From Observational Studies to Causal Rule Mining

JIUYONG LI, THUC DUY LE, LIN LIU, and 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, and hence they can only be applied to problems with available domain knowledge and a handful of variables. In practice, many datasets 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 datasets. The problem is that 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, whereas observational studies can be utilised to eliminate noncausal associations. In this article, we propose the concept of causal rules (CRs) and develop an algorithm for mining CRs in large datasets. 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 datasets have demonstrated the effectiveness and efficiency of CR mining. In comparison with the commonly used causal discovery methods, the proposed approach generally 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. 7, 2, Article 14 (November 2015), 27 pages.
DOI: http://dx.doi.org/10.1145/2746410

This work has been partially supported by Australian Research Council Discovery Project DP130104090 and DP140103617.
A preliminary version of this work was published in the *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops, the 1st IEEE ICDM Workshop on Causal Discovery 2013 (CD'13)*, 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; emails: {jiuyong.li, thuc.le, lin.liu, jixue.liu}@unisa.edu.au, saisai.ma@mymail.unisa.edu.au; Z. Jin, Department of Automation, University of Science and Technology, Hefei 230026, China; email: manjinzhou@gmail.com; B. Sun, Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China; email: bysun@iim.ac.cn.

2157-6904/2015/11-ART14 $15.00
DOI: http://dx.doi.org/10.1145/2746410

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 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, and 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, e.g., 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 nonexposure 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. To achieve a convincing result, an observational study must try to replicate an 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 past 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 datasets [Han and Kamber 2005]. An association rule represents interesting associations among variables, e.g., $pizza \rightarrow garlic\ bread$; $\{strong\ wind,\ high\ temperature\} \rightarrow falling\ trees$. Although statistical associations do not necessarily mean causality, i.e., buying garlic bread and pizza together does not indicate that buying one is the cause of buying the other, as 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 datasets. 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 (CR).

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, e.g., the rule shown earlier $\{strong\ wind,\ high\ temperature\} \rightarrow falling\ trees$. In this case, we consider the two attributes as one variable/exposure in our observational studies, and hence the validity of the combined cause can be tested.

In the rest of the article, we will present the definition of CRs 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 work.

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 nonexposed 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 work 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 article, 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 datasets.

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, as they represent random associations in a dataset [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, as discovering causal relationships is the goal of the majority of applications.

This article 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 datasets for evaluating the performance and scalability of the proposed method, and new experiments on investigating the effect of different matching methods (Section 5).

3. CAUSAL RULES

3.1. Notations

Let D be a dataset 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 dataset makes the conceptual discussions in the article easier, and it does not lose the generality of a dataset that contains attributes of multiple discrete values. For example, a multivalued dataset for the variables (Gender, Age, ...) is equivalent to a binary dataset for the variables (Male, Female, 0–19, 20–39, 40–69, ...). In this article, both the Male and Female variables are kept to allow us to have combined variables that involve them separately, e.g., (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 CR.

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.

The 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, $p \rightarrow z$,

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

where $\text{supp}(x)$ indicates the support of pattern X, the count of value x in the given data set, D, and 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 dataset, then the odds ratio of the rule $p \rightarrow z$ on D is defined as

$$\text{oddsratio}_{\text{D}}(p \rightarrow z) = \frac{\text{supp}(pz) * \text{supp}(\neg p\neg z)}{\text{supp}(p\neg z) * \text{supp}(\neg pz)}. \quad (1)$$

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 $\text{supp}(p \rightarrow z) = \text{supp}(pz)$. Given a dataset D, let *min_supp* and *min_oratio* be the minimum support and odds ratio, respectively; $p \rightarrow z$ is an association rule if $\text{supp}(p \rightarrow z) > \textit{min\_supp}$ and $\text{oddsratio}_{\text{D}}(p \rightarrow z) > \textit{min\_oratio}$; and $\text{LHS}(p \rightarrow z) = p$ and $\text{RHS}(p \rightarrow z) = z$.

In the definition, we consider z as the RHS (right-hand side) 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 ($\text{prob}(z|p)$) is replaced by the odds ratio. Furthermore, only positive association between a predictor variable and the response variable is considered in the preceding 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 preceding definition, we use the 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 follows.

Let ω be the odds ratio of the rule $p \to z$ on the given dataset 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]:

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

and

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

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, and hence P and Z are associated. Equivalently, $p \to z$ is an association rule.

An important advantage of the preceding process is that it is automatically adaptive to the size of a dataset. For a large dataset, the confidence interval of an odds ratio is small, and hence a small odds ratio can be significantly higher than 1. For a small dataset, 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" → "Salary = low" from a dataset 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, whereas 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 datasets are as follows:

	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 preceding 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 subdataset or a superdataset. 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 RCTs 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 article.

A retrospective cohort study selects individuals who have been exposed and have not been exposed to a suspected risk factor but are alike within many other aspects. For example, middle-aged males 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 article, 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 *nonexposure group*, and the set of variables determining the common characteristics of the two groups the *control variable set*.

From the preceding description, the core requirement for a cohort study is to obtain the matched exposure and nonexposure 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 the following variables: education, profession, experience, and location. From a given dataset, we will need to select samples for the exposure and nonexposure 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 CRs using the idea of retrospective cohort studies.

3.4. Causal Rule Definition

Given an association rule as a hypothesis such 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 Datasets. Given a dataset D, for an exposure variable, we use the following process to select samples for the exposure and nonexposure 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 of the control variables. Then, t_i is added to the exposure group, t_j is added to the nonexposure group, and both are removed from the original dataset. This process repeats until no more matched pairs can be found. As a result, the distributions of the control variables in the exposure and nonexposure groups are identical or similar to each other.

We formulate the preceding 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 article 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 dataset D and a set of control variables C, the fair dataset D_f for the rule is the maximum subdataset of D that contains only matched record pairs from D.

Example 3.6. Given an association rule $a \rightarrow z$ identified using the following dataset, 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. Then with exact matching, records (#1, #5), (#2, #6) and (#3, #7) form

three matched pairs and thus a fair dataset for $a \rightarrow z$ includes records (#1, #2, #3 #5, #6, #7).

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

In the preceding definition, the requirement of the maximum subdataset of D is for the best utilisation of the dataset.

Matches in a dataset 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 CRs 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 datasets (one to two rule difference in three runs). Even in a small dataset, more than 80% rules are consistent over different runs.

Since with a fair dataset for a rule the exposure and nonexposure 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 nonexposure groups, which will provide us the method for testing whether an association rule is a CR 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 dataset for rule $p \rightarrow z$ can be represented as follows:

	$P = 0$	
$P = 1$	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 nonexposure groups, n_{12} is the number of matched pairs containing z in the exposure group and $\neg z$ in the non-exposure group, n_{21} is the number of matched pairs containing $\neg z$ in the exposure group and z in the nonexposure group, and n_{22} is the number of matched pairs containing $\neg z$ in both the exposure and nonexposure groups. In Example 3.6, $n_{11} = 0$, $n_{12} = 2$, $n_{21} = 0$, and $n_{22} = 1$.

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

Definition 3.7 (*Odds Ratio of a Rule on its Fair Dataset*). The odds ratio of an association rule $p \to z$ on its fair dataset D_f is

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

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

The preceding definition leads to the definition of a CR.

Definition 3.8 (*Causal Rule*). An association rule ($p \to z$) 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 dataset, $\text{oddsratio}_{D_f}(p \to z) > min\_oratio$, where *min_oratio* is the minimum odds ratio.

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

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

and

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

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, and we then conclude that P is a cause of Z.

Based on Definition 3.8, testing if an association rule is a CR becomes the problem of finding the fair dataset for the rule. A fair dataset 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 dataset is significantly greater than 1, it means that a change of the response variable results 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 is widely used in social science, health, and medical research. In this section, we will demonstrate that the CRs defined over a fair dataset are 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, i.e., 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 the *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 follows, where $E[.]$ is the expectation of a random variable:

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

In the preceding 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 preceding discussion to our definition of CRs, treatment T and Z_i^j ($j \in \{0, 1\}$) are the exposure variable P and the response variable Z, respectively, in the CR 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 dataset 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

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

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

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

The preceding 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 dataset in fact are perfectly stratified, as samples in the exposure and nonexposure 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 nonexposure group is random. Therefore, according to Equations (5) and (6), for a fair dataset D_f, we have

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

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

Let us now show how to estimate the causal effect, $E[\delta_i]$, with a fair dataset. In a fair dataset, 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 follows:

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

In the preceding 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 dataset is estimated as follows:

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

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 follows:

$$\Delta = E[Z \mid T = 1, D_f] - E[Z \mid T = 0, D_f]. \quad (11)$$

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 (nonexposure) group in a fair dataset. In our definition of a CR, we also use the difference of outcomes in different groups to identify CRs, except we use the odds ratio to represent the difference as a cohort study does instead of the preceding arithmetic difference. Therefore, the definition of a CR over a fair dataset 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 CR mining (referred to as CR-CS in the rest of this article). The algorithm integrates association rule mining with a causal relationship test based on cohort studies. In the following, we firstly discuss two antimonotone properties for efficient generation of candidate CRs, and we then discuss the selection of control variables for building a fair dataset. Finally, we introduce the details of detecting CRs from the candidate CRs.

4.1. Antimonotone Properties

Antimonotone properties are at the core of efficient association rule mining. For example, a well-known antimonotone 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 antimonotone properties that we will apply to candidate CR pruning.

In the following discussions, we say that rule $px \to z$ is more specific than rule $p \to z$, or $p \to z$ is more general than $px \to z$. 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.

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 CRs/relationships, more specific candidate CRs 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

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

Input: Dataset 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 dataset 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 subpatterns respectively // Observation 2
end while
output R_C

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 $p \rightarrow z$ and $px \rightarrow z$ cover the same set of records. There will be the same fair dataset for both rules. Therefore, if $p \rightarrow z$ is a CR, so is $px \rightarrow z$. If $p \rightarrow z$ is not a CR, 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 $\text{supp}(px) = \text{supp}(p)$, then $\text{supp}(pxy) = \text{supp}(py)$. Using the same reasoning as earlier, we conclude that rule $pxy \rightarrow z$ is redundant with respect to rule $px \rightarrow z$. □

Since there are two antimonotone properties in addition to the antimonotone property of support, it is efficient to use a level-wise algorithm like Apriori [Agrawal et al. 1996]. Both antimonotone properties 1 and 2 can be used in the same way as the antimonotone property of support.

4.2. Control Variables

The set of control variables determines the size of a fair dataset. If the control variable set is large, the chance of finding a nonempty fair dataset 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, and 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 $= m$" $\rightarrow$ "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, as 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 nonexposure 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 nonexposure group may be empty, and 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 nonexposure 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 dataset, 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, as 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 dataset is empty.

Exclusive variables can be confounding variables, e.g., $P = thunder$ and $Q = storm$ may be mutually exclusive in a dataset 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 postprocessing, as they are strongly associated.

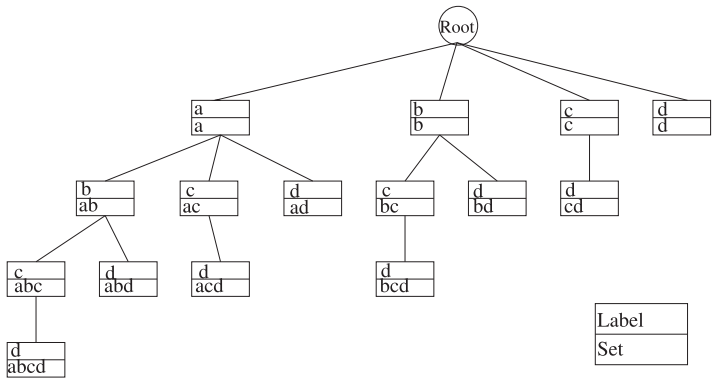

Fig. 1. An example prefix tree. a, b, c, d stand for nonzero values of variables $A, B, C,$ and D.

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 CRs. A node stores the prefix set of the 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 it easy to search, insert, and delete a node. This property also makes the counting of supports of value sets efficient. 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 either. This property avoids many unnecessary searches in the counting process. In the counting process, each node store two counts, e.g., supp(pz) and supp($p\neg z$). As a result, the contingency table of $p \rightarrow z$ is determined.

All supersets with the same prefix are stored under a parent node which stores the prefix. When the parent node is removed, so are all supersets. This suits the forward pruning for candidate CR 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 a link will facilitate the pruning by antimonotone property 2.

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

4.4. Causal Rule Detection

The causal rule detection process involves three steps, as discussed next.

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 $y \rightarrow z$. The

Function 1 Create a fair dataset for rule $p \rightarrow z$
Input: Dataset D, rule $p \rightarrow z$, and control variable set C
Output: a fair dataset for rule $p \rightarrow z$, D_f

```
1: find the covering set of c(C = 1), D_c
2: split D_c into D_cp and D_c¬p // D_cp contains value p and D_c¬p does not
3: let D_f = ∅
4: for each record t_i in D_cp // assuming |D_cp| ≤ |D_c¬p|. If not, swap D_cp and D_c¬p. do
5:    for each record t_j in D_c¬p do
6:       if t_i and t_j are matched w.r.t with respect to 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
```

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

4.4.2. Creating a Fair Dataset. We select the samples from the given dataset D to get the fair dataset 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 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 is more than one matched record, choose one randomly. Add the pair of records to the fair dataset. 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 $p \rightarrow z$ is a CR, we firstly follow Definition 3.7 to calculate the odds ratio of the rule on its fair dataset 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 $p \rightarrow z$ is a CR. 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 fair dataset. 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. Datasets and Parameters

To evaluate CR-CS, the proposed CR mining algorithm, 24 synthetic datasets and 8 frequently used public datasets were employed in the experiments. A summary of the datasets is given in Table I. The number of variables in the table refers to the number of predictor variables in a dataset. All predictor variables and the response variable are binary variables, with values of 1 or 0 indicating the presence or absence of an

Table I. A Summary of Datasets Used in Experiments

Name	Record (#)	Variables (#)	Distributions	Ground Truth
V20-2K	2,000	19	41.9% & 58.1%	7
V20-5K	5,000	19	41.9% & 58.1%	7
V20-10K	10,000	19	41.9% & 58.1%	7
V40-2K	2,000	39	37.6% & 62.4%	7
V40-5K	5,000	39	37.6% & 62.4%	7
V40-10K	10,000	39	37.6% & 62.4%	7
V60-2K	2,000	59	52.5% & 47.5%	7
V60-5K	5,000	59	52.5% & 47.5%	7
V60-10K	10,000	59	52.5% & 47.5%	7
V80-2K	2,000	79	50.6% & 49.4%	8
V80-5K	5,000	79	50.6% & 49.4%	8
V80-10K	10,000	79	50.6% & 49.4%	8
V100-2K	2,000	99	48.1% & 51.9%	6
V100-5K	5,000	99	48.1% & 51.9%	6
V100-10K	10,000	99	48.1% & 51.9%	6
V200-10K	10,000	199	19.8% & 80.2%	20
V400-10K	10,000	399	19.8% & 80.2%	40
V600-10K	10,000	599	19.8% & 80.2%	60
V800-10K	10,000	799	19.8% & 80.2%	80
V1000-10K	10,000	999	19.8% & 80.2%	100
Name	Records (#)	Variables (#)	Distributions	Known Combined Rules
V8-2K	2,000	7	45.1% & 54.9%	2
V12-2K	2,000	11	72.1% & 27.9%	3
V16-2K	2,000	15	45.2% & 54.8%	3
V20-2K-cmb	2,000	19	55.6% & 44.4%	4
German	1,000	60	30.0% & 70.0%	—
Kr-vs-kp	3,196	74	47.8% & 52.2%	—
Mushroom	8,124	215	48.2% & 51.8%	—
Tic-tac	958	27	34.7% & 65.3%	—
Adult	48,842	99	23.9% & 76.1%	—
Hypothyroid	3,163	51	4.8% & 95.2%	—
Sick	2,800	58	6.1% & 93.9%	—
Census income	299,285	495	6.2% & 93.8%	—

attribute correspondingly. The class variable in each of the 8 public datasets 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 datasets. For the synthetic datasets, the ground truth column represents the number of true single causes and known combined causes, each consisting of two predictor variables.

The first 15 synthetic datasets in Table I were used to evaluate the performance of CR-CS in finding single CRs in comparison with the Bayesian network based 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 Bayesian network using the "simulate data from IM" template. The conditional probability table was also randomly assigned, which will be used to simulate the data. The datasets 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 variables satisfy the constraints in the conditional probability tables. We selected a node in each of the Bayesian networks as the fixed target for running the algorithms.

The next five synthetic datasets (V200-10K, . . . , V1000-10K) were used to assess the efficiency of the algorithms. To generate those large datasets with a fixed proportion of nodes being the parents of the target node (which is not practical with TETRAD), we firstly draw simple Bayesian networks where some predictor variables are parents of the response variable and some are not. We then use logistic regression to simulate the datasets for those Bayesian networks. The total number of causes in each Bayesian network is given in Table I.

Meanwhile, the four datasets 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 datasets have been generated with the following procedure. We firstly generate a dataset for a random Bayesian network 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 Bayesian network 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 dataset due to the combinations of noncausal 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 datasets, Hypothyroid and Sick are two medical datasets, 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 dataset is an extraction of the U.S. 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 dataset. These three datasets 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) dataset was also sourced from the UCI Machine Learning Repository. We combined the training and test datasets and then sampled 50K, 100K, 150K, 200K, and 250K records for the experiments. Continuous attributes have been removed. The dataset and the last five synthetic datasets (with 10K records) were used to assess the efficiency of CR-CS (Section 5.6). Other real-world datasets are also from the 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 dataset for 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 datasets and to 95% for real-world datasets considering the noises in the real-world datasets.

5.2. Causal Rules Versus Association Rules

CRs 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 (whereas a CR 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 CR mining, we compared the results obtained by CR-CS to the results of various types of association

Table II. Comparison of the Numbers of Association Rules (AR), Nonredundant Rules (NRR), Optimal Rules (OR), and Causal Rules (CR); Many Association and Interesting Rules Are Not Causal

	AR (#)	NRR (#)	OR (#)	CR (#)
Adult	3,108	2,863	976	46
Hypothyroid	39,476	17,692	3,237	30
Sick	56,183	28,698	3,917	21

rule mining. From Table II, the number of CRs is significantly smaller than the numbers of other types of (association) rules, including association rules [Agrawal et al. 1996], nonredundant rules [Zaki 2004], and optimal rules [Li 2006]. Associations are measured by the odds ratio defined in Definition 3.1, and their significance is tested using the method discussed in Section 3.2; all methods used the same minimum local support. The maximum length of rules is 4.

The number of CRs obtained from a dataset is very small. They may not be enough for classification, as not every record in the data is covered by a CR. However, they are more reliable relationships, as each CR is tested by the cohort study in data.

Most discovered CRs (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 Versus Findings of Other Causal Discovery Methods

To evaluate the performance of CR-CS, we conducted a set of experiments with the first 15 synthetic datasets and 3 real-world datasets (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 datasets. PC-Select [Colombo et al. 2014] is a local causal discovery method that finds all 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 an 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, a 95% confidence level was used. With our method (CR-CS), since there are small variations in the CRs discovered in different runs due to random selection of matched pairs when a record has multiple matches, in the experiments with one dataset, we generated CRs, 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 datasets with different numbers 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 datasets.

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 15 datasets. Table IV shows the pairwise test results for

Table III. Performance of CCC, CCU, PC-Select, and CR-CS in Finding Single Rules with Synthetic Datasets

	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

Note: P, R, and F_1 represent precision, recall, and F_1-measure, respectively.

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	—

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 better than the other. However, note that PC-Select is only suitable for datasets with small number of nodes or sparse datasets with a small number of causes of the target. It took more than 2 hours for PC-Select to complete when it was applied to the synthetic dataset with 100 nodes and 20 causes of the target, and it failed to return results for the dataset with 120 nodes with 26 causes of the target within 24 hours.

To evaluate the ability of CR-CS in recovering combined CRs, we use synthetic datasets with known combined rules as described in Section 5.1. We applied CR-CS to the four datasets 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 datasets, 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 dataset, as shown in Table V, CR-CS, CCC, and CCU discovered a similar number of rules, whereas PC-Select found a relatively small number of rules.

When we look into the rules discovered by these methods, they are quite different. In Table VI, we list the most similar and dissimilar rule groups found in the Adult dataset using CR-CS and the other methods. We can see that overall, CR-CS and PC-Select obtained similar results, whereas 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 Datasets

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

Table VI. Similar and Dissimilar Causal Rule Groups Discovered by CR-CS and the Other Methods in the Adult Dataset

Causal Rules	CR-CS	CCC	CCU	PC-Select
Education=doctorate $\rightarrow >$ 50K	√	√	√	√
Education=masters $\rightarrow >$ 50K	√	√	√	√
Education=bachelors $\rightarrow >$ 50K	√	√	√	√
Education=prof-School $\rightarrow >$ 50K	√	√	√	√
Education=some-college $\rightarrow \leq$ 50K		√	√	
Education=HS-grad $\rightarrow \leq$ 50K	√	√	√	
Education=12th $\rightarrow \leq$ 50K	√	√	√	
Education=11th $\rightarrow \leq$ 50K	√	√	√	√
Education=10th $\rightarrow \leq$ 50K	√	√	√	√
Education=9th $\rightarrow \leq$ 50K	√	√	√	√
Education=7-8th $\rightarrow \leq$ 50K	√	√	√	√
Education=5-6th $\rightarrow \leq$ 50K	√	√	√	
Education=1-4th $\rightarrow \leq$ 50K		√	√	
Education=preschool $\rightarrow \leq$ 50K		√		
Occupation=exec-managerial $\rightarrow >$ 50K	√			√
Occupation=prof-specialty $\rightarrow >$ 50K	√			
Occupation=tech-support $\rightarrow >$ 50K	√	√	√	
Occupation=sales $\rightarrow >$ 50K	√			
Occupation=handlers-cleaners $\rightarrow \leq$ 50K	√			√
Occupation=machine-op-inspct $\rightarrow \leq$ 50K	√			
Occupation=adm-clerical $\rightarrow \leq$ 50K	√			
Occupation=other-service $\rightarrow \leq$ 50K	√			√
Occupation=farming-fishing $\rightarrow \leq$ 50K	√			√
Occupation=transport-moving $\rightarrow \leq$ 50K	√			
Occupation=craft-repair $\rightarrow \leq$ 50K	√			
Workclass=sel-emp-inc $\rightarrow >$ 50K	√	√		√
Workclass=sel-emp-not-inc $\rightarrow >$ 50K		√	√	
Workclass=federal-gov $\rightarrow >$ 50K	√	√	√	√
Workclass=state-gov $\rightarrow >$ 50K	√			
Workclass=local-gov $\rightarrow >$ 50K	√	√	√	
Workclass=private $\rightarrow \leq$ 50K		√	√	
Native Country=USA $>$ 50K	√	√	√	
Native Country=various countries	1	22	17	2
Education=Some-college & Workclass=Private $\rightarrow \leq$ 50K	√			

Note: 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, and 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.

Table VII. Number of Combined Causal Rules Discovered by CR-CS in Real-World Datasets

	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
Mushroom	26	61	0	1
Sick	13	7	1	0
Tic-tac	8	30	3	0

Table VIII. Some Combined Causal Rules in the Mushrooms Dataset

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

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 income. We see that people with higher education have a better chance for a high salary, i.e., doctorate, masters, bachelors, and professional school (prof-School). In contrast, people with lower education more likely receive a low salary, i.e., 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 CRs, but they have been missed by CCC and CCU. PC-Select, however, found fewer numbers of rules in this group (Occupation), and the rules found by it are reasonable. On the other hand, 22 rules related to the Native Country attributes are discovered by CCC and 17 rules are discovered by CCU, but only 1 rule is discovered 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 CRs.

The combined CR 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 CR.

To investigate the number of combined CRs in real-world cases, we run CR-CS for eight real-world datasets with up to level-4 rules. Table VII shows the number of single and combined CRs discovered by CR-CS with a 95% confidence level. We can see that the combined CRs at levels 3 and 4 are rare, but CR-CS found a number of combined CRs 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 dataset, 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

Table IX. Numbers of Causal Rules of Different Runs and Frequent Causal Rules

Fair Dataset	1	2	3	Frequent
Adult	46	46	45	46
Hypothyroid	31	30	30	30
Sick	21	20	21	21

Table X. Results of CR-CS Using Different Matching Methods

Dataset	Exact Matching	Jaccard Distance
Adult	46	46
Hypothyroid	30	31
Sick	21	22

brightly colored mushrooms are poisonous. These combined CRs 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 dataset is subject to selection bias. Usually, there are significantly more exposed cases than nonexposed cases, so the data distribution is often skewed for the exposure and nonexposure conditions. When we choose pairs of matched records to form a fair dataset, we pick up one record from the exposure group and find a matched record from the nonexposure group. In this process, the values of the response variable are blinded. When there is 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 dataset is affected by the random selection. This will cause misses or false discoveries of CRs. This situation is the same as the real-world sample process, which is subject to sampling bias.

To reduce the impact of selection bias, we run the method on a dataset multiple times and select consistent rules in multiple CR sets as the final CRs. The variance is not big, and the causal discovery is quite stable. The numbers of CRs from different runs and the rules supported by two CR sets are listed in Table IX. On a large dataset, i.e., the Adult dataset, the change of rules between different runs is very small. There is only one rule difference in the three runs. Even in a small dataset, such as the Sick dataset, nearly 90% of rules are consistent over three runs.

5.5. Results Obtained Using Different Matching Methods

As described in Section 3.4 (Definition 3.4), when creating a fair dataset, different similarity measures can be used for finding matched pairs of records. In the experiments described so far, exact matching has been used. 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 datasets with exact matching and the matching using Jaccard distance.

5.6. Efficiency

To test the time efficiency of CR-CS, we applied it to the Census Income (KDD) dataset and the last five synthetic datasets (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.

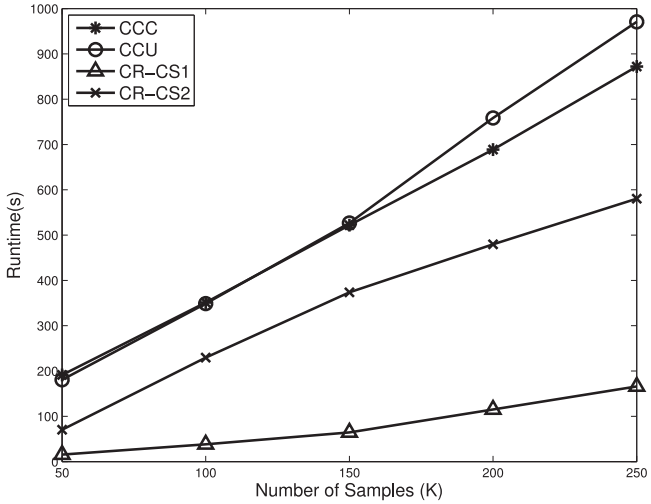

Fig. 2. Scalability with respect to number of records. Note: PC-Select is not included, as it did not return results after 2 hours of execution.

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. 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 (a quad-core CPU of 3.4GHz and 16GB 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 2 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 noneligible 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 was more scalable than CCC and CCU, whereas CR-CS2 was much slower when the number of attributes became big, as the number of association rules increased significantly with the increase in the number of attributes, leading to additional time for testing CRs. Although PC-Select can achieve high quality of causal discovery (see Table III), from 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 article, we have proposed the concept of CRs and have developed a method to find CRs from observational data by integrating association rule mining with retrospective cohort studies. Through the integration, our method has been able to take the

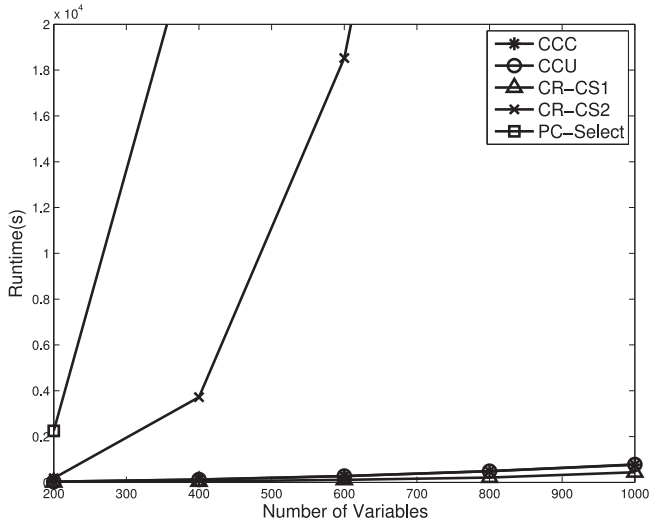

Fig. 3. Scalability with respect to number of attributes.

advantage of the high efficiency of association rule mining to produce candidate causal relationships from large datasets and then to utilise the idea of cohort studies to obtain reliable CRs based on the candidates. The validity of the definition of CRs has been justified to be consistent with the potential outcome model. Experimental results have shown that the proposed method is able to find more reasonable causal relationships compared to existing causal discovery methods. Moreover, our method was able to find causes consisting of combined variables, which cannot be uncovered by other existing methods. We have shown that the method is faster than 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 datasets. With the proposed method, selection of the control variable set is key to discovering quality CRs. The validation of the control variable set in real-world applications will ensure the quality of CRs discovered.

The proposed CR mining method and 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 focus on how they complement each other, exploring integrated methods for efficient and quality causal relationship discovery.

REFERENCES

R. Agrawal, T. Imieliński, and A. Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (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*. American Association for Artificial Intelligence, Menlo Park, CA, 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, 171–234.

K. Bache and M. Lichman. 2013. UCI Machine Learning Repository. Retrieved November 4, 2015, from http://archive.ics.uci.edu/ml.

C. C. Blackmore and P. Cummings. 2004. Observational studies in radiology. *American Journal of Roentgenology* 183, 5, 1203–1208.

C. Borgelt. 2003. Efficient implementations of Apriori and Eclat. In *Proceedings of the 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 the ACM SIGMOD International Conference on Management of Data (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, 1287–1330.

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

J. Concato, N. Shah, and R. I. Horwitz. 2000. Randomized, controlled, trials, observational studies, and the hierarchy of research design. *New England Journal of Medicine* 342, 25, 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, 203–224.

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

A. M. Euser, C. Zoccali, K. Jager, and F. W. Dekker. 2009. Cohort studies: Prospective versus retrospective. *Nephron Clinical Practice* 113, 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, 1–38.

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

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, 1–26.

S. Kleinberg and G. Hripcsak. 2011. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics* 44, 6, 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, Los Alamitos, CA, 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, 610–626.

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

J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, and B. Sun. 2013. Mining causal association rules. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW'13)*. IEEE, Los Alamitos, CA, 114–123.

B. Liu, W. Hsu, and Y. Ma. 1998. Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (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)*. 314–323.

S. L. Morgan and C. 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*. Cambridge University Press.

J. P. Pellet. 2008. Using Markov blankets for causal structure learning. *Journal of Machine Learning Research* 9, 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, MA.

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

J. W. Song and K. C. Chung. 2010. Observational studies: Cohort and case-control studies. *Plastic and Reconstructive Surgery* 126, 6, 2234–2242.

P. Spirtes. 2010. Introduction to causal inference. *Journal of Machine Learning Research* 11, 1643–1662.

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

H. O. Stolberg, G. Norman, and I. Trop. 2004. Randomized controlled trials. *American Journal of Roentgenology* 183, 6, 1539–1544.

E. A. Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25, 1, 1–21.

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

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

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

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

Received August 2014; revised January 2015; accepted March 2015