Discovery of Causal Rules Using Partial Association

Zhou Jin[1,3]
[1]*Department of Automation*
University of Science and Technology
of China
Hefei 230026, China
manjinzhou@gmail.com

Jiuyong Li[2], Lin Liu[2] and Thuc Duy Le[2]
[2]*School of Computer and Information Science*
University of South Australia
Mawson Lakes, SA 5095, Australia
{jiuyong.li, lin.liu}@unisa.edu.au
thuc_duy.le@mymail.unisa.edu.au

Bingyu Sun[3], Rujing Wang[3]
[3]*Institute of Intelligent Machines*[3]
Chinese Academy of Sciences
Hefei 230031, China
{bysun, rjwang}@iim.ac.cn

***Abstract*—Discovering causal relationships in large databases of observational data is challenging. The pioneering work in this area was rooted in the theory of Bayesian network (BN) learning, which however, is a NP-complete problem. Hence several constraint-based algorithms have been developed to efficiently discover causations in large databases. These methods usually use the idea of BN learning, directly or indirectly, and are focused on causal relationships with single cause variables. In this paper, we propose an approach to mine causal rules in large databases of binary variables. Our method expands the scope of causality discovery to causal relationships with multiple cause variables, and we utilise partial association tests to exclude noncausal associations, to ensure the high reliability of discovered causal rules. Furthermore an efficient algorithm is designed for the tests in large databases. We assess the method with a set of real-world diagnostic data. The results show that our method can effectively discover interesting causal rules in large databases.**

***Keywords*-data mining; causality; partial association; causal rule**

I. Introduction

The discovery of causal relationships between variables is important in revealing intricate interactions among the components of a system or process. Causal relationships are more powerful than correlative relationships, in that causal relationships indicate not only that the variables are related, but also how varying a cause variable is likely to induce a change of an effect variable, therefore they are more useful in prediction and reasoning. For example, with identified causal relationships, we can predict potential consequences (effects) before actually carrying out any actions (causes), which is extremely useful in preventing erroneous decision or policy making. In medical science, causal relationships help us understand the causes of diseases and deliver better diagnosis and cures for diseases. In brief, the studies of causality will fully display their values in various areas, such as economics, physical, behavioral, social and biological sciences.

The common approach to identify causal relationships is to conduct randomised controlled experiments, which unfortunately, is often expensive and infeasible with large number of variables. Therefore much attention has been focused on discovering causal relationships from observational data, particularly with the rapidly increased accumulation of large volume of observational data.

Discovering causal relationships in large observational databases, however, is also a challenging task. Firstly, there is not a generally acceptable definition for causal relationships. Causality is more a philosophical phenomenon, and it may have different meanings in different areas. This makes it difficult to expound causality in a unified form. Pearl [1] proposed a framework that connected conditional independence with causal structures, based on which some methods have been developed to define and identify the causality. However, it is still far from satisfaction with discovering causal relationships effectively and efficiently from large databases. Secondly, the computational cost for the discovery is high. Causal relationships are typically represented by probabilistic dependence.

Probabilistic causality has been proposed and discussed in some early literature [2], [3], [4], [5]. More recently, Bayesian networks [6] have emerged as a major platform for discovering causal structures [7], [8], [9], [10], [11], [12], [13], given the methods they have provided for representing, inferring and learning probabilistic independence among variables. However, discovering complete causal models with Bayesian network learning is a NP-complete problem [14]. Constraint-based approaches which do not search for a general Bayesian network are more efficient. Currently, several constraint-based algorithms have been used to discover causal structures in large databases and have produced some good results [15], [16], [17], [18], [19]. These methods use observational data to determine conditional independence of variables and learn local causal structures. It is worth noting that these constraint-based methods apply the idea of Bayesian network learning directly or indirectly, with the goal of generating a directed acyclic graph (DAG) to represent the conditional independence between variables.

Although constraint-based approaches have shown promise with large data sets, they normally are designed to discover causal relationships with some fixed structures in a DAG, e.g. CCC [15], CCU [16] and Y structures [17]. Furthermore, they are not designed to discover *combined* *causal factors*. In practice, the combination of two or more cause variables may strengthen the degree of effects. Even when each variable individually does not cause any effect, their combination may do. For example, someone who has eaten raw fish and drunk cold milk separately may feel well, but he/she may feel sick after having them together. To the best of our knowledge, there is no previous work on discovering combined causal relationships with constraint-based causal relationship discovery.

We should note that it is insufficient to discover causal relationships in observational data only. Ultimately, the identified relationships have to be validated with controlled experiments. However, it is sufficient to exclude noncausal relationships based on the discoveries from data. Causal relationship discovery in data is to find a short list of rules that are most likely causal. These causal rules represent a small set of statistically reliable relationships that are likely to embed cause and effect relationships. This distinguishes the causal rule discovery from the normal rule discovery. For example, association rule mining normally finds a large set of rules, many of which are redundant and spurious, but causality has not been used as an interestingness criterion before. However, we argue that it would be inefficient to find causal rules in a large collection of association rules as a secondary discovery process, and new approaches are required for causal rule discovery.

In this paper, we propose a general approach to causal relationship discovery using association and partial association. In comparison with previous work, we have made three main contributions. Firstly, the approach does not rely on Bayesian network structures, and it does not define causal relationships in such a way that the relationships are restricted to some specific structures in a DAG. Our method directly searches for potential causal relationships among variables based on association and partial association, which also have strong theoretical support for causal relationship discovery. Secondly, our approach can discover causal relationships with a single cause factor (variable), as well as the relationships with combined cause variables, from observational data. Existing constraint-based discovery methods do not consider combined cause variables. Furthermore, with Bayesian network learning based causal relationship discovery, domain experts are needed to create the structure of a network to identify combined causal variables if each of the individual variables of the combination is independent of the outcome (effect) variable. Thirdly, we have designed an efficient algorithm for causal rule discovery. The algorithm reduces the memory space requirement considerably to make it computational feasible.

The rest of the paper is organised as follows. In Section II, we give the relevant definitions and present the problem statement. Section III describes our algorithm for causal rule discovery and discusses its time complexity. Then in Section IV we present the implementation of the algorithm and the experimental results to show the validity of the algorithm. In Section V we conclude the paper and suggest future work.

II. Definitions and Problem Statement

In this section, we firstly define the notation to be used in the paper and an informal definition of causal rules (Section II-A). We then define the concepts and formulas for presenting our causal rule discovery method (Sections II-B and II-C). Finally, we give the formal definition of causal rules in Section II-D, and discovering such rules is the goal of this paper.

A. Notation and an Informal Definition

Individual random variables are represented using upper case letters, e.g. X and Y. We use bold-faced upper case letters, e.g. $\mathbf{X}$, to denote a set of random variables. Lower case letters, e.g. x and y, denote the value assignments to variables X and Y respectively. We use the symbol "\" to denote the set difference operator, e.g. $\mathbf{X}\backslash\{X_i\}$ represents the set of all variables in $\mathbf{X}$ except for X_i.

Particularly in this paper, we use the bold-faced upper case letter, $\mathbf{V}$, where $\mathbf{V} = \{V_1, V_2, \ldots, V_m\}$ to represent a set of binary predictive variables, and use letter Y to represent a binary outcome (target or effect) variable. Suppose that value 1 indicates yes, and value 0 indicates no, an example data set for 6 predictive variables A, B, C, D, E, F and one target variable Y is shown in Table I, where the last column shows the number of repeats of a sample in the data set.

Table I
An example data set for six predicate variables and the target variable

A	B	C	D	E	F	Y	#Repeats
1	1	1	1	1	1	1	14
1	0	1	1	1	1	1	8
1	1	0	1	0	1	1	15
0	1	1	1	1	1	1	8
0	1	0	0	0	0	0	5
0	0	0	0	1	0	1	6
1	0	0	0	0	1	0	4
1	0	1	1	1	0	0	3
0	1	0	1	1	0	0	3
0	1	0	0	1	0	0	5

Let $\mathbf{X} \subset \mathbf{V}$, a causal rule $\mathbf{X} \rightarrow Y$ indicates that: 1) variables in $\mathbf{X}$ are associated with Y, and 2) the association remains given any $\mathbf{Z} \subset \mathbf{V}$ and $\mathbf{X} \cap \mathbf{Z} = \emptyset$. In other words, the association is persistent.

The justification of the above (informal) definition of a causal rule is as follows. If two variables are not associated, they cannot form a cause and effect relationship. If two variables (X and Y) are associated, but the association disappears when a third variable Z is given, this means that: both X and Y are caused by Z; or X causes Z and Z causes Y. In either case, X is not a direct cause of Y. As mentioned before, it is insufficient to validate causal relationships in data only, but it is sufficient to exclude noncausal relationships with the findings in data. Hence the idea is to exclude the noncausal relationships based on the definition of causal rules, and keep the remainder as potential causal relationships. The same strategy is employed in [16]. In Section II-D, we will define causal rules formally.

B. Association

In this section, we firstly review the concept of the association of two variables, by following the standard statistical definition. Then we discuss the case with more than two variables, and provide several new definitions, which will be used in developing and presenting our approach to causal rule discovery.

Table II provides the general form of a contingency table for two binary variables X and Y. In the table, n_{ij} $(i, j \in \{0, 1\})$ represents the counts or frequencies of $X = i$ and $Y = j$ in the data set, and $n_{i.} = n_{i1} + n_{i2}$, $n_{.j} = n_{1j} + n_{2j}$.

Table II
A 2×2 CONTINGENCY TABLE

	Y=1	Y=0	Total
X=1	n_{11}	n_{12}	$n_{1.}$
X=0	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

If X and Y are independent, the count values will be distributed in the four cells (the cells that are not totals) based on the independent principle. In this case, the expected values of the four cells are: $E(n_{11}) = \frac{n_{1.}}{n} * \frac{n_{.1}}{n} * n = \frac{n_{1.} * n_{.1}}{n}$, $E(n_{12}) = \frac{n_{1.} * n_{.2}}{n}$, $E(n_{21}) = \frac{n_{2.} * n_{.1}}{n}$, and $E(n_{22}) = \frac{n_{2.} * n_{.2}}{n}$.

The Chi-square statistic (Equation (1)) indicates the deviation of the observed values from the expected values when the two variables are independent. The higher the Chi-square statistic value, the more the two variables deviate from the independence. Normally, when $\chi^2 > 3.84$, we have 95% confidence to reject the hypothesis that the two variables are independent and consider that they are associated.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})} \qquad (1)$$

When we have more than two variables, e.g. three variables X_p, X_q, and Y, following the standard statistics method, to compute the Chi-square statistic for the three variables, a three way contingency table is constructed with 8 cells: $\{(X_p = 1, X_q = 1, Y = 1), (X_p = 1, X_q = 1, Y = 0), (X_p = 1, X_q = 0, Y = 0), \ldots, (X_p = 0, X_q = 0, Y = 0)\}$. One major problem with using such a multi-way contingency table is that it is very difficult to obtain a reliable Chi-square estimation for three or more variables. To obtain a reliable Chi-square statistic, the count value in each cell has to be 5 or larger, and this is often violated when the number of cells becomes large. For example, with the 3 binary variables representing gender, having breast cancer and having prostate cancer, the cells with “gender=female, prostate cancer=yes” and “gender=male, breast cancer=yes” may each have a count close to zero, as in reality it is highly unlikely for a female to have prostate cancer or for a male to have breast cancer.

Therefore for our purpose of discovering causal rules, we propose to use a simplified contingency table, as illustrated in Table III with three variables X_p and X_q (two predicative variables) and Y (the target variable). We call the simplified contingency table *contingency table with combined variables*. In the table the predictive variables are considered to be combined into one variable, and all the cases of the predictive variables having zero values (negative outcomes) are considered as “others”.

Table III
THE GENERAL FORM OF A CONTINGENCY TABLE WITH COMBINED VARIABLES

	Y=1	Y=0	Total
$X_p = 1, X_q = 1$	n_{11}	n_{12}	$n_{1.}$
others	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

It is important to note that such a contingency table captures the semantics of our intended causal relationships better. In most cases of investigating causal relationships, we are more interested in the positive outcomes (value 1) of binary variables. For example, “smoking = yes” is more meaningful to doctors than “smoking = no”. We would have to introduce lots of redundancies if we considered all cells with close to zero counts. We call the association identified by using such a simplified contingency table *positive association* (Definition 1).

Definition 1. (POSITIVE ASSOCIATION AND ZERO ASSOCIATION) Given an attribute set (predictive variables) **X**, an outcome attribute (target variable) Y, and M (the contingency table with combined variables for **X**), **X** is positively associated with Y if $\chi^2_M \geq \chi^2_p$, where χ^2_M is the Chi-square statistic obtained from M using Equation (1), and χ^2_p is the Chi-square value corresponding to the specified p value. When $\chi^2_M < \chi^2_p$, **X** and Y are said to be zero-associated.

With the definition, a commonly used $p-$value is 0.05, and $\chi^2_{0.05} = 3.84$.

We consider only positive associations since in most cases, e.g. investigating the causes of a diseases, only positive outcomes are of interest. In practice, the affirmative statements are of interest to many research areas. For example, medical researchers would like to know the direct associations between different symptoms and a disease. However surveying for the relationships between the absence of a symptom and the disease or the absence of a symptom and the absence of the disease have little interest to most users. Furthermore, we only consider positive associations in order to reduce the search space. Empirically, a positive association does not mean causality, but it is logical to say that if there is no positive association between two variables, there will be no causality between them.

Definition 2. (SUPPORT FOR POSITIVE ASSOCIATION) In the contingency table with combined variables for attribute set **X** and outcome attribute Y, n_{11} denotes the counts of all positive assignments to the variables in **X**, and it is called the support for the positive association between **X** and Y.

The support for positive association, n_{11}, indicates the number of cases supporting the positive association between the predictive variables and the outcome variable. If n_{11} is small, the association may not be of interest to users since the relationship explains only very few cases. It is sensible to set a threshold on the support for positive association, to only keep positive associations with high supports, thus the following definition.

Definition 3. (FREQUENT POSITIVE ASSOCIATION) Assume that **X** and Y are positively associated, the association is frequent if $n_{11} > n^*$, where n^* is the minimum support threshold.

The value of n^* can be set to be a constant or relative to $n_{.1}$. The rate $n^*/n_{.1}$ indicates the percentage of cases being explained in $n_{.1}$. Consider the example in Table IV, where $n_{11} = 27$ and $n_{.1} = 30$. It is easy to see that $\{B, E\}$ has frequent positive associative relationship with Y if we set $n^*/n_{.1} = 50\%$, i.e. $n^* = 15$.

Table IV
AN EXAMPLE CONTINGENCY TABLE WITH COMBINED VARIABLES

	Y=1	Y=0	Total
B=1, E=1	27	3	30
others	29	12	41
Total	56	15	71

Frequent positive associations make it possible for the causal rule discovery to use frequency pruning in association rule mining, which could improve the efficiency of the algorithm. Based on the definition of frequent positive association, we have the following definition.

Definition 4. (FREQUENT ASSOCIATION RULE) $\mathbf{X} \rightarrow Y$ is a frequent association rule if: 1) **X** and Y are positively associated; and 2) The support of the positive association $\mathbf{X} \rightarrow Y$ is greater than the minimum support threshold.

C. Partial Association

Positive associations are a starting point for identifying causal relationships. However, a positive association between two variables may disappear when other variables influence the relationship. Causal relationships, however, are more persistent and precise than positive associations. Causal relationships indicate the essential relationships that will not be affected by other factors. In other words, causal relationships are persistent given all other observable variables.

A positive association may imply a causal relationship, but not every association is a causal relationship. If a positive association is not persistent, it will be unwise to use it for predicting direct causal relationships between variables. For example, with three binary variables: having high blood pressure (HB), smoking (S) and having cancer (C), S and C may have positive association, but when we take the factor of HB into consideration, we cannot establish the causal relationship that S is the cause of C, as it could be possible that HB is a common cause of both S and C. If we use associations only, we may miss the real causes. As a result, it is imperative to remove those noncausal associations effectively and efficiently. Our idea is to exclude noncausal relationships from associative relationships and produce a small set of potential causal relationships.

To filter out noncausal relationships from a set of positive associations, we may conduct randomised controlled experiments. However, these experiments are usually costly and sometimes impossible or irrational to perform. On the other hand, when only having observational data, partial association tests can be employed to remove noncausal relationships.

Assume that a set of predictive variables, **X** and an outcome variable, Y, are positively associated. When taking another set of variables **Z** into account, we can test the partial association between **X** and Y. If the partial association is zero, it means the association between **X** and Y is not persistent and it is obstructed by other variables. In causality discovery, zero partial association can be explained by considering **Z** as a common cause of both **X** and Y, or **X** causes **Z**, which then causes Y. So we are more interested in nonzero partial associations which indicate constant relationships and potential causal relationships. Therefore we should determine that the association between **X** and Y is persistent using partial association test before we are able to conclude that $\mathbf{X} \rightarrow Y$ is a causal rule. In other words, we need to test the partial association between **X** and Y given any $\mathbf{Z} \subset \mathbf{V}$ and $\mathbf{X} \cap \mathbf{Z} = \emptyset$.

There are some existing methods for testing partial associations, but no all-powerful tests are well-known. Mantel and Haenszel [20] proposed a refined method for testing partial association between two variables I and J with a three-way contingency table. It has the power against alternatives no matter the sign of the deviation varies or not. Furthermore, it has been proved to be valid when the data set is large (but the individual counting number is small), so it is suitable for discovering causal relationships in large data sets. Birch [21] showed that the Mantel-Haenszel test was optimal for testing against alternatives when the strength of partial associations keep constant. Therefore, the Mantel-Haenszel test is a good tool for testing partial associations for the purpose of causal relationship discovery.

Assume that I and J are two binary variables, and K has t possible values, $k_1 \ldots k_t$. To test the partial association of I and J given K, let n_{ijk} be the number of $I = i$, $J = j$ and $K = k$. A dot in the subscript indicates the sum of all possible values i, j or k. Based on the basic idea of Mantel-Haenszel test, the partial association, $PA(I, J, K)$, between I and J can be computed using the following equation.

$$PA(I,J,K) = (|\sum_k \frac{n_{11k}n_{00k} - n_{10k}n_{01k}}{n_{..k}}| - \frac{1}{2})^2 / \sum_k \frac{n_{1.k}n_{.10k}n_{.0k}n_{0.k}}{n^2_{..k}(n_{..k}-1)} \tag{2}$$

Based on this equation, we can define nonzero partial association as follows.

Definition 5. (Nonzero Partial Association) Let $\alpha \in [0, 1]$ be a significance level threshold for a partial association. There is a nonzero partial association between I and J given K if the following inequality holds, otherwise it is a zero partial association between I and J given K.

$$PA(I, J, K) \geq \chi^2_\alpha$$

From the definition, we see that a nonzero partial association means that the association between I and J is persistent. Otherwise, the association is considered to be non-persistent. We will search for all persistent associations for the purpose of causal relationship discovery.

In order to discover the causal rules included in the positive association rules of the predictive variables and the outcome variable in large data sets, we transplant the Mantel-Haenszel test to compute the partial associations, and to identify the nonzero partial associations.

Assume that $\mathbf{V} = \{V_1, V_2 \ldots V_m\}$ is a set of binary predictive variables, Y is a binary outcome variable. The number of all the possible combinations of the variables in $\mathbf{V}$ (i.e. the power set of $\mathbf{V}$ excluding the empty set and $\mathbf{V}$ itself) is $2^m - 2$ altogether. In order to find if a variable $V_i \in \mathbf{V}$ has a partial association with Y conditioned on any of the combinations, we need to test the partial association between V_i and Y given each of the combinations of $\mathbf{V} \backslash V_i$.

It is worthing noting that although some combinations may not have records of values in the data set, we see that on average, the number of the combinations will increase exponentially when the number of variables, i.e. m, increases. This will hinder the performance of partial association tests. In the next section, we will present a method as part of our algorithm to alleviate this problem.

Let n_{11k} be the number of V_i and Y both having value 1 given the k-th combination, n_{10k} be the number of V_i being 1 and Y being 0 given the k-th combination, n_{01k} be the number of V_i being 0 and Y being 1 given the k-th set of conditions, n_{00k} be the number of V_i and Y each having value 0 given the k-th combination. Then we can use Equation (2) to obtain the value of the partial association between V_i and Y. Finally based on Definition 5, we can test if V_i has a nonzero partial association with Y or not.

For example, referring to the example data set shown in Table I, let us test if $\{B, F\}$ has a partial association with Y. The contingency tables for testing the partial association between $\{B, F\}$ and Y are listed in Table V. To simplify the representation, in the following we use BF to represent the set $\{B, F\}$, and similar notation is used for a set with multiple elements. Also we use $[B = 1, F = 1]$ to indicate that only B and F are 1 and all other variables in $\mathbf{V}$ are zero.

Table V
Contingency tables for testing partial association between BF and Y given $K = ACDE$

	$ACDE$	
	$-$	$+$
$\widetilde{BF}$	3	8
BF	0	14

In order to investigate the partial association of BF and Y in this case, variable K is used to represent the possible combinations of the rest 4 variables, A, C, D, and E. Then we can obtain from the original data set a set of 2×2 three-way contingency tables, each corresponding to a possible value of K. Note that a contingency table that has all zeros in one row or one column can be eliminated because the contribution to the partial association value from such a table is going to be zero. As a result, the actual number of valid contingency tables could be smaller. In this example, there is only one table left, as shown in Table V, BF represents both B and F are ones, and $\widetilde{BF}$ represents either B or F is zero, or both are zero, where "-" and "+" represent $Y = 0$ and $Y = 1$ respectively. Then based on the contingency table and using Equation (2), the partial association between I and J given K is calculated as follows (note that in this case, $I = BF$, $J = Y$ and $K = ACDE$ and k has just one value):

$$\sum_k \frac{n_{11k}n_{00k} - n_{10k}n_{01k}}{n_{..k}} = 1.6800$$

$$\sum_k \frac{n_{1.k}n_{.10k}n_{.0k}n_{0.k}}{n^2_{..k}(n_{..k}-1)} = 0.6776$$

$$PA(BF, Y, K) = 2.418, \quad \alpha = 0.880$$

So we conclude that the partial association between BF and Y is not significant, thus $BF \rightarrow Y$ is not a causal rule.

With our causal rule discovery approach, we firstly find all the variables (or sets of variables) that are positively associated with Y. Then we perform partial association tests as illustrated above on these positive associations, and keep the nonzero partial associations, as they may imply causal relationships. This idea is formalised in the following section II-D.

D. *Causal rules*

Definition 6. (CAUSAL RULES) $\mathbf{X} \rightarrow Y$ is a causal rule if: 1) $\mathbf{X}$ and Y are positively associated; 2) The support of the association $\mathbf{X} \rightarrow Y$ is greater than the minimum support threshold; and 3) there exists a nonzero partial association between $\mathbf{X}$ and Y.

The support requirement makes it possible to use an efficient frequent pattern discovery algorithm as a base for causal rule discovery.

Definition 7. (REDUNDANT CAUSAL RULES) Assume that $\mathbf{X}\subset\mathbf{W}$, if $\mathbf{X} \rightarrow Y$ is a causal rule, rule $\mathbf{W} \rightarrow Y$ is redundant as it does not provide new information.

If $\mathbf{X} \rightarrow Y$ is a causal rule, $\mathbf{W} \rightarrow Y$ may or may not be a causal rule. However, in either case it is not of interest to us. Therefore, we can terminate a search for more complicated (longer) causal rules when a causal rule has been discovered. This reduces the search space.

If $\mathbf{X}$ and Y are positively associated, then $\mathbf{W}$, a superset of $\mathbf{X}$, may or may not positively associated with Y. If $\mathbf{W}$ and Y are not positively associated, they cannot have a causal relationship, and we should not test its partial association. If $\mathbf{W}$ and Y are associated, then the association attributes to the association between $\mathbf{X}$ and Y. Therefore, no matter $\mathbf{W}$ and Y are associated or not, the rule $\mathbf{W} \rightarrow Y$ is not of interest. This leads to the following condition for testing the causal rules with combined factors.

Definition 8. (CONDITION FOR TESTING CAUSAL RULES WITH COMBINED FACTORS) We only test a combined causal rule $\mathbf{XV} \rightarrow Y$ if $\mathbf{X}$ and Y have a zero association and $\mathbf{V}$ and Y have a zero association.

This definition will serve as a forward pruning criterion where all variables positively associated with the target variable are excluded from the combination of future search. This condition and the minimum support requirement make the search space manageable.

III. ALGORITHM

In this section, we describe our algorithm for causal rule discovery and discuss its complexity.

The algorithm (Algorithm 1) is designed based on our definition of causal rules (Definition 6), as well the conditions for removing redundant rules and testing rules with combined factors, as discussed in Section II-D.

Algorithm 1 Mining Causal Rules
Input: variable set U, data set T, support threshold s, significance threshold α, target Z.
Output: Causal Rules Set R.

```
set P=$\emptyset$, N=$\emptyset$, R=$\emptyset$, V=$U$
while (set V is not empty)
  Prune(V)
  for each variable X in V
    create contingency table
    if (X is frequent)
     Calculate $\chi^2_{X,Z}$
     if ($\chi^2_{X,Z}$>$\chi^2_{\alpha}$)
      insert X into P
      if PAssociation(X) is nonzero
        insert X into R
      end
     else insert X into N
     end
    end
  set V $\leftarrow$ Generate(N),P=$\emptyset$, N=$\emptyset$
return R
```

To save space, at above, we have omitted the details of the three key functions: Generate(), Prune() and PAssociation().

The algorithm makes multiple passes over the data. In the first pass, we count the support of all the individual variables together with the outcome variable and summarise them in the corresponding contingency tables to conduct the Chi-square tests. Positive associations and zero associations identified from the tests are kept. Associations with insufficient support will be eliminated directly. Next we use a well-designed method to do the partial association test with 2×2 contingency tables for each positive association. The records of data set will first be sorted and repeats are counted before the detection of partial associations. If a record contains the value of a variable, such as X, the rest variable with value 1 will be treated as a condition set. Conditioned on this set, one pass over the records will produce a 2×2 contingency table as illustrated in table V. Then the remaining tables for X will be generated according to the ordered records. Hence the nonzero partial association, by Definition 5, could be detected. The causal rules in current combined case can be determined from the nonzero partial associations. At the end of the pass, the zero associations found in the first step are combined for the next pass until no causal rules are found.

There are several favorable properties to improve the performance of the algorithm. The efficiency of the algorithm lies in the redundant rule property specified in Definition 7. Based on it, the pruning technique is used in generating combinations of two or more variables to reduce the search space. Suppose that the data set is a complete binary matrix and has no missing values, each column of the matrix indicates a predictive variable and the outcome variable is listed in the last column, QuickSort algorithm can be used to rank the records and merge the equivalence classes based on the features of the data, so the total number of the records will be reduced remarkably, especially in a database with large number of records. Regarding efficiency, not all the combinations are considered as a condition during the tests of partial associations. Instead, we only investigate the combinations appearing in the data which may be much smaller than the totality. As a result, the complexity of the algorithm is reduced.

To analyse the performance of the algorithm with respect to time and space complexity, and the number of passes over the data set, we denote the number of variables n, the number of records m. Suppose that there are l different records in the database. The complexity of the method is discussed based on the mining of causal rules in the form of $[X = 1, Y = 1] \rightarrow Z = 1$.

Consider the naive method for discovering the causal rule $XY \rightarrow Z$. The single variables are combined and the support is counted with $O(n)$ passes over the database. Each combination needs a Chi-square test to determine the positive association, which requires $O(n^2)$ passes. In the process of testing partial associations, a positive association will be examined conditioned on all other combinations. The total number of possible combinations is $O(2^n)$, so it needs to scan the database as many as $O(2^n n^2)$ times. To conclude, the passes over the database using naive method is $O(2^n n^2)$, and the time it takes is $O(m2^n n^2)$.

As mentioned above, we use QuickSort to rank the records and merge the equivalence classes, so the number of different records is reduced to l. The QuickSort algorithm takes O(mlgm) on average. Then the database is scanned for $O(1)$ times to generate positive association if $O(n^2)$ memory space is available. In the worst case, it costs $O(l)$ passes over the data set to test a partial association.

At the end, it takes $O(l^2)$ time and $O(l)$ passes over the database. In addition, it can be proved that $l \leq 2^n$. Therefore, this method is more efficient comparing with naive method. The result is shown in Table VI.

Table VI
COMPARISON IN SPACE AND TIME WITH THE NAIVE METHOD

Algorithm	Memory space	Time	DB passes
naive	$O(1)$	$O(m2^n n^2)$	$O(2^n n^2)$
efficient method	$O(n^2)$	$O(l^2)$	$O(l)$

IV. IMPLEMENTATION AND EXPERIMENTS

We implement the algorithm and apply it to an Arrhythmia data set [22], which distinguishes between the presence and absence of cardiac arrhythmia and classifies them into different groups. The data set contains 452 records and each record obtains 279 data attributes and one class attribute. All the data attributes have been translated into binary variables in which 1 indicates yes and 0 indicates no. Here we consider the class named “Right bundle branch block” as target and aim to find its causes in the attributes. Hence the patients are classified into two categories: Right bundle branch block or not, and they are labeled in the class attribute respectively with 1 and 0. Our goal is to discover the potential causal relationships between the data attributes and the target.

Table VII
SINGLE CAUSAL RULES DISCOVERED WITH OUR METHOD
(α=0.02, n^*=5)

Causal Rule in form of X→Y		
P18→P280	P53→P280	P57→P280
P89→P280	P105→P280	P138→P280
P150→P280	P160→P280	P177→P280
P189→P280	P198→P280	P217→P280
P228→P280	P229→P280	P238→P280
P267→P280	P277→P280	

A. Causal rule discovery

We firstly find all the single causal rules using our method, and compare it with Verstein’s method [16]. The minimum support threshold n^* is set as 5, and the significance level threshold α is set as 0.02 considering the numbers of causal rules. To facilitate the description, the data attribute with index i is indicated as Pi, for example, the third attribute is $P3$. The discovered causal rules with our algorithm are shown in Table VII.

CCC and CCU are two causal structures presented in [16]. We implement the methods in [16] for discovering the CCC and CCU structures, and apply it to the Arrhythmia Data Set with the same values for α and n^*. Suppose that A, B and C are three attributes and they form a CCC structure. That is, A, B and C are pairwise dependent, meanwhile A and C are independent given B. With the assumption of no hidden and confounding variables, three different causal paths could be inferred: $A \rightarrow B \rightarrow C$, $A \leftarrow B \leftarrow C$, $A \leftarrow B \rightarrow C$.

Causal rules like $A \rightarrow B$, $C \rightarrow B$ can be deducted when B is the target. The causal rules contained in the CCC structures are found and illustrated in Table VIII.

Table VIII
CAUSAL RULES BASED ON CCC STRUCTURES
(α=0.02, n^*=5))

CCC Causal Rule X→Y		
P5→P280	P17→P280	P18→P280
P29→P280	P30→P280	P42→P280
P53→P280	P57→P280	P71→P280
P77→P280	P78→P280	P89→P280
P90→P280	P91→P280	P93→P280
P95→P280	P102→P280	P103→P280
P104→P280	P105→P280	P107→P280
P115→P280	P116→P280	P138→P280
P150→P280	P160→P280	P63→P280
P173→P280	P177→P280	P178→P280
P189→P280	P192→P280	P197→P280
P198→P280	P213→P280	P217→P280
P222→P280	P224→P280	P227→P280
P228→P280	P229→P280	P234→P280
P235→P280	P237→P280	P238→P280
P244→P280	P245→P280	P267→P280
P273→P280	P277→P280	

With respect to a CCU structure, A and B are dependent, and C and B are dependent. A and C are independent, but they become dependent conditioned on B. The only causal path is $A \rightarrow B \leftarrow C$. The CCU causal rules found using method [16] are listed in Table IX.

Table IX
CAUSAL RULES BASED ON CCU STRUCTURES
(α=0.02, n^*=5)

CCU Causal Rule X→Y		
P5→P280	P17→P280	P18→P280
P30→P280	P42→P280	P57→P280
P78→P280	P89→P280	P90→P280
P91→P280	P93→P280	P102→P280
P105→P280	P107→P280	P163→P280
P192→P280	P197→P280	P213→P280
P217→P280	P222→P280	P224→P280
P229→P280	P237→P280	

Table X
COMBINED CAUSAL RULES DISCOVERED WITH OUR METHOD
(α=0.05, n^*=5)

Association	+/−	Association	+/−
P3→P280	−	P11→P280	−
P12→P280	−	P14→P280	−
P89→P280	+	P160→P280	+
P171→P280	−	P195→P280	−
P215→P280	−	P221→P280	−
P225→P280	−	P231→P280	−
P241→P280	−	P245→P280	−

Causal Rule	+/−	Partial Association Value
P3&P11→P280	+	5.0698
P11&P171→P280	+	4.9543
P11&P231→P280	+	2.2982
P171&P221→P280	+	0.8767
P171&P231→P280	+	3.1126

Comparing with the set of CCC causal rules, the set of single casual rules discovered with our method is a subset of it. Quite a lot of inconstant associations include in the CCC causal rules should have been removed using partial association test. It is infeasible to find plausible alternative explanation for the effect other than the cause [23]. In this paper, we use partial association tests to reduce the plausibility of other explanations. Therefore, there is no cause and effect relationship between two variables when their association is zero partial. Based on this, the CCC rules may include many non-causal rules. For example, P5→P280 in Table VIII is a causal relationship based on the CCC structure. When the partial association test is applied to check if the casual relationship is persistent, it fails to pass the test due to a low partial association value (3.250). As a result, P5→P280 should be eliminated.

The result of the CCU method shows a small number of causal rules. It is also a subset of the CCC causal rules, but there are several different causal rules in the result which do not appear in Table VIII. The same reason for the set of CCU rules, it also contains a few non-causal rules. In addition, the CCU method works in an opposite direction compared to partial association test. In the view of CCU, we need to examine the association between different predictive variables. Take the relation A→B for instance, we have to find another predictive variable C, which is independent of A but becomes dependent on A given B, before we can determine A→B to be a causal rule. In other words, to find causal rules in CCU structures we must investigate the influence of the target variable on the associations among predictive variables. Instead, the partial association test investigates the influence of other predictive variable on the associations between predictive variables and target variable. In conclusion, it is reasonable for the two methods CCC and CCU to make a difference in the result, especially in a large data set.

In discovering of causal rules, we concern the global persistence of causal relationships and remove the influence of other predictive variables. A lot of the positive associations may not pass the partial associations tests thus are eliminated, so the causal rules defined in this paper should be more reliable. Besides, the proposed algorithm is more efficient according to the experimental comparison (see the next section for details).

Furthermore, another contribution of the proposed method is that we extend the causal rules for the situation with combine cause variables. Two or more variables are combined to test if they are both positive associated and have nonzero partial association with the target. In consideration of the small sample size of the data set, we mainly investigate 15 attributes. For simplicity, we only list the combined positive associations and summarise them in Table X. The symbol "+" or "−" indicates positive or zero association for the single rules in the top part of the table, and in the lower part of the table "+" indicates both positive association and nonzero partial association for the combined rules.

As $\chi^2_{0.05}$=3.84, we determine that P3&P11→P280 and P11&P171→P280 are the newly discovered causal rules because they also have nonzero partial associations. The result shows that the combinations of zero associated variables can generate combined causal rules, such as P3&P11→P280 in which P3 and P11 are both zero associated with P280 respectively. The combination of positive associated variables are thought to be redundant and provide little new information. However, P3&P11→P280 and P11&P171→P280 are new causal rules, which imply that the interaction of zero associations can produce causal relationships. Considering the CCC and CCU structures, the zero associated variables are ignored directly. Based on this, our method could generate new causal rules which are not considered by the CCC or CCU method.

B. *Performance*

Apriori [24] is the most fundamental algorithm for generating association rules in large databases. It is implemented distinctively here to find the association rules in regards to a given target whose causes we are only interested in discovering. The scalable techniques [16] for discovering the CCC and CCU causal structures are also used for comparison. We select two subsets with record size 20K and

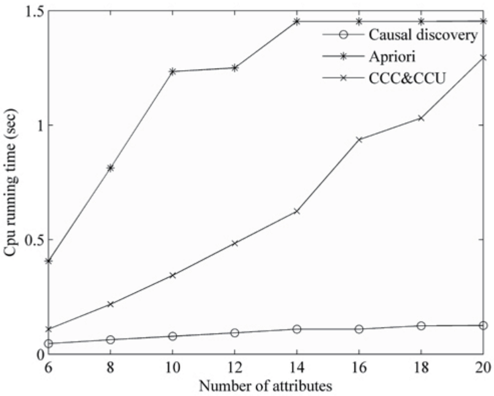

Figure 1: Extraction Time Comparison (20K Records)

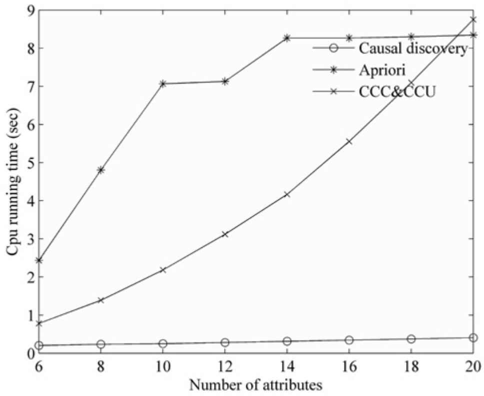

Figure 2: Extraction Time Comparison (100K Records)

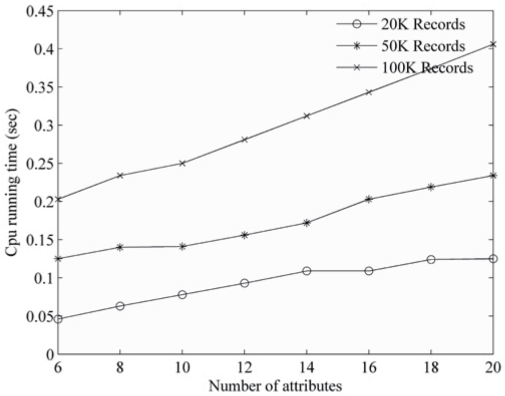

Figure 3: Scale-up of records

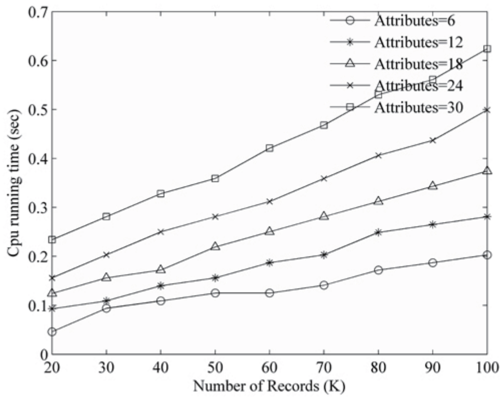

Figure 4: Scale-up of attributes

100K from the census income data [22]. Each set is sampled to their 8 subsets with transaction sizes: 6, 8, 10, 12, 14, 16, 18 and 20. We run all the algorithms for these subsets in sequence to show the superiority of our algorithm. Both Figure 1 and Figure 2 show the performance of the causal discovery proposed in this paper relative to Apriori and the CCC&CCU algorithm. For the two groups of subsets, the Causal discovery algorithm to be the most efficient one and it finds all single causal rules. As shown in the figures, the extraction time of the other two algorithms is nearly larger by 1 order of magnitude when the attributes size becomes large. For the subsets with small number of attributes, the CCC&CCU algorithm has better performance than Apriori. The growth of attribute size leads to clear performance degradation, and the Causal discovery algorithm degenerates most slowly for both the two scales of 20K and 100K.

Further, we do experiments to evaluate the scalability of the algorithms with the record size and the number of attributes. Figure 3 shows that the Causal discovery algorithm scales up with the number of records. We examine the performance degradation of the algorithm for three different scales: 20K, 50K, 100K. The significance level threshold and the minimum support threshold are respectively set as 0.05 and 5, and they remain the same in all the experiments. As shown in Figure 3, the extraction time increases gently with the number of attributes. More important, the curve is linear, which means that the performance of our algorithm is linearly related to the increase of attribute size.

Next, the increase of extraction time for different record size with different number of attributes is evaluated and the curves are shown in Figure 4. 8 sample points ranging from 20K to 100K are extracted to create the variation curves. Each curve indicates the trend of extraction time along with record size. They clearly illustrate that the time of discovering single causal rules by my Causal discovery algorithm is linear in the size of data set for all attributes scales. As a result, the algorithm should be adaptable to discovery casual rules in large data set efficiently.

V. Conclusions and Future work

In this paper we have developed a general approach to discover causal rules in large databases of binary variables. Based on the idea of a causal relationship being persistent, our method takes both associations and partial associations into account to detect the causal relationships. Also the method extends previous work by considering combined cause factors, and the identified combined causal rules can be potentially useful in a wide range of areas. To cope with the computational complexity of partial association tests, we have proposed to use QuickSort in the algorithm, which has significantly improved the efficiency. When applying our approach to a set of real-world diagnostic data, the algorithm is able to efficiently produce both single and combined causal rules.

In the future, we plan to evaluate the performance of this algorithm with a wider range of large scale data sets. In addition, we will conduct further comparisons of our approach with other causality discovery methods. The algorithm proposed aims at discovering causality from observed binary data with known target, so we intend to extend the algorithm to make it adapt for various data types and discover causal relationship effectively from more complex cases.

References

[1] J. Pearl and T. S. Verma, “A theory of inferred causation”, *In Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, Morgan Kaufmann, San Mateo, pp. 441-452, 1991.

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

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

[4] I. J. Good, “A theory of causality”, *British Journal for the Philosophy of Science*, Vol. 9, No. 36, pp. 307-310, 1959.

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

[6] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, San Mateo, CA, 1988.

[7] J. Pearl, “From Bayesian network to causal networks”, *In Mathematical models for handling partial knowledge in artificial intelligence*, Plenum Press, pp. 157-181, 1995.

[8] D. Heckerman, “A Bayesian approach to learning causal networks”, *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 285-295, 1995.

[9] N. L. Zhang and D. Poole, “Exploiting Causal Independence in Bayesian Network Inference”, *Journal of Artificial Intelligence Research*, Vol. 5, No. 1, pp. 301-328, 1996.

[10] D. Heckerman, “Bayesian Networks for Data Mining”, *Data Mining and Knowledge Discovery*, Vol. 1, No. 1, pp. 79-119, 1997.

[11] G. Cooper, D. Heckerman and C. Meek, “A Bayesian approach to casual discovery”, Technical Report, Microsoft Research, 1997.

[12] M. R. Waldmann and L. Martignon, “A Bayesian network model of causal learning”, *In Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pp. 1102-1107, 1988.

[13] S. Nadkarni and P. P. Shenoy, “A Bayesian network approach to making inferences in causal maps”, *European Journal of Operational Research*, Vol. 128, No. 3, pp. 479-498, 2001.

[14] D. M. Chickering, “Learning Bayesian Networks is NP-Complete”, Technical Report, Microsoft Research, 1996.

[15] G. F. Cooper, “A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationship”, *Data Mining and Knowledge Discovery*, Vol. 1, No. 2, pp. 203-224, 1997.

[16] C. Silverstein, S. Brin, R. Motwani and J. Ullman, “Scalable Techniques for Mining Causal Structures”, *Data Mining and Knowledge Discovery*, Vol. 4, No. 2-3, pp. 163-192, 2000.

[17] S. Mani, P. L. Spirtes and G. F. Cooper, “A theoretical study of Y structures for causal discovery”, *In Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence*, pp. 314-323, 2006.

[18] J. P. Pellet and A. Elisseeff, “Using Markov Blankets for Causal Structure Learning”, *The Journal of Machine Learning Research*, Vol. 9, pp. 1295-1342, 2008.

[19] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani and X. D. Koutsoukos, “Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithm and Empirical Evaluation”, *Journal of Machine Learning Research*, Vol. 11, pp. 171-234, 2010.

[20] N. Mantel and W. Haenszel, “Statistical aspects of the analysis of data from the retrospective analysis of disease”, *Journal of the National Cancer Institute*, Vol. 22, No. 4, pp. 719-748, 1959.

[21] M. W. Birch, “The Detection of Partial Association, I: The 2×2 Case”, *Journal of the Royal Statistical Society*, Vol. 26, No. 2, pp. 313-324, 1964.

[22] A. Frank and A. Asuncion, *UCI Machine Learning Repository*, http://archive.ics.uci.edu/ml , 2010.

[23] W. R. Shadish, T. D. Cook and D. T. Campbe, *Experimental and quasi-experimental designs for generalized causal inference*, Wadsworth Publishing, 2002.

[24] R. Agrawal and R. Srikant, “Fast Algorithms for mining association rules”, *In Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499, 1994.