000	599	19.8% & 80.2%	60
V800-10K	10000	799	19.8% & 80.2%	80
V1000-10K	10000	999	19.8% & 80.2%	100
Name	#Records	#Variables	Distributions	Known combined rules
V8-2K	2000	7	45.1% & 54.9%	2
V12-2K	2000	11	72.1% & 27.9%	3
V16-2K	2000	15	45.2% & 54.8%	3
V20-2K-cmb	2000	19	55.6% & 44.4%	4
German	1000	60	30.0% & 70.0%	-
Kr-vs-kp	3196	74	47.8% & 52.2%	-
Mushroom	8124	215	48.2% & 51.8%	-
Tic-tac	958	27	34.7% & 65.3%	-
Adult	48842	99	23.9% & 76.1%	-
Hypothyroid	3163	51	4.8% & 95.2%	-
Sick	2800	58	6.1% & 93.9%	-
Census income	299285	495	6.2% & 93.8%	-

methods, PC-Select, CCC, and CCU. Those synthetic data sets of random Bayesian networks were generated using the TETRAD software (http://www.phil.cmu.edu/tetrad/). In TETRAD, we firstly generate randomly the structure of the BN using the "simulate data from IM" template. The conditional probability table was also randomly assigned, which will be used to simulate the data. The data sets were then generated using the built-in Bayes Instantiated Model (Bayes IM). In the Bayes IM, the data of each binary variable was randomly generated so that the distributions of all the variables satisfy the constraints in the conditional probability tables. We selected a node in each of the BNs as the fixed target for running the algorithms.

The next five synthetic data sets (V200-10K, ..., V1000-10K) were used to assess the efficiency of the algorithms. To generate those large data sets with a fixed propotion of nodes being the parents of the target node (which is not practical with TETRAD), we firstly draw simple BNs where some predictor variables are parents of the response

variable, and some are not. We then use logistic regression to simulate the data sets for those BNs. The total number of causes in each BN is given in Table I.

Meanwhile the four data sets, V8-2K, V12-2K, V16-2K, and V20-2K-cmb, are for assessing the ability of CR-CS to discover combined causes. These four synthetic data sets have been generated with the following procedure. We firstly generate a data set for a random BN using TETRAD and choose a node as the target. To create a known combined cause, we randomly select a parent variable, X, of the target in the generated BN to split it into two new variables, X_a, X_b. The new variables must satisfy two conditions: (1)$X = X_a \wedge X_b$ (i.e. $X = 1$ if and only if $X_a = 1$ and $X_b = 1$), and (2) X_a and X_b are not associated with the response variable. The number of known combined causes are shown in Table I. Note that we do not have a complete ground truth of all combined causes, as there may be other combined causes in the data set due to the combinations of non-causal single variables. In the experiments, we investigate the performance of CR-CS in terms of the ability to recover known combined causes.

Among real world data sets, Hypothyroid and Sick are two medical data sets and they were originally retrieved from the Thyroid Disease folder of the UCI Machine Learning Repository [Bache and Lichman 2013] and then discretised by using the MLC++ discretisation utility [Kohavi et al. 1996]. The Adult data set is an extraction of the USA census database in 1994 and it was also retrieved from the same repository. In our experiments, all continuous attributes have been removed from the original Adult data set. These three data sets were used in the experiments for testing the effectiveness of CR-CS, in comparison with other methods (see Sections 5.2 and 5.3). They were also used for evaluating the stability (Section 5.4) of CR-CS and the impact of different matching methods (Section 5.5).

The Census Income (KDD) data set was also sourced from the UCI Machine Learning Repository. We combined the training and test data sets and then sampled 50K, 100K, 150K, 200K and 250K records for the experiments. Continuous attributes have been removed. The data set and the last five synthetic data sets (with 10K records) were used to assess the efficiency of CR-CS (Section 5.6). Other real world data sets are also from UCI Machine Learning Repository and are used to investigate the number of combined causes discovered by CR-CS.

In the experiments, while the default minimum local support (δ in Algorithm 1) was 0.05, we set it to 0.01 for the Adult data set in the comparison with the other three methods, CCC [Cooper 1997], CCU [Silverstein et al. 2000] and PC-select [Colombo et al. 2014]. The confidence level was set to 99% for calculating the confidence interval (lower bounds and upper bounds) of the odds ratio for synthetic data sets, and 95% for real world data sets considering the noises in the real world data sets.

5.2. Causal rules vs. association rules

Causal rules have advantages over association rules. An association rule may represent a spurious relationship between two variables as a statistical association does not necessarily mean that the two variables are related or directly related (while a causal rule indicates that the two variables have a direct relationship given the observed variables). Those spurious association rules could not be removed by increasing thresholds. They can only be identified by analysing the relationship by shielding the effects of other variables.

To investigate the difference between association rule mining and causal rule mining, we compared the results obtained by CR-CS with the results of various types of association rule mining. From Table II, the number of causal rules is significantly smaller than the numbers of other types of (association) rules, including association rules [Agrawal et al. 1996], non-redundant rules [Zaki 2004], and optimal rules [Li 2006]. Associations are measured by the odds ratio defined in Definition 3.1, and their

Table II: Comparison of the numbers of association rules (AR), non-redundant rules (NRR), optimal rules (OR) and causal rules (CR). Many association and interesting rules are not causal.

	#AR	#NRR	#OR	#CR
Adult	3108	2863	976	46
Hypothyroid	39476	17692	3237	30
Sick	56183	28698	3917	21

significance is tested using the method discussed in Section 3.2, and all the methods used the same minimum local support. The maximum length of rules is 4.

The number of causal rules obtained from a data set is very small. They may not be enough for classification since not every record in the data is covered by a causal rule. However, they are more 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, which makes it easy for these rules to be easily interpreted and applied to solve real world problems where only short rules are preferred.

5.3. Causal rules vs. findings of other causal discovery methods

To evaluate the performance of CR-CS, we conducted a set of experiments with the first 15 synthetic data sets and 3 real world data sets, Adults, Hypothyroid and Sick, and compared the performance of CR-CS with the constraint based methods, CCC [Cooper 1997], CCU [Silverstein et al. 2000], and PC-select [Colombo et al. 2014].

As mentioned in Section 2, CCC [Cooper 1997] and CCU [Silverstein et al. 2000] are two efficient constraint based causal discovery methods. Both of them learn the simple structures involving three variables with certain dependence/independence relationships among them, and infer causal relationships from the structures. Both methods assume no hidden and no confounding variables in data sets. PC-select [Colombo et al. 2014] is a local causal discovery method that finds all the parents and children of a given node. It is similar to the well-known PC algorithm [Spirtes et al. 2001] for learning a Bayesian network, except that it only finds the local causal relationships around a given response variable. The PC algorithm can return optimal result

In the experiments, CCC and CCU were restricted to identify the structures involving the response variables only. When a statistical significance test was involved, 95% confidence level is used. With our method (CR-CS), since there are small variations in the causal rules discovered in different runs due to random selection of matched pairs when a record has multiple matches, in the experiments, with one data set, we generated causal rules (i.e. ran the algorithm) three times and chose the rules occurring at least twice in the three runs.

5.3.1. Experiment results of synthetic data. Table III shows the precision (P), recall (R), and F_1-measure (F_1) of the four methods for the 15 synthetic data sets with different number of variables and samples. As we can see from the table, PC-select and CR-CS are significantly better than CCC and CCU in precision, recall, and F_1 measure. CR-CS and PC-select achieve good results with more than 70% in precision and F_1 measure for most of the synthetic data sets.

To investigate if a method performs better than the other, for each pair of methods, we conduct the Wilcoxon test [Demšar 2006] of the F_1-measures of the results obtained by the pair of methods with the fifteen data sets. Table IV shows the pairwise test results for the four methods. Overall, PC-select and CR-CS are significantly better than CCC and CCU, but there is no evidence to conclude that CR-CS or PC-select is

Table III: Performance of CCC, CCU, PC-select, and CR-CS in finding single rules with synthetic data sets. P, R and F_1 represent precision, recall and F_1-measure, respectively.

	CCC			CCU			PC-select			CR-CS		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
V20-2K	0.75	0.86	0.80	1.00	0.57	0.73	0.83	0.71	0.77	1.00	0.57	0.73
V20-5K	0.63	1.00	0.78	0.50	0.43	0.46	1.00	1.00	1.00	0.86	0.86	0.86
V20-10K	0.55	0.86	0.67	0.40	0.29	0.33	1.00	0.86	0.92	1.00	0.86	0.92
V40-2K	0.50	0.86	0.63	0.50	0.43	0.46	0.83	0.71	0.77	1.00	0.52	0.73
V40-5K	0.57	1.00	0.74	0.57	0.57	0.57	1.00	1.00	1.00	1.00	1.00	1.00
V40-10K	0.41	1.00	0.58	0.30	0.43	0.35	0.88	1.00	0.93	1.00	1.00	1.00
V60-2K	0.27	0.57	0.36	0.00	0.00	0.00	0.80	0.57	0.67	1.00	0.57	0.73
V60-5K	0.38	0.86	0.52	0.40	0.29	0.33	0.86	0.86	0.86	1.00	0.86	0.92
V60-10K	0.30	0.86	0.44	0.33	0.57	0.42	0.86	0.86	0.86	0.83	0.71	0.77
V80-2K	0.75	0.75	0.75	1.00	0.38	0.55	1.00	0.75	0.86	1.00	0.63	0.77
V80-5K	0.55	0.75	0.63	1.00	0.50	0.67	1.00	0.75	0.86	1.00	0.75	0.86
V80-10K	0.66	0.88	0.74	0.67	0.25	0.36	0.88	0.88	0.88	1.00	0.88	0.93
V100-2K	0.43	1.00	0.60	0.25	0.33	0.29	0.75	1.00	0.86	0.80	0.67	0.73
V100-5K	0.29	0.83	0.44	0.17	0.17	0.17	0.63	0.83	0.71	0.80	0.67	0.73
V100-10K	0.35	1.00	0.52	0.57	0.67	0.62	1.00	0.83	0.91	0.71	0.83	0.77

better than the other. However, note that PC-select is only suitable for data sets with small number of nodes or sparse data sets with small number of causes of the target. It took more than two hours for PC-select to complete when it was applied to the synthetic data set with 100 nodes and 20 causes of the target, and it failed to return results for the data set with 120 nodes with 26 causes of the target within 24 hours.

Table IV: Wilcoxon signed ranks test results for the four methods with F_1 measure listed in Table III

p-value	CR-CS	PC-select	CCC	CCU
CR-CS	-	0.769	**3.74E-04**	**2.51E-06**
PC-select	0.244	-	**2.49E-05**	**2.63E-06**
CCC	1.000	1.000	-	**0.002**
CCU	1.000	1.000	0.998	-

To evaluate the ability of CR-CS in recovering combined causal rules, we use synthetic data sets with known combined rules as described in section 5.1. We applied CR-CS to the four data sets, V8-2K, V12-2K, V16-2K, and V20-2K-cmb to discover level 2 rules with the 99% confidence level. The experiment results have shown that CR-CS can recover all known combined rules, including 2 rules in V8-2K, 3 rules in V12-2K, 3 rules in V16-2K, and 4 rules in V20-2K. There are also 5, 3, 16, and 15 extra combined rules discovered by CR-CS in the four data sets, respectively. We do not have a means to test if extras are real combined causes. However, the results show that the method is able to uncover known combined causes.

5.3.2. Experiment results of real world data. With the Adult data set, as shown in Table V, CR-CS, CCC and CCU discovered similar number of rules, while PC-select found a relatively small number of rules.

When we look into the rules discovered by these methods, they are quite different. We list in Table VI the most similar and dissimilar rule groups found in the Adult data set using CR-CS and the other methods. We can see that overall CR-CS and PC-select obtained similar results, while only for the variables related to the Education

Table V: Number of causal rules/relationships discovered by CR-CS, CCC, CCU and PC-select with real world data sets

	CR-CS	CCC	CCU	PC-select
Adult	46	53	46	19
Hypothyroid	30	14	10	4
Sick	21	13	3	5

attribute, rules discovered by CR-CS and PC-select are similar to those discovered by CCC and CCU.

Intuitively, Education is the major factor 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 some college but no degree and lower.

Rules discovered by CR-CS and PC-select are dissimilar to those found by CCC and CCU in relation to the Occupation, Workclass and Native-country attributes. There are 11 rules discovered by CR-CS with respect to the Occupation attributes, but only one rule is discovered by CCC and CCU in this group. 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, but they have been missed by CCC and CCU. PC-select, although found fewer number of rules in this group (Occupation), the rules found by it are reasonable. On the other hand, 22 rules related to the Native Country attributes are discovered by CCC, 17 rules by CCU, but only 1 rule by CR-CS in this group. PC-select found two rules in this group, again performing more consistently with CR-CS. Intuitively, Native Country should not a factor for high/low salary. This shows that CR-CS is able to discover more reasonable causal rules.

The combined causal rule discovered by CR-CS is also reasonable. As shown in Table VI, people with some-college education but without any degree and working in a private sector would have low salaries. CR-CS did not discover that people with some-college or with private work-class would have low income at the single rule level as found by CCC and CCU, but it provides more details with the combined causal rule.

To investigate the number of combined causal rules in real world cases, we run CR-CS for eight real world data sets with up to level 4 rules. Table VII shows the number of single and combined causal rules discovered by CR-CS with the 95% confidence level. We can see that the combined causal rules at level 3 and 4 are rare, but CR-CS found a number of combined causal rules at the second level. Although we do not have a ground truth to validate all of the combined rules, some rules are reasonable based on common knowledge. For example, with the Mushroom data set, we can see from Table VIII that poisonous mushrooms are pink and have either evanescent ring type or white spore print. Our common understanding is that poisonous mushrooms are normally in bright color, but not all brightly colored mushrooms are poisonous. These combined causal rules provide more detail on the poisonous mushrooms than just based on their colors, and therefore they are useful in practice. Similarly, CR-CS discovers that mushrooms without bright color and odor are edible, and these rules are also reasonable.

5.4. Stability

The creation of a fair data set is subject to selection bias. Usually there are significantly more exposed cases than non-exposed cases so the data distribution is often skewed for

Table VI: The similar and dissimilar causal rule groups discovered by CR-CS and the other methods in the Adult data set. (some-college: Some college but no degree; exec-managerial: Executive admin and managerial; prof-specialty: Professional specialty; handlers-cleaners: Handlers equip cleaners etc.; machine-op-inspct: Machine operators assemblers & inspectors; adm-clerical: Admin support including clerical; other-service: Other services; farming-fishing: Farming forestry and fishing. sel-emp-inc: Self-employed-incorporated; sel-emp-not-inc: Self-employed-not incorporated.)

Causal rules	CR-CS	CCC	CCU	PC-select
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=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	2
Education=Some-college & Workclass=Private $\rightarrow \leq 50$K	√			

the exposure and non-exposure conditions. 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

Table VII: Number of combined causal rules discovered by CR-CS in real world data sets

	1st Level	2nd Level	3rd Level	4th Level
Adult	45	1	0	0
Census income	77	6	0	0
Germand	8	38	12	5
Hypothyroid	20	7	3	0
Kr-vs-kp	3	15	0	0
Mushrom	26	61	0	1
Sick	13	7	1	0
Tic-tac	8	30	3	0

Table VIII: Some combined causal rules in the Mushrooms data set.

Combined causal rules
Stalk-color-below-ring = pink & Ring-type = evanescent $\rightarrow$ poisoners
Stalk-color-below-ring = pink & Spore-print-color = white $\rightarrow$ poisoners
Odor = none & Stalk-shape = tapering $\rightarrow$ edible
Cap-color = gray & Odor = none $\rightarrow$ edible

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 causal 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 IX. On a large data set, such as the Adult data set, the change of rules between different runs is very small. Only one rule difference in the three runs. Even in a small data set, such as the Sick data set, nearly 90% rules are consistent over three runs.

Table IX: The numbers of causal rules of different runs and the frequent causal rules

fair data set	1	2	3	frequent
Adult	46	46	45	46
Hypothyroid	31	30	30	30
Sick	21	20	21	21

5.5. Results obtained using different matching methods

As described in Section 3.4 (Definition 3.4), when creating a fair data set, different similarity measures can be used for finding matched pairs of records. In the experiments described so far, exact matching has been used. In order to gain some insights into the impact of different similarity measures, we also experimented on our method when Jaccard distance is used in matching a pair of records. Jaccard distance [Han and Kamber 2005] is a commonly used measure of the similarity between records with binary attributes. From Table X, we see that the numbers of rules discovered are very similar across the three data sets with exact matching and the matching using Jaccard distance.

Table X: Results of CR-CS using different matching methods

Data set	Exact matching	Jaccard distance
Adult	46	46
Hypothyroid	30	31
Sick	21	22

5.6. Efficiency

To test the time efficiency of CR-CS, we applied it to the Census Income (KDD) data set and the last five synthetic data sets (with 10K records), to observe its scalability in terms of the number of records and the number of attributes respectively. The experiments were also done in comparison with the other three methods.

As the original CCC and CCU algorithms do not assume a fixed response variable, we ran them with the restriction of only looking for the triplets that contain the response variable. For our method, we ran it in two different versions: CR-CS1 and CR-CS2 respectively. With CR-CS1, we constrained the length of rules to 1, making it comparable with CCC, CCU and PC-select. With CR-CS2, the length of rules was restricted to 2 to allow the discovery of combined causes. CR-CS1 and CR-CS2 were implemented in Java, CCC and CCU were implemented in Matlab, and for PC-select, we used the pcSelect() function of the R package *pcalg* [Colombo et al. 2014; Kalisch et al. 2012]. The comparisons were carried out using the same desktop computer (Quad core CPU 3.4 GHz and 16 GB of memory).

The execution time (in seconds) of CR-CS1, CR-CS2, CCC and CCU with respect to the number of records in the Census Income (KDD) data is shown in Figure 2. The execution time of PC-select is not included as it did not return results on any data set after two hours of execution. From the figure, we can see that CR-CS1 was much faster than CCC and CCU consistently for different record sizes, and even CR-CS2 was also faster than the other methods. The main reason is that our method employs association rule mining to remove non-eligible rules and thus to reduce the search space significantly.

The execution time of CR-CS1, CR-CS2, CCC, CCU and PC-select with respect to the number of attributes is shown in Figure 3 (only the results returned within 6 hours are shown). Similarly, CR-CS1 is more scalable than CCC and CCU, while CR-CS2 is much slower when the number of attributes became big as the number of association rules increased significantly with the increase of the number of attributes, leading to additional time for testing causal rules. Although PC-select can achieve high quality of causal discovery (see Table III), from the Figure 3, we can see that PC-select is inefficient or even infeasible, especially when the number of variables is large.

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed the concept of causal rules and have developed a method to find causal rules from observational data by integrating association rule mining with retrospective cohort studies. Through the integration, our method has been able to take the advantage of the high efficiency of association rule mining to produce candidate causal relationships from large data sets, and then to utilise the idea of cohort studies to obtain reliable causal rules based on the candidates. The validity of the definition of causal rules has been justified to be consistent with the potential outcome model. Experiments results have shown that the proposed method is able to find more reasonable causal relationships comparing to the existing causal discovery methods. Moreover, our method was able to find causes consisting of combined variables, which are not possible to be uncovered by the other existing methods. We have

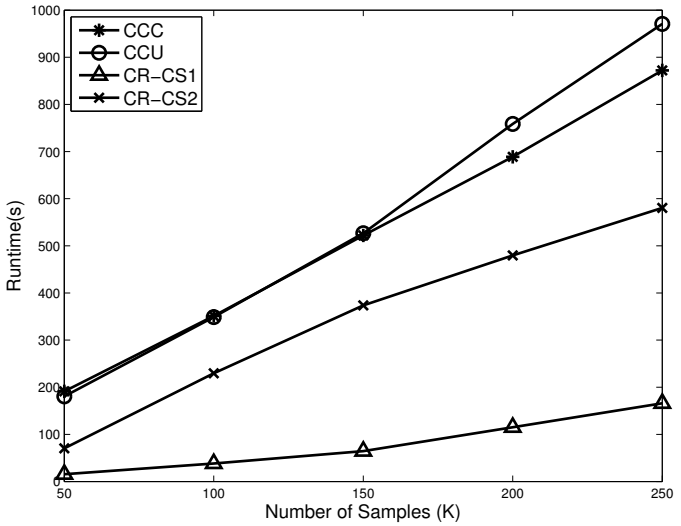

Fig. 2: Scalability with respect to number of records (note:PC-select is not included since it did not return results after two hours of execution)

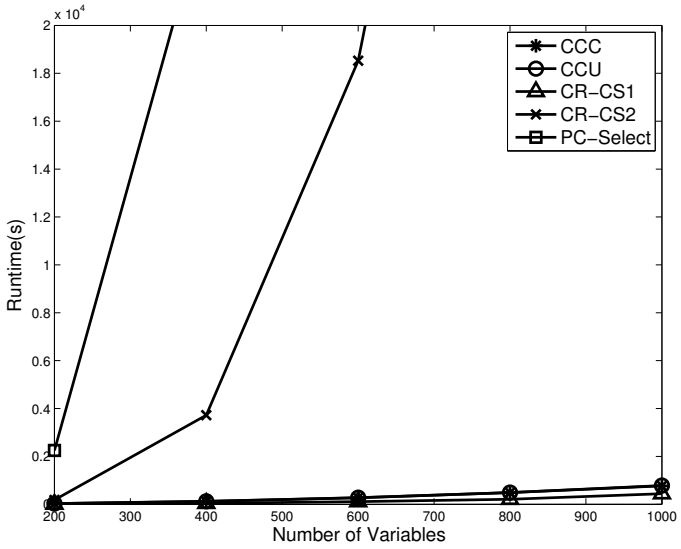

Fig. 3: Scalability with respect to number of attributes

shown that the method is faster than the efficient constraint based causal relationship discovery methods. Hence our method can be used as a promising alternative for causal discovery in large and high dimensional data sets. With the proposed method, the selection of control variable set is a key to discovering quality causal rules. The validation of the control variable set in real world applications will ensure the quality of causal rules discovered.

The proposed causal rule mining method and the constraint based causal discovery approaches tackle the problem of causal discovery from different directions. They each have their own strengths and limitations. Our future work will be studying how they complement each other and exploring integrated methods for efficient and quality causal relationship discovery.

ACKNOWLEDGMENTS

This work has been partially supported by Australian Research Council Discovery Project DP130104090 and DP140103617.

REFERENCES

R. Agrawal, T. Imieliński, and A. Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of SIGMOD'93*. 207–216.

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

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. 2010. Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research* 11 (2010), 171–234.

K. Bache and M. Lichman. 2013. UCI Machine Learning Repository. (2013). http://archive.ics.uci.edu/ml

C. C. Blackmore and P. Cummings. 2004. Observational Studies in Radiology. *American Journal of Roentgenology* 183, 5 (2004), 1203–1208.

C. Borgelt. 2003. Efficient implementations of Apriori and Eclat. In *Proceedings of IEEE ICDM Workshop on Frequent Item Set Mining Implementations*. 24–32.

S. Brin, R. Motwani, and C. Silverstein. 1997. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of SIGMOD'97*. 265–276.

D. Chickering, D. Heckerman, and C. Meek. 2004. Large-Sample Learning of Bayesian Networks is NP-Hard. *Journal of Machine Learning Research* 5 (2004), 1287–1330.

D. Colombo, A. Hauser, M. Kalisch, and M. Maechler. 2014. Package 'pcalg'. (2014). Retrieved March 13, 2014 from http://cran.r-project.org/web/packages/pcalg/pcalg.pdf

J. Concato, Shah N, and R. I. Horwitz. 2000. Randomized, controlled, trials, observational studies, and the hierarchy of research design. *The New England Journal of Medicine* 342, 25 (June 2000), 1887–1892.

G. F. Cooper. 1997. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery* 1 (1997), 203–224.

J. Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* 7 (2006), 1–30.

A. M. Euser, C. Zoccali, K. Jager, and F. W. Dekker. 2009. Cohort studies: prospective versus retrospective. *Nephron Clinical Practice* 113 (2009), 214–217.

J. L. Fleiss, B. Levin, and M. C. Paik. 2003. *Statistical Methods for Rates and Proportions* (3rd ed.). Wiley.

I. Guyon, D. Janzing, and B. Schölkopf. 2010. Causality: Objectives and Assessment. *Journal of Machine Learning Research Workshop and Conference Proceedings* 6 (2010), 1–42.

J. Han and M. Kamber. 2005. *Data Mining: Concepts and Techniques* (2nd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. 2012. Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software* 47, 11 (5 2012), 1–26.

S. Kleinberg and G. Hripcsak. 2011. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics* 44, 6 (2011), 1102–1112.

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

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

J. Li. 2006. On optimal rule discovery. *IEEE Transactions on Knowledge and Data Engineering* 18, 4 (2006), 460–471.

Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, and Bingyu Sun. 2013. Mining causal association rules. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. IEEE, 114–123.

B. Liu, W. Hsu, and Y. Ma. 1998. Integrating classification and association rule mining. In *Proceedings of KDD'98*. 27–31.

S. Mani, G.F. Cooper, and P. Spirtes. 2006. A theoretical study of y structures for causal discovery. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI'06)*. AUAI Press, 314–323.

Stephen L Morgan and Christopher Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.

R. E. Neapolitan. 2003. *Learning Bayesian Networks*. Prentice Hall.

J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. ambridge University Press.

J. P. Pellet. 2008. Using Markov blankets for causal structure learning. *Journal of Machine Learning Research* 9 (2008), 1295–1342.

P. R. Rosenbaum. 2010. *Design of Observational Studies*. Springer.

W. R. Shadish, T. D. Thomas, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2nd. ed.). Houghton Mifflin, Boston.

Silverstein, S. Brin, R. Motwani, and J. Ullman. 2000. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery* 4 (2000), 163–192.

J. W. Song and K. C. Chung. 2010. Observational Studies: Cohort and Case-Control Studies. *Plastic & Reconstructive Surgery* 126, 6 (December 2010), 2234–2242.

P. Spirtes. 2010. Introduction to Causal Inference. *Journal of Machine Learning Research* 11 (2010), 1643–1662.

P. Spirtes, C. C. Glymour, and R. Scheines. 2001. *Causation, Predication, and Search* (2nd. ed.). The MIT Press.

H. O. Stolberg, G. Norman, and I. Trop. 2004. Randomized Controlled Trials. *American Journal of Roentgenology* 183, 6 (2004), 1539–1544.

E. A. Stuart. 2010. Matching methods for causal inference: a review and a look forward. *Statist. Sci.* 25, 1 (2010), 1–21.

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

G. I. Webb. 2008. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning* 71 (2008), 307–323.

G. I. Webb. 2009. Discovering significant patterns. *Machine Learning* 71 (2009), 1–31.

M. J. Zaki. 2004. Mining non-redundant association rules. In *Advances in Knowledge Discovery and Data Mining*. Vol. 9. 223–248.

Received xx 2014; revised xx 2014; accepted xx 2015