



## □ DISCOVERING CAUSALITY IN LARGE DATABASES

SHICHAO ZHANG

University of Technology, Sydney, Australia;  
School of Computing, Guangxi University, Nanning,  
People's Republic of China

CHENGQI ZHANG

University of Technology, Sydney, Australia

*A causal rule between two variables,  $X \rightarrow Y$ , captures the relationship that the presence of  $X$  causes the appearance of  $Y$ . Because of its usefulness (compared to association rules), techniques for mining causal rules are beginning to be developed. However, the effectiveness of existing methods (such as the LCD and CU-path algorithms) are limited to mining causal rules among simple variables, and are inadequate to discover and represent causal rules among multi-value variables. In this paper, we propose that the causality between variables  $X$  and  $Y$  be represented in the form  $X \rightarrow Y$  with conditional probability matrix  $M_{Y|X}$ . We also propose a new approach to discover causality in large databases based on partitioning. The approach partitions the items into item variables by decomposing “bad” item variables and composing “not-good” item variables. In particular, we establish a method to optimize causal rules that merges the “useless” information in conditional probability matrices of extracted causal rules.*

### INTRODUCTION

In current literature, there are mainly three representing forms for discovering knowledge in large databases, which are item-based association rules, quantitative association rules, and causality. The work on mining item-based association rules (Agrawal, Imielinski, and Swami 1993; Agrawal and Srikant 1994; Brin, Motwani, and Silverstein 1997; Piatetsky-Shapiro 1991; Shintani and Kitsuregawa 1998; Srikant and Agarwal 1997) and quantitative association rules (Han, Cai, and Cercone 1993; Miller and Yang 1997; Srikant and Agrawal 1996) has formed a richer and well-considered framework. The work on mining causality among variables in large databases has

We would like to thank the anonymous reviewers for their good comments on this paper.

Address correspondence to Shichao Zhang, Faculty of Information Technology, University of Technology, P.O. Box 123, Broadway NSW 2007, Sydney, Australia. E-mail: zhangsc@it.uts.edu.au

also begun in Cooper (1997); Heckerman, Geiger, and Chickering (1995); Silverstein et al. (1998) on account of its usefulness to practical applications such as decision and planning. The mining models for causality, such as the LCD algorithm (Cooper 1997) and the CU-path algorithm (Silverstein et al. 1998), which are referred to as constraint-based causal discovery, have been proposed for mining causal relationships in market basket data. In fact, the CU-path algorithm is an improved model of the LCD algorithm, which applies chi-squared formula to test the dependence, independence, and conditional independence between variables so as to discover the possible causal relationships between these variables.

However, these models are only of utility to mine causal rules among simple variables such as “*states*  $\rightarrow$  *united*” for words in the *clari-world* news hierarchy (Silverstein et al. 1998). They are inadequate to discover causal rules among multi-value variables in large databases and to represent them. Actually, mining causality among multi-value variables in many applications such as decision, diagnosis, and planning, is useful for solving problems in applications. Accordingly, we propose a model of mining causality among multi-value variables in large databases based on partition in this paper, in which the causality is represented by the form  $X \rightarrow Y$  with conditional probability matrix  $M_{Y|X}$  (Pearl 1988). The task of mining causality can be simply regarded as:

- (1) partitioning item variables, and
- (2) estimating conditional probability matrices of rules of interest.

The second sub-task will be implemented in Piatetsky-Shapiro’s argument. The first subtask is more difficult than the second sub-task. For the problems in the second sub-task, an equi-depth partitioning model and method of calculating the number of partitions required were posed in Srikant and Agrawal (1996). Unfortunately, because previous partitions on data are generally blind relative to a given database, the generated quantitative items and item variables are sometimes bad or not good. One of our main contributions in this paper is to advocate a new partitioning model to determine all item variables for a given database, which decomposes the “bad quantitative items” and “bad item variables”, and composes the “not-good quantitative items” and “not-good item variables”.

Though causal rules among item variables have both useful and expressive, mining item-based association rules and quantitative association rules are still necessary to many practical applications. For example, when item variable  $X$  can impact item variable  $Y$  only at fewer point-values, this knowledge is represented by the item-based association rule and the quantitative association rule more efficient than in causality.

In order to mine better structured rules, the other main contribution in this paper is to present a method of merging useless (unnecessary) information

in extracted causal rules. Apparently, this model of merging useless information is extreme utility in optimizing the knowledge in intelligent systems.

The rest of this paper is organized as follows. First, we present some needed concepts that will be used throughout the following sections. In the next section, we first define a “good partition” to generate item variables from items, and then present a method of mining causality of interest from large databases. In the following section, we build a way of improving the causal rules so as to reduce useless information in their matrices once causal rules are extracted. In the last section, a simple comparison with previous work and a summary of this paper is presented.

## BASIC DEFINITIONS

Assume  $I$  is a set of items in database  $D$ . A subset of the same type of items in  $I$  is referred to as a *quantitative item*. For convenience, we use the term of quantitative item as a set and a name interchangeably. Certainly, an item  $A \in I$  can be taken as a special quantitative item. An *item variable* denotes a variable to represent a quantitative item in the set of quantitative items of the same domain.

An *item-based association rule* is a relationship of the form:

$$A \Rightarrow B,$$

where  $A$  and  $B$ , are itemsets and  $A \cap B = \emptyset$ . It has both support and confidence greater than or equal to some user specified minimum support (*minsupp*) and minimum confidence (*minconf*) thresholds, respectively.

A *quantitative association rule* is a relationship of the form:

$$\langle \text{attribute1}, \text{value1} \rangle \Rightarrow \langle \text{attribute2}, \text{value2} \rangle,$$

where *attribute1* and *attribute2* are attributes, *value1* and *value2* are subsets of the domains of *attribute1* and *attribute2* respectively,  $\langle \text{attribute1}, \text{value1} \rangle$  and  $\langle \text{attribute2}, \text{value2} \rangle$  are quantitative items. We now illustrate mining quantitative association rules using an example.

*Example 1: Consider the personnel database in some university. The interest data is the set of records with “educational level”, “salary” of first job. We extract 30,000 such records from the database, and the statistical results are listed in Table 1.*

In Table 1, partitioning the domain of *Education* into *Doctor*, *Master*, and *UnderMaster*; or *Doctor*, *Master*, and *UnderMaster* are quantitative items; partitioning the domain of *Salary* into three quantitative items  $[3500, +\infty)$ ,  $[2100, 3500)$  and  $[0, 2100)$ ; *Number* is the statistical results such as the

**TABLE 1** Statistical Results of Interest Data

Education	Salary	Number
Doctor	[3500, +∞)	8500
	[2100, 3500)	1400
	[0, 2100)	100
Master	[3500, +∞)	1900
	[2100, 3500)	7100
	[0, 2100)	1000
UnderMaster	[3500, +∞)	200
	[2100, 3500)	3000
	[0, 2100)	6800

number of transactions that contain quantitative items *Master* and [2100, 3500) is 7100. In the light of the models in Han, Cai, and Cercone (1993); Srikant and Agrawal (1996) we can extract quantitative association rules as follows.

Rule1:  $Education = Doctor \Rightarrow Salary \geq 3500$   
with confidence 0.85

Rule2:  $Education = Master \Rightarrow Salary \in [2100, 3500)$   
with confidence 0.71

Rule3:  $Education = UnderMaster \Rightarrow Salary < 2100$   
with confidence 0.68

where Rule1, Rule2, and Rule3 are three quantitative association rules. This means that discovering such quantitative rules is beneficial to mine databases with categorical attributes.

A *causal rule* is a relationship between  $X$  and  $Y$  of the form:

$$X \Rightarrow Y,$$

where  $X$  and  $Y$  are variables with values in *ranges*  $R(X)$  and  $R(Y)$ , respectively.  $x \in R(X)$  is called a *point-value* of  $X$ , where  $x$  is a quantitative item in data mining.

In order to mine these causal rules, we propose a new model to discover causality in large databases based on “good partition”. And the causality is represented of the form  $X \rightarrow Y$  with conditional probability matrix  $M_{Y|X}$  according to Bayesian rules, which  $M_{Y|X}$  is as:

$$M_{Y|X} \stackrel{\Delta}{=} P(y|x) \stackrel{\Delta}{=} P(Y = y|X = x)$$

$$= \begin{bmatrix} p(y_1|x_1) & p(y_2|x_1) & \dots & p(y_n|x_1) \\ p(y_1|x_2) & p(y_2|x_2) & \dots & p(y_n|x_2) \\ \dots & \dots & \dots & \dots \\ p(y_1|x_m) & p(y_2|x_m) & \dots & p(y_n|x_m) \end{bmatrix}$$

where,  $p(y_j|x_i) = p(Y = y_j|X = x_i)$  are conditional probabilities,  $i = 1, 2, \dots, m$ ,  $j = 1, 2, \dots, n$ .

For example if we let  $X$  and  $Y$  be two item variables with  $\{Doctor, Master, UnderMaster\}$  and  $\{[3500, +\infty), [2100, 3500), [0, 2100)\}$  respectively, for Example 1, we can obtain causal rule  $X \Rightarrow Y$  with conditional probability matrix as follows:

$$M_{Y|X} = \begin{bmatrix} 0.85 & 0.14 & 0.01 \\ 0.19 & 0.71 & 0.1 \\ 0.02 & 0.3 & 0.68 \end{bmatrix}$$

where,  $p(Y = [3500, +\infty)|X = Doctor) = 0.85$ ,  $p(Y = [2100, 3500)|X = Doctor) = 0.14$ ,  $p(Y = [0, 2100)|X = Doctor) = 0.01$ ,  $p(Y = [3500, +\infty)|X = Master) = 0.19$ ,  $p(Y = [2100, 3500)|X = Master) = 0.71$ ,  $p(Y = [0, 2100)|X = Master) = 0.1$ ,  $p(Y = [3500, +\infty)|X = UnderMaster) = 0.02$ ,  $p(Y = [2100, 3500)|X = UnderMaster) = 0.3$ ,  $p(Y = [0, 2100)|X = UnderMaster) = 0.68$ .

This nice result is due to a better partition. However, because partitions on data are generally blind relative to a given database, constructing a reasonable partition is difficult for applications. For example, if  $\{Doctor, Under Doctor\}$  is a partition on  $R(Education)$  for Example 1, then Rule2 and Rule3 can be neither generalized in casual rule  $X \Rightarrow Y$  nor discovered as a valid rules. This means, *Under Doctor* is a **bad quantitative item** under the partition. Furthermore, if two quantitative items can compose a new quantitative item under a partition and the new quantitative item is not a bad quantitative item, then the two quantitative items are referred to as **not-bad quantitative items**. If a quantitative item can't be composed with any other quantitative item into a not-bad quantitative item under the partition, it is called a **good quantitative item**.

Also, if an item variable causes that a quantitative rule can't be generalized in a certain causal rule, it is called a **bad item variable**. If two item variables can compose a new item variable under a partition and the new item variable is not a bad item variable, then the two item variables are referred to as **not-bad item variables**. If an item variable can't be composed with any other item variable into a not-bad item variable under the partition, it is called a **good item variable**.

A partition which can cause bad quantitative items or bad item variables, is called a **bad partition**. A partition can cause not-good quantitative items or not-good variables, is called a **not-good partition**. If all quantitative items and item variables are good under a partition, this partition is called a **good partition**.

For any two itemsets  $i_1$  and  $i_2$ ,  $i_1$  and  $i_2$  are *property tolerant* if and only if  $i_1$  and  $i_2$  have a same property, or attribute, or constraint.  $i_1$  and  $i_2$  are *associated tolerant* if and only if for any itemset  $i_3$ ,  $p(i_3|i_1) \approx p(i_3|i_2)$ . Again,

two quantitative items  $q_1$  and  $q_2$  are *property tolerant* if and only if  $q_1$  and  $q_2$  have a same property, or attribute, or constraint.  $q_1$  and  $q_2$  are *associated tolerant* if and only if for any quantitative item  $q_3$ ,  $p(q_3|q_1) \approx p(q_3|q_2)$ . And two item variables  $X_1$  and  $X_2$  are *property tolerant* if and only if  $X_1$  and  $X_2$  have a same property, or attribute, or constraint.  $X_1$  and  $X_2$  are *associated tolerant* if and only if for any item variable  $Y$ ,  $p(Y|X_1) \approx p(Y|X_2)$ .

## CAUSALITY DISCOVERY IN LARGE DATABASES

In this section, we propose a new model of discovering probabilistic causality based on a “good partition” and Piatetsky-Shapiro’s argument. We shall first discuss the partition on data, followed by the new model of mining probabilistic dependencies from large relational databases, and finally, we shall present the algorithm of discovering causal rules.

### Partitioning Domains of Attributes

Though there are a lot of data partitioning models, a major issue in mining causality is still the partitioning technique on data and domains of attributes for specific applications. In data mining, there are two partitioning models: the knowledge-based partitioning model (Han, Cai, and Cercone 1993) and the equi-depth partitioning model (Srikant and Agrawal 1996). Han, Cai, and Cercone (1993) put forward a knowledge based partitioning, which requires background knowledge, such as concept hierarchies, data relevance, and expected rule forms. This partitioning is efficient to discover a kind of quantitative association rules from relational databases with using attribute-oriented induction method. The other model, called as equi-depth partitioning model, is proposed by Srikant and Agrawal (1996), which is optimal for the measure of partial completeness. This partitioning is useful for mining quantitative association rules in databases. In particular, the number of partitions required can be calculated as:

$$\text{Number of Intervals} = \frac{2n}{m(K-1)}$$

where,  $n$  is the number of quantitative attributes,  $m$  is the minimum support, and  $K$  is the partial completeness level. However, such a partition on data is blind relative to a given database. It is possible that some quantitative items are bad and others are not-good because of the blindness of a partition to a given database. In fact, the above number of partitions doesn’t concern the associated tolerant. We will use a so-called “good partition” to generate quantitative items and item variables for a given database, which decomposes the “bad item variables” and composes the “good item variables.”

Let  $D$  be a given database,  $I$  the set of all items in  $D$ . Our partitioning model is as follows:

- (1) Generating relative properties, attributes, and constraint conditions for  $D$ .
- (2) Generating the set  $QI$  of all quantitative items by these relative constraint conditions, which all quantitative items are formed a partition of  $I$ .
- (3) Optimizing all the quantitative items using the decomposition and composition for quantitative items.
- (4) Generating the set  $IV$  of all item variables by these relative properties and attributes, which all item variables are formed a partition of  $QI$  and each item variable is taken some quantitative items as its point-values. This means that each item variable can be viewed as a set of some quantitative items with the same property (or attribute) in some sense.
- (5) Optimizing all the item variables using the decomposition and composition for item variables.

### **Quantitative Items**

Generally, there are different partitions for the domain of an attribute in different applications. Thus, we must consider users' requirements and the reasonable of problem to determine a partition.

In previous sections, we partition the domains of *Education* and *Salary* as

$$\{Doctor, Master, UnderMaster\}$$

and

$$\{[3500, +\infty), [2100, 3500), [0, 2100)\}$$

respectively, and  $X$  and  $Y$  stand for *Education* and *Salary*, respectively. In this example, the constraint condition on quantitative item *Doctor* is  $Education = Doctor$ , the constraint condition on quantitative item *UnderMaster* is that *Education* is lower than *Master*, the constraint condition on quantitative item  $[3500, +\infty)$  is  $3500 \leq Salary < +\infty$ , and the constraint condition on quantitative item  $[2100, 3500)$  is  $2100 \leq Salary < 3500$ , etc. For  $R(Education)$ ,  $q_1 = [x]_{Education=Doctor}$ ,  $q_2 = [x]_{Education=Master}$ ,  $q_3 = [x]_{Education=UnderMaster}$  are three subsets of  $R(Education)$ . For  $R(Salary)$ ,  $q_4 = [x]_{3500 \leq Salary < +\infty}$ ,  $q_5 = [x]_{2100 \leq Salary < 3500}$  and  $q_6 = [x]_{0 \leq Salary < 2100}$  are three subsets of  $R(Salary)$ , which each subset is a set of discrete real numbers. For simplicity, such as  $q_4, q_5$  and  $q_6$  will be denoted as three intervals:  $[3500, +\infty)$ ,  $[2100, 3500)$  and  $[0, 2100)$ , respectively.

According to this, we can formally define quantitative items and partition as follows.

*Definition 1:* Assume  $I$  is a set of items. A quantitative item over  $I$  is a set of all items satisfied constraint condition  $CR$ . A consequence of quantitative items:  $q_1, q_2, \dots, q_k$  is a partition of  $I$  if it satisfies:

- (1)  $I = q_1 \cup q_2 \cup \dots \cup q_k$ ;
- (2)  $q_i = \{A | A \in I \wedge [A]_{CR_i}\}$ , where  $CR_i$  is a constraint relation,  $[A]_{CR_i}$  means that item  $A$  satisfied the constraint  $CR_i$ ;
- (3)  $q_i \cap q_j = \emptyset$  for  $i \neq j$ ,  $1 \leq i, j \leq k$ .

In fact, a quantitative item is the generalization of some items with the same constraint condition. For the above example, the consequence of quantitative items:  $q_1, q_2, \dots, q_6$  is partition of  $I$ . And  $q_1$  is a generalization of *Doctor* with education *Doctor*,  $q_3$  is a generalization of *Bachelor*, *Under Bachelor* with education lower than *Master*,  $q_4$  is a generalization of items with salary over 3500, and  $q_6$  is a generalization of items with salary less than 2100.

According to different requirements in applications, we can divide them into different sets of quantitative items. For example, we can partition  $R(\text{Education})$  and  $R(\text{Salary})$  as:

$$\begin{aligned} & \{\text{Doctor}, \text{UnderDoctor}\}, \\ & \{\text{Doctor}, \text{Master}, \text{Bachelor}, \text{UnderBachelor}\}, \\ & \{\text{Doctor}, \text{Master}, \text{UnderMaster}\} \end{aligned}$$

and

$$\begin{aligned} & \{[7200, +\infty), [3500, 7200), [2100, 3500), [0, 2100)\}, \\ & \{[3500, +\infty), [2100, 3500), [0, 2100)\}, \\ & \{[3500, +\infty), [0, 3500)\}, \end{aligned}$$

respectively. However, a reasonable partition on data for data mining also needs to consider the identity on supports of items and the associated degree with other items. We apply the decomposition and composition for quantitative items to generate good partition.

### **Decomposition and Composition for Quantitative Items**

In order to find out a “good partition” for a given database, the quantitative items partitioned in properties must be optimized. We now define a method to decompose and compose quantitative items.

*Lemma 1: Let  $I$  be the set of all items of a given database,  $QI$  the set of all quantitative items under a partition.*

- (i) *For  $q \in QI$ , is a bad quantitative item if and only if there are at least two items  $i_1$  and  $i_2$  in  $I$  such that  $i_1 \rightarrow i_3$  and  $i_2 \rightarrow i_4$  can all be extracted as valid rules and,  $i_3 \cap i_4 = \emptyset$ , where  $i_3$  and  $i_4$  are itemsets over  $I$ .*
- (ii) *For  $q \in QI$ ,  $q$  is a good quantitative item if and only if all items in  $q$  are associated tolerant. More intuitively, for any two items  $i_1, i_2 \in q$  and  $i_3 \in I$ ,  $p(i_3|i_1) \approx p(i_3|i_2)$  holds.*
- (iii) *For  $q_1, q_2 \in QI$ ,  $q_1$  and  $q_2$  are not-good quantitative items if and only if  $q_1$  and  $q_2$  can be composed into a new quantitative item  $q_3$  and  $q_3$  is a good quantitative item.*

*Proof:* It can directly be proven according to previous definitions.

Certainly, bad quantitative items are not allowed in mining quantitative association rules. And not-good quantitative items are also avoided unless it is required. Accordingly, we now build the decomposition of bad quantitative items and the composition of not-good quantitative items as follows:

*Procedure 1: DecComposeQI;*

**Input:**  $I$ : set of all items,  $QI$ : set of all quantitative items in property,

**Output:**  $OQI$ : set of optimized quantitative items;

- (1) **let**  $OQI \leftarrow \text{emptyset}$ ;  
**let**  $qset \leftarrow QI$ ;  
**for any element**  $q$  **in**  $qset$  **do begin**  
**if**  $q$  **is a bad quantitative item then**  
**if**  $i_1, i_2 \in q$  **and they are not associated tolerant then beginif**  
**decompose**  $q$  **into two sub-quantitative items**  $q_1$  **and**  $q_2$  **such that**  
 $q_1 \cup q_2 = q$ ,  $i_1 \in q_1$  **and**  $i_2 \in q_2$ ;  
**let**  $qset \leftarrow (qset - \{q\}) \cup \{q_1, q_2\}$ ;  
**endif**;  
**enddo**;
- (2) **for any two elements**  $q_1$  **and**  $q_2$  **in**  $qset$  **do begin**  
**if**  $q_1$  **and**  $q_2$  **are property tolerant then**  
**if**  $q_1$  **and**  $q_2$  **are associated tolerant then beginif**  
**compose**  $q_1$  **and**  $q_2$  **into a new quantitative item**  $q$  **such that**  $q = q_1 \cup q_2$  **and**  
 $q$  **is not a bad quantitative item**;  
**let**  $qset \leftarrow (qset - \{q_1, q_2\}) \cup \{q\}$ ;  
**endif**;  
**enddo**;

(3) **let**  $OQI \leftarrow qset$ ;  
**output**  $OQI$ ;

### **Item Variables**

According to our partitioning model, an attribute such as *Education* and *Salary* can be taken as an item variable. For the above item variables  $X$  and  $Y$ ,  $X$  is the set of quantitative items: *Doctor*, *Master* and *UnderMaster* with same attribute *Education*. That is, any element of  $X$  is denoted a degree of education. And  $Y$  is the set of quantitative items:  $[3500, +\infty)$ ,  $[2100, 3500)$  and  $[0, 2100)$  with same attribute *Salary*. That is, any element of  $y$  is denoted an order of salary. However, an attribute sometimes needs to be divided into several different variables for specific applications according to different properties. For example, let the domain of the attribute “Weather” in a system is  $\{strongsun, middlesun, weaksun, largerrain, middlerain, smallrain, \dots\}$ . In different application, “Weather” is sometimes taken as a variable, and sometimes divided into different variables as  $X_1, X_2, \dots$ , where, the domain of  $X_1$  is  $\{strongsun, middlesun, weaksun\}$ , the domain of  $X_2$  is  $\{largerrain, middlerain, smallrain\}, \dots$ . In this partitioning, item variable  $X_1$  is used for describing the degree of sun, or all elements of  $X_1$  have the same property: rain. We can now formally define item variable as follows.

*Definition 2: Assume  $QI$  is a set of quantitative items. An item variable over  $QI$  is a set of all quantitative items satisfied a property (or attribute). A consequence of item variables:  $X_1, X_2, \dots, X_m$  is a partition of  $QI$  if it satisfies:*

- (1)  $QI = X_1 \cup X_2 \cup \dots \cup X_m$ ;
- (2) the range of  $X_i$  is the set  $\{q | q \in QI \wedge P1(q)\}$ , where  $P1$  is a property (or attribute),  $P1(q)$  means that quantitative item  $q$  satisfies the property  $P1$ ;
- (3)  $X_i \cap X_j = \emptyset$  for  $i \neq j, 1 \leq i, j \leq m$ .

An item variable is the generalization of some quantitative items with the same properties. For previous examples,  $QI = \{q_1, q_2, q_3, q_4, q_5, q_6\}$ , we take  $X$  and  $Y$  as two item variables with domains  $R(X) = \{q_1, q_2, q_3\}$  and  $R(Y) = \{q_4, q_5, q_6\}$ , respectively. Then the consequence of item variables,  $X$  and  $Y$ , is a partition of  $QI$  according to the above definition. Also, we apply the decomposition and composition for item variables to generate good partition.

### **Decomposition and Composition for Quantitative Items**

*Lemma 2: Let  $I$  be the set of all items of a given database,  $QI$  the set of all quantitative items under a partition.  $IV$  the set of all item variables under a partition.*

- (1) For  $X \in IV$ ,  $X$  is a bad item variable if and only if there are at least two quantitative items  $q_1$  and  $q_2$  in  $X$  such that  $q_1 \rightarrow q_3$  and  $q_2 \rightarrow q_4$  can all be extracted as valid quantitative rules and,  $q_3 \cap q_4 = \emptyset$ , where  $q_3$  and  $q_4$  are quantitative itemsets over  $QI$ .
- (2) For  $X \in IV$ ,  $X$  is a good item variable if and only if all quantitative items in  $X$  are associated tolerant. More intuitively, for any two items  $q_1, q_2 \in X$  and  $q_3 \in QI$ ,  $p(q_3|q_1) \approx p(q_3|q_2)$  holds.
- (3) For  $X_1, X_2 \in IV$ ,  $X_1$  and  $X_2$  are not-good item variables if and only if  $X_1$  and  $X_2$  can be composed into a new item variable  $X_3$  and  $X_3$  is a good item variable.

*Proof:* It can directly be proven according to previous definitions.

Certainly, bad item variables are not allowed in mining causal rules. And not-good item variables are also avoided unless they are required. Accordingly, we now build the decomposition of bad item variables and the composition of not-good item variables as follows.

*Procedure 2: DecComposeIV;*

**Input:**  $QI$ : set of all quantitative items,  $IV$ : set of all item variables in property,

**Output:**  $OIV$ : set of optimized item variables;

- (1) **let**  $OIV \leftarrow \text{emptyset}$ ;  
**let**  $vset \leftarrow IV$ ;  
**for any element**  $X$  **in**  $vset$  **do begin**  
**if**  $X$  **is bad item variable** **then**  
**if**  $q_1, q_2 \in R(X)$  **and they are not associated tolerant** **then beginif**  
**decompose**  $X$  **into two item variable**  $X_1$  **and**  $X_2$  **such that**  $R(X_1) \cup R(X_2) = R(X)$ ,  $q_1 \in R(X_1)$  **and**  $q_2 \in R(X_2)$ ;  
**let**  $qset \leftarrow (vset - \{X\}) \cup \{X_1, X_2\}$ ;  
**endif**;  
**enddo**;
- (2) **for any two elements**  $X_1$  **and**  $X_2$  **is**  $vset$  **do begin**  
**if**  $X_1$  **and**  $X_2$  **are property tolerant** **then**  
**if**  $X_1$  **and**  $X_2$  **are essentially tolerant** **then beginif**  
**compose**  $X_1$  **and**  $X_2$  **into a new item variable**  $X$  **such that**  $R(X) = R(X_1) \cup R(X_2)$  **and**  $X$  **is not bad item variable**;  
**let**  $vset \leftarrow (vset - \{X_1, X_2\}) \cup \{X\}$ ;  
**endif**;  
**enddo**;

(3) **let**  $OIV \leftarrow vset$ ;  
**output**  $OIV$ ;

We now build the algorithm of partitioning model as follows. Let  $D$  be a given database,  $I$  the set of all items in  $D$ .

*Procedure 3: PartitionData;*

**Input:**  $D$ : database,  $Table$ : a concept hierarchy table,  $I$ : set of all items in  $D$ ;

**Output:**  $OQI$ : the set of quantitative items,  $OIV$ : the set of item variables;

(1) generate  $S_p$  the set of properties,  $S_a$  the set of attributes by  $Table$ , and  $S_c$  the set of constraint conditions for  $D$ ;  
(2) **let**  $I \leftarrow$  all items in  $D$ ;  
**for**  $c \in S_c$  **do begin**  
**generate**  $q_c$  over  $I$   
**if**  $q_c \neq \emptyset$  **then**  
**let**  $QI \leftarrow QI \cup \{q_c\}$ ;  
**end**;  
**for each** item  $i$  of  $I$  **then**  
**if**  $I$  is not contained in any quantitative item **then**  
**let**  $r \leftarrow r \cup \{i\}$ ;  
**let**  $QI \leftarrow QI \cup \{r\}$ ;  
(3) optimize  $QI$  by Procedure 1;  
(4) **for**  $a \in S_p \cup S_a$  **do begin**  
**generate**  $x_a$  over  $QI$ ;  
**if**  $x_a \neq \emptyset$  **then**  
**let**  $IV \leftarrow IV \cup \{x_a\}$ ;  
**for each** quantitative item  $q$  of  $QI$  **then**  
**if**  $q$  is not a value of any item variable **then**  
**let**  $z \leftarrow z \cup \{q\}$ ;  
**let**  $IV \leftarrow IV \cup \{z\}$ ;  
(5) optimize  $IV$  by Procedure 2;  
**end**;

### Discovering Probabilistic Dependencies

In this subsection, we first present the method of acquiring conditional probabilities of the point-values (quantitative items) of an item variable given another item variable, and then propose a way to identify causal rules of interest in this subsection.

### Acquiring Conditional Probabilities

After quantitative items and item variables are generated, the work of mining causal rules in databases becomes easy. We can statistic the probabilities:  $p(X = a)$ ,  $p(Y = b)$ , and  $p(Y = b \wedge X = a)$  for any two item variables  $X$  and  $Y$  as follows:

$$\begin{aligned} p(X = a) &= N(X = a)/n; \\ p(Y = a) &= N(Y = b)/n; \\ p(Y = b \wedge X = a) &= N(Y = b \wedge X = a)/n. \end{aligned}$$

where,  $n$  is the total number of tuples in the database,  $N(X = a)$  denotes the number of tuples in the database that contain the quantitative item  $a$ ,  $N(Y = b)$  denotes the number of tuples in the database that contain the quantitative item  $b$ , and  $N(Y = b \wedge X = a)$  denotes the number of tuples in the database that contain the quantitative items  $a$  and  $b$ . Now we can solve the conditional probability of  $Y = b$  given  $X = a$  as:

$$p(Y = b|X = a) = \frac{p(Y = b \wedge X = a)}{p(X = a)}$$

In order to illustrate the use of the above method, we choose 10 tuples from a relational database to show this procedure in the following example:

*Example 2: 10 tuples selected from a relational database is shown in Table 2. The supports and probabilities of single quantitative items and sets of quantitative items are shown in Table 2. Because there are 10 tuples in the database, the support is the number of tuples in which the items or sets of quantitative items occur divided by 10.*

Let  $n$  be the number of all tuples in the above database,  $N(a)$  stand for the number of tuples in the database that quantitative item or set of quantitative item  $a$  occurs in. For example,  $N(a)$  of quantitative item  $[3500, +\infty)$  and set of quantitative item  $\{Doctor, [3500, +\infty)\}$  can be counted as

$N([3500, +))$  is the number of all tuples in the database that its projection on *Salary* is great than or equal to 3500;

$N(\{Doctor, 3500, +\infty\})$  is the number of all tuples in the database that its projection on *Education* is *Doctor* and its projection on *Salary* is great than or equal to 3500.

Accordingly, we can obtain the  $N(a)$  of quantitative itemsets from Table 2 as shown in Table 3.

**TABLE 2** Some Data in the Database

EMP #	Education	Salary
25	doctor	5000
26	doctor	4500
27	doctor	3500
28	doctor	3500
29	doctor	4100
30	doctor	3500
31	doctor	4200
32	doctor	5400
33	doctor	2600
34	doctor	2000

According to previous definitions,

$$p(Y = [3500, +\infty) | X = Doctor) = \frac{p(Y = [3500, +\infty) \wedge X = Doctor)}{p(X = Doctor)} = 0.8,$$

$$p(Y = [2100, 3500) | X = Doctor) = \frac{p(Y = [2100, 3500) \wedge X = Doctor)}{p(X = Doctor)} = 0.1,$$

$$p(Y = [0, 2100) | X = Doctor) = \frac{p(Y = [0, 2100) \wedge X = Doctor)}{p(X = Doctor)} = 0.1.$$

### **Causal Rules of Interest**

As you will see in Table 4, when item variable  $X$  can impact item variable  $Y$  only at one or fewer point-values, a conditional probability matrix will contain much unnecessary information. In this case, an association rule based on items or quantitative association rule is more efficient than the above causal rule. In other words, we would like to discover causal rules of interest from large databases. For this reason, if a causal rule  $X \rightarrow Y$  with  $M_{Y|X}$  is of interest, it must satisfy three conditions: (1) there are enough

**TABLE 3** Statistical Results for Table 2

Itemset	Number of Tuples	Support $sup(X)$ (%)	Probability $p(X)$
[3500, +∞)	8	80	0.8
[2100, 3500)	1	10	0.1
[0, 2100)	1	10	0.1
Doctor	10	100	1
Doctor, [3500, +∞)	8	80	0.8
Doctor, [2100, 3500)	1	10	0.1
Doctor, [0, 2100)	1	10	0.1

TABLE 4 Statistical Results for Interest Data

Education	Salary	Number
PostDoctor	[3500, +∞)	9000
	[2100, 3500)	900
	[0, 2100)	100
Doctor	[3500, +∞)	8000
	[2100, 3500)	1900
	[0, 2100)	100
PostMaster	[3500, +∞)	1000
	[2100, 3500)	7500
	[0, 2100)	1500
Master	[3500, +∞)	3100
	[2100, 3500)	3100
	[0, 2100)	3800
Bachelor	[3500, +∞)	2400
	[2100, 3500)	3800
	[0, 2100)	3800
UnderBachelor	[3500, +∞)	2000
	[2100, 3500)	4000
	[0, 2100)	4000

conditional probabilities  $p(y_i|x_j)$  in  $M_{Y|X}$  that is greater than or equal to  $minconf$ ; (2) for these point-values such as  $p(y_i|x_i) \geq minconf$  is general with  $p(x_j \cup y_i) \geq minsup$  in  $M_{Y|X}$ ; and (3) these point-values also match Piatetsky-Shapiro's argument [15], or  $p(y_i|x_j) - p(x_j)$  is greater than or equal to a threshold  $\lambda$  given by users. This means that, let  $QI$  be the set of quantitative items in database  $D$ ,  $X$  and  $Y$  be item variables,  $R(X) \subset QI, |R(X)| = n$ ,  $R(Y) \subset QI, |R(Y)| = m$ .  $minsup, minconf, \gamma > 0, \alpha > 0, \lambda > 0$  and  $\eta > 0$  are given by users or experts, where  $\gamma$  is the minimum number of itemsets with supports greater than or equal to  $minsup$ ,  $\eta$  is the minimum number of conditional probabilities, and  $\alpha$  is the minimum number of probabilities satisfying Piatetsky-Shapiro's argument, then  $X \Rightarrow Y$  can be extracted as a causal rule of interest if there are enough point-pairs  $(x_j, y_i)$  in  $M_{Y|X}$  such that  $p(x_j \cup y_i) \geq minsup, p(Y = y_i|X = x_j) \geq minconf$  and  $(p(Y = y_i|X = x_j) - p(Y = y_i)) \geq \lambda$ .

For example, let  $minsup = 0.3, minconf = 0.6, \gamma > 2, \alpha > 2, \lambda = 0.1$  and  $\eta = 2$ .  $QI = \{Doctor, Master, UnderMaster, [3500, +\infty), [2100, 3500), [0, 2100)\}$  in Example 1.  $X = \{Doctor, Master, UnderMaster\}, |R(X)| = 3, Y = 5\{3500, +\infty), [2100, 3500), [0, 2100)\}$  and  $|R(Y)| = 3$ . On the basis of the above definition,  $X \Rightarrow Y$  with  $M_{Y|X}$  can be extracted as a causal rule of interest.

*Theorem 1: Let  $QI$  be the set of quantitative items in database  $D$ ,  $X$  and  $Y$  be item variables,  $R(X) \subset QI, |R(X)| = n, R(Y) \subset QI, |R(Y)| = m, R(X) \cap R(Y) = \emptyset$ .  $minsup, minconf, \gamma > 0, \alpha > 0, \lambda > 0$  and  $\eta > 0$  are given by users or experts. Let*

- (1)  $S_{\text{support}} = \{(x_j, y_i) | p(x_j \cup y_i) \geq \text{minsupp} \wedge (1 \leq i \leq m) \wedge (1 \leq j \leq n)\}$ ,  
 (2) Let  $S_{\text{conf}} = \{(x_j, y_i) | p(Y = y_i | X = x_j) \geq \text{minconf} \wedge (1 \leq i \leq m) \wedge (1 \leq j \leq n)\}$ , and  
 (3) Let  $S_{\text{depend}} = \{(x_j, y_i) | (p(Y = y_i | X = x_j) - p(Y = y_i)) \geq \lambda \wedge (1 \leq i \leq m) \wedge (1 \leq j \leq n)\}$ .

The causal rule  $X \Rightarrow Y$  is of interest if and only if  $|S_{\text{support}} \cap S_{\text{conf}} \cap S_{\text{depend}}| \geq \min\{n, m, \gamma, \eta, \alpha\}$ .

*Proof:* We first prove ( $\Rightarrow$ ). According to the definition of rule of interest, there are enough point-pairs  $(x_j, y_i)$  in  $M_{Y|X}$  such that  $p(x_j \cup y_i) \geq \text{minsupp}$ ,  $p(Y = y_i | X = x_j) \geq \text{minconf}$  and  $(p(Y = y_i | X = x_j) - p(Y = y_i)) \leq \lambda$ . Or:

$$|S_{\text{support}} \cap S_{\text{conf}} \cap S_{\text{depend}}| \geq \min\{n, m, \gamma, \eta, \alpha\}.$$

Now we prove ( $\Leftarrow$ ). Because  $|S_{\text{support}} \cap S_{\text{conf}} \cap S_{\text{depend}}| \geq \min\{n, m, \gamma, \eta, \alpha\}$ , so:

$$\begin{aligned} |S_{\text{support}}| &\geq \min\{n, m, \gamma\}, \\ |S_{\text{conf}}| &\geq \min\{n, m, \eta\}, \\ |S_{\text{depend}}| &\geq \min\{n, m, \alpha\}. \end{aligned}$$

That is, the causal rule  $X \Rightarrow Y$  is of interest.

## Algorithms

Let  $D$  be the given database,  $\text{minsupp}$ ,  $\text{minconf}$ ,  $\alpha$ ,  $\lambda$ ,  $\eta$ : threshold values given by user. Our algorithm of mining causal rules in databases is as follows.

*Algorithm 1: CausalityDB*

**Input:**  $D$ : database,  $\text{minsupp}$ ,  $\text{minconf}$ ,  $\alpha$ ,  $\lambda$ ,  $\eta$ : threshold values;

**Output:**  $X \Rightarrow Y$ : causal rule,  $M_{Y|X}$ : the conditional probability matrix of  $Y$  given  $X$ ;

- (1) call Procedure 3;
- (2) for  $X, Y \in OIV$  do
  - for each element  $a$  in  $R(X)$  and  $b$  in  $R(Y)$  do
    - let  $p(Y = b | X = a) \leftarrow p(Y = b \wedge X = a) / p(X = a)$ ;
    - let  $CRSET \leftarrow$  the rule  $X \Rightarrow Y$  as a candidate rule;
    - with conditional probability matrix of  $Y$  given  $X$ :  $M_{Y|X}$ ;
    - endif
  - endfor

- (3) **for** each extracted rules  $R$  with  $M_{Y|X}$  in  $CRSET$  **do**egin  
**let**  $S_{\text{support}} \leftarrow \{(x_j, y_i) | p(x_j \cup y_i) \geq \text{minsupp} \wedge (1 \leq i \leq m) \wedge (1 \leq j \leq n)\}$ ;  
**let**  $S_{\text{conf}} \leftarrow \{(x_j, y_i) | p(Y = y_i | X = x_j) \geq \text{minconf} \wedge (1 \leq i \leq m) \wedge (1 \leq j \leq n)\}$ ;  
**let**  $S_{\text{depend}} \leftarrow \{(x_j, y_i) | (p(Y = y_i | X = x_j) - p(Y = y_i)) \geq \alpha \wedge (1 \leq i \leq m) \wedge (1 \leq j \leq n)\}$ ;  
**if**  $|S_{\text{support}}| < \min\{n, m, \gamma\}$  **then**  
**generate** item-based rules or quantitative rules for  $S_{\text{support}}$ ;  
**else if**  $|S_{\text{conf}}| < \min\{n, m, \eta\}$  **then**  
**generate** item-based rules or quantitative rules for  $S_{\text{conf}}$ ;  
**else if**  $|S_{\text{depend}}| < \min\{n, m, \alpha\}$  **then**  
**generate** item-based rules or quantitative rules for  $S_{\text{depend}}$ ;  
**else let**  $NewCRSET \leftarrow$  the rule  $X \Rightarrow Y$  as an interest rule;  
with conditional probability matrix of  $Y$  given  $X : M_{Y|X}$ ;  
**enddo**;  
(4) **call**  $\text{RefineRules}(NewCRSET, RSET)$ ;  
(5) **output**  $RSET$ ;  
**endall**.

*Theorem 2: For given large database  $D$ ,  $\text{minsupp}$  and  $\text{minconf}$  are given by users. If  $A \rightarrow B$  is extracted as a rule in Algorithm 1, then it is of interest under the partition.*

*Proof:* There are three kinds of rules extracted in Algorithm 1. We need to prove each kind of rules are of interest under our partition.

- (i) If  $A$  and  $B$  itemsets,  $A \rightarrow B$  is of interest under the partition according to Piatetsky-Shapiro's argument in Piatetsky-Shapiro (1991) and Theorem 1.
- (ii) If  $A$  and  $B$  are quantitative items,  $A \rightarrow B$  is of interest under the partition according to Piatetsky-Shapiro's argument in Piatetsky-Shapiro (1991) and Theorem 1.
- (iii) If  $A$  and  $B$  are item variables, it can be directly proved that  $A \rightarrow B$  is of interest under the partition by (4) Algorithm 1 and Piatetsky-Shapiro's argument Piatetsky-Shapiro (1991).

*Theorem 3: For given large database  $D$ ,  $\text{minsupp}$  and  $\text{minconf}$  are given by users. Let  $R_I$  and  $R_C$  be the set of item-based rules or quantitative rules, and rules in Algorithm 1, respectively, then  $R_I$  is generalized by  $R_C$ .*

*Proof:* According to Procedure 3, the partition on items and the partition on quantitative items are all good partitions. This means that, for any rule  $A \rightarrow B$  in  $R_I$ , it satisfies

- (a)  $A$  and  $B$  are all itemsets,  $A \rightarrow B$  is in  $R_C$  or, generalized in a certain quantitative association rule in  $R_C$  under the partition for items, or generalized in a certain causal rule  $X \rightarrow Y$  under the partition for quantitative items, where  $A$  and  $B$  are generalized in two certain quantitative items in  $R(X)$  and  $R(Y)$ , respectively.
- (b)  $A$  and  $B$  are all quantitative items,  $A \rightarrow B$  is in  $R_C$ , or generalized in a certain causal rule  $X \rightarrow Y$  under the partition for quantitative items.

Consequently, then  $R_I$  is generalized by  $R_C$ .

## OPTIMIZING CAUSAL RULES

Because our partition on data is blind relative to a given database, the space of point-values is difficult to fit into applications. Or, if only probabilities of some point-values are greater than or equal to *minconf*, then others can be taken as unnecessary information in the conditional probability matrix. Hence, we present a method to reduce such unnecessary information in their matrices once causal rules are extracted in this section.

### Unnecessary Information

For a conditional probability matrix, if the probabilities of a row (or a column) satisfy  $p(Y = y_i | X = x_j) < \text{minconf}$  for  $i = i_0$  and  $j = 1, 2, \dots, m$  (or  $i = 1, 2, \dots, n$  and  $j = j_0$ ), then this row (or column) is called *unnecessary information* in a relationship matrix. For example, let us partition the domains of *Education* and *Salary* as:

$$\{PostDoctor, Doctor, PostMaster, Master, Bachelor, UnderBachelor\}$$

and

$$\{3500, +\infty), [2400, 3500), [0, 2400)\}$$

respectively, and  $X$  and  $Y$  stand for *Education* and *Salary*, respectively. And the statistical results are from a database as shown in Table 4.

According to the above model,  $X \rightarrow Y$  can be extracted as a causal rule with the conditional probability matrix  $M_{Y|X}$  as follows:

$$M_{Y|X}^1 = \begin{bmatrix} 0.9 & 0.09 & 0.01 \\ 0.8 & 0.19 & 0.01 \\ 0.1 & 0.75 & 0.15 \\ 0.31 & 0.31 & 0.38 \\ 0.24 & 0.38 & 0.38 \\ 0.2 & 0.4 & 0.4 \end{bmatrix}$$

In this conditional probability matrix, if let  $minconf = 0.6$ , then  $p(Y = y_i | X = Master) < minconf$ ,  $p(Y = y_i | X = Bachelor) < minconf$  and  $p(Y = y_i | X = UnderBachelor) < minconf$ . That is, rows: 4, 5, and 6 are certainly unnecessary information. In fact, when given evidences are as: (0,0,0,1,0,0), (0,0,0,0,1,0), and (0,0,0,0,0,1), the reasoning results: (0.31, 0.31, 0.38), (0.24, 0.38, 0.38) and (0.2, 0.4, 0.4) can't be useful to applications. They would be reduced as possible. We now define a model to polish the extracted causal rules as follows:

### Merging Unnecessary Information

The problem of merging unnecessary information can be formally described as follows: Let  $X \rightarrow Y$  be an extracted causal rule with conditional probability matrix  $M_{Y|X}$ , the domain of  $X$  be  $R(X) = \{x_1, x_2, \dots, x_m\}$ , and the domain of  $Y$  be  $R(Y) = \{y_1, y_2, \dots, y_n\}$ .

- (1) finding out all columns  $i_1, i_2, \dots, i_s$  that any column  $i_k$  holds  $p(y_{i_k} | x_j) < minconf$  for  $j = 1, 2, \dots, m$  and  $k = 1, 2, \dots, s$ .
- (2) Merging all columns  $i_1, i_2, \dots, i_s$  into column  $i_1$  if  $s > 1$ , and delete columns  $i_2, \dots, i_s$  from  $M_{Y|X}$ .
- (3) finding out all rows  $i_1, i_2, \dots, i_t$  that any row  $i_k$  holds  $p(y_i | x_{i_k}) < minconf$  for  $j = 1, 2, \dots, n$  and  $k = 1, 2, \dots, t$ .
- (4) Merging all rows  $i_1, i_2, \dots, i_t$  into row  $i_1$  if  $t > 1$ , and delete rows  $i_2, \dots, i_t$  from  $M_{Y|X}$ .

*Procedure 4: RefineRules (NewCRSET: the set of interest causal rule, RSET: the extracted rule)*

**Input:** *NewCRSET: the set of interest causal rule;*

**Output:** *RSET: the set of optimized causal rules;*

- (1) **begin**  
**let**  $RSET \leftarrow \theta$ ;  
**for** each rule  $X \rightarrow Y$  with  $M_{Y|X}$  in *NewCRSET* **do begin**

- (2) **let**  $col \leftarrow \theta$ ;  
**for** each column  $i$  of  $M_{Y|X}$  **beginfor**  
**for**  $j := 1$  **to**  $m$  **do**  
**if**  $p(y_i|x_j) \geq minconf$  **then**  
**next**  $i$   
**let**  $col \leftarrow col \cup \{i\}$ ;  
**endfor**
- (3) **for**  $j := 1$  **to**  $m$  **do**  
**let**  $p_j \leftarrow 0$ ;  
**for** each  $i \in col$  **do**  
**for**  $j := 1$  **to**  $m$  **do**  
**let**  $p_j \leftarrow p_j + p(y_i|x_j)$ ;  
**for**  $j := 1$  **to**  $m$  **do**  
**let**  $p_j \leftarrow p_j/|col|$ ;  
**for** each  $i \in col$  **do**  
**delete** column  $i$  from matrix  $M_{Y|X}$ ;  
**add**  $(p_1, p_2, \dots, p_m)$  as a new column of  $M_{Y|X}$ ;
- (4) **let**  $r \leftarrow \theta$ ;  
**for** each row  $i$  of  $M_{Y|X}$  **beginfor**  
**for**  $j := 1$  **to**  $n$  **do**  
**if**  $p(y_i|x_i) \geq minconf$  **then**  
**next**  $i$   
**let**  $r \leftarrow r \cup \{i\}$ ;  
**endfor**
- (5) **for**  $j := 1$  **to**  $n$  **do**  
**let**  $p_j \leftarrow 0$ ;  
**for** each  $i \in r$  **do**  
**for**  $j := 1$  **to**  $n$  **do**  
**let**  $p_j \leftarrow p_j + p(y_j|x_i)$ ;  
**for**  $j := 1$  **to**  $n$  **do**  
**let**  $p_j \leftarrow p_j/|r|$ ;  
**for** each  $i \in r$  **do**  
**delete** row  $i$  from matrix  $M_{Y|X}$ ;  
**add**  $(p_1, p_2, \dots, p_n)$  as a new row of  $M_{Y|X}$ ;
- (6) **let**  $RSET \leftarrow$  optimized rule  $X \rightarrow Y$  with  $M_{Y|X}$ ;  
**enddo**  
**end**;

We now demonstrate the use of this model with the above rule as follows. Actually, when  $X = Master$ ,  $X = Bachelor$ , and  $X = UnderBachelor$ , we cannot determine which salary he is possible to earn. In order to reduce this unnecessary information, we can merge these three quantitative items into a

TABLE 5 Statistical Results of Interest Data

Education	Salary	Number
PostDoctor	[3500, +∞)	9000
	[2100, 3500)	900
	[0, 2100)	100
Doctor	[3500, +∞)	8000
	[2100, 3500)	1900
	[0, 2100)	100
PostMaster	[3500, +∞)	1000
	[2100, 3500)	7500
	[0, 2100)	1500
M&U	[3500, +∞)	7500
	[2100, 3500)	10900
	[0, 2100)	11600

quantitative item  $M \& U$ . Hence, the above data can be reduced as shown in Table 5.

Hence, for causal rule  $X \rightarrow Y$  is with the conditional probability matrix  $M_{Y|X}$  as follows. And the domain of  $X$  is  $R(X) = \{PostDoctor, Doctor, PostMaster, M\&U\}$ , the domain of  $Y$  is  $R(Y) = \{[3500, +\infty), [2100, 3500), [0, 2100)\}$ .

$$M_{Y|X}^2 = \begin{bmatrix} 0.9 & 0.09 & 0.01 \\ 0.8 & 0.19 & 0.01 \\ 0.1 & 0.75 & 0.15 \\ 0.25 & 0.363 & 0.387 \end{bmatrix}$$

*Theorem 4: The merge of row unnecessary information is reasonable.*

*Proof:* In the above merge, rows  $i_1, i_2, \dots, i_t$  are called unnecessary information, if all of them satisfy  $p(y_j|x_{i_k}) < minconf$  for  $j = 1, 2, \dots, n$  and  $k = 1, 2, \dots, t$ . Rows  $i_1, i_2, \dots, i_t$  are all merged as row  $i_1$  with  $p(y_j|x_i) = (p(y_j|x_{i_1}) + p(y_j|x_{i_2}) + \dots + p(y_j|x_{i_t}))/t$ , for  $j = 1, 2, \dots, n$ . Certainly,

$$\begin{aligned} \sum_{j=1}^n p(y_j|x_{i_1}) &= (p(y_1|x_{i_1}) + p(y_1|x_{i_2}) + \dots + p(y_1|x_{i_t}))/t \\ &\quad + (p(y_2|x_{i_1}) + p(y_2|x_{i_2}) + \dots + p(y_2|x_{i_t}))/t \\ &\quad + \dots \\ &\quad + (p(y_n|x_{i_1}) + p(y_n|x_{i_2}) + \dots + p(y_n|x_{i_t}))/t \\ &= (1 + 1 + \dots + 1)/t = 1. \end{aligned}$$

Hence, the merge of row unnecessary information is reasonable.

### Merging Items with Identical Properties

As we have seen, this merging quantitative items can improve the probability matrix. On the other hand, for when  $X = PostDoctor$  and  $X = Doctor$ , we can determine that his/hers salary is in  $[3500, +\infty)$  with higher confidence. Such quantitative items are called **items with identical property**. In the same reason of reducing this redundant, we can merge these two quantitative items into a quantitative item  $P \& D$ . Hence, the above data can be reduced as shown in Table 6.

**TABLE 6** Statistical Results of Interest Data

Education	Salary	Number
P&D	$[3500, +\infty)$	17000
	$[2100, 3500)$	2800
	$[0, 2100)$	200
PostMaster	$[3500, +\infty)$	1000
	$[2100, 3500)$	7500
	$[0, 2100)$	1500
M&U	$[3500, +\infty)$	7500
	$[2100, 3500)$	10900
	$[0, 2100)$	11600

Hence, for causal rule  $X \rightarrow Y$  is with the conditional probability matrix  $M_{Y|X}$  as follows. And the domain of  $X$  is  $R(X) = \{P \& D, PostMaster, M \& U\}$ , the domain of  $Y$  is  $R(Y) = \{[3500, +\infty), [2100, 3500), [0, 2100]\}$

$$M_{Y|X}^3 = \begin{bmatrix} 0.85 & 0.14 & 0.01 \\ 0.1 & 0.75 & 0.15 \\ 0.25 & 0.363 & 0.387 \end{bmatrix}$$

Also, the problem of merging such quantitative items can be formally described as follows. Let  $X \rightarrow Y$  be an extracted causal rule with conditional probability matrix  $M_{Y|X}$ , the domain of  $X$  be  $R(X) = \{x_1, x_2, \dots, x_m\}$ , the domain of  $Y$  be  $R(Y) = \{y_1, y_2, \dots, y_n\}$ .

- (1) finding out all columns  $i_1, i_2, \dots, i_s$  that column  $i_k$  holds  $p(y_{i_k}|x_j) \geq minconf$  at some  $j$  ( $1 \leq j \leq m$ ) and  $k = 1, 2, \dots, s$ .
- (2) Merging all columns  $i_1, i_2, \dots, i_s$  into column  $i_1$  if  $s > 1$ , and delete columns  $i_2, \dots, i_s$  from  $M_{Y|X}$ .
- (3) finding out all rows  $i_1, i_2, \dots, i_t$  that row  $i_k$  holds  $p(y_j|x_{i_k}) \geq minconf$  at some  $j$  ( $1 \leq j \leq m$ ) and  $k = 1, 2, \dots, t$ .
- (4) Merging all rows  $i_1, i_2, \dots, i_t$  into row  $i_1$  if  $t > 1$ , and delete row  $i_2, \dots, i_t$  from  $M_{Y|X}$ .

The algorithm for merging quantitative items with identical properties is similar to the above procedure of reducing unnecessary information, so we omit it here.

*Theorem 5: The merge of column unnecessary information is reasonable.*

*Proof:* The proof is as similar to Theorem 4.

Previously, we have illustrated that causal rules are useful for reasoning and decision under certainty. This usefulness continues in our polished rules. When such rules are applied, the probabilities of the merged point-values are added together as the probability of a new point-value that is used to substitute for the merged point-values. For example, let an evidence be  $(0.7, 0.1, 0.08, 0.02, 0.02)$  for  $M_{Y|X}^1$ . If it is used to inference in  $M_{Y|X}^3$ , the evidence  $(0.7, 0.1, 0.08, 0.08, 0.02, 0.02)$  needs to merge into  $(0.8, 0.08, 0.12)$ . In this way, we can obtain  $Y = (0.718, 0.216, 0.066)$ . Or if 0.718 is the probability that he/she earned salary in  $[3500, +\infty)$ , 0.216 is the probability that he/she earned salary in  $[2100, 3500)$ , and 0.066 is the probability that he/she earned salary in  $[0, 2100)$ .

*Theorem 6: For given large database  $D$ , minsupp and minconf are given by users. If  $X \rightarrow Y$  is extracted as a rule in our model, then  $X \rightarrow Y$  is an optimized rule.*

*Proof:* There are two kinds of causal rules.

- (1) If  $X$  and  $Y$  are all either itemsets or quantitative items,  $X \rightarrow Y$  is certainly an optimized rule.
- (2)  $X$  and  $Y$  are all item variables, then a conditional probability matrix  $M_{Y|X}$  is attached to this rule. And unnecessary information in  $M_{Y|X}$  is merged on rows and columns of the matrix. And items with identical properties in  $M_{Y|X}$  are merged on rows and columns of the matrix. Consequently,  $M_{Y|X}$  is an optimal matrix in unnecessary information and identical properties. This means,  $X \rightarrow Y$  is certainly an optimized rule.

## COMPARISON AND CONCLUSION

As have seen, we proposed a new approach for mining causality among multi-value variables in large databases based on partition in this paper, in which the causality is represented in the form  $X \rightarrow Y$  with a conditional probability matrix  $M_{Y|X}$  (Pearl 1988). To end this paper, a simple comparison with previous work and a summary of this paper are presented in this section.

## Related Work

Data mining (or KDD knowledge discovering in databases) has recently centered around the association rules (Agrawal, Imielinski, and Swami 1993; Brin, Motwani, and Silverstein 1997; Park, Chen, and Yu 1997; Srikant and Agrawal 1996; 1997). Mining association rules (Agrawal, Imielinski, and Swami 1993; Brin, Motwani, and Silverstein 1997) as the main task in KDD is to discover a set of strong association rules in the form of  $A \Rightarrow B$ , where  $A$  and  $B$  are disjoint sets of items (or  $A \cap B = \phi$ ). In order to implement this task, a wide range of problems have been investigated over such diverse topics as models for discovering generalized association rules (Brin, Motwani, and Silverstein 1997; Srikant and Agrawal 1997), efficient algorithms for mining association rules (Park, Chen, and Yu 1997), measurements of interest (Agrawal, Imielinski, and Swami 1993; Brin, Motwani, and Silverstein 1997; Srikant and Agrawal 1997), mining quantitative rules (Han, Cai, and Cercone 1993; Miller and Yang 1997; Srikant and Agrawal 1996), and association rule-based query languages and parallel data mining for association rules (Shintani and Kitsuregawa 1998). There are some excellent surveys (Chen, Han, and Yu 1996; Frawley, Piatetsky-Shapiro, and Matheus 1992) on these research findings.

As we have seen, the efforts on mining association rules (Agrawal, Imielinski, and Swami 1993; Piatetsky-Shapiro 1991) and quantitative association rules (Han, Cai, and Cercone 1993; Srikant and Agrawal 1996) assist causal discovery. On the other hand, there has been significant work in mining causal relationships using Bayesian analysis (Cooper 1997; Heckerman, Geiger, and Chickering 1995).

In Bayesian learning techniques such as in Cooper (1997); Heckerman, Geiger, and Chickering (1995); Pearl (1988), the user typically specifies a prior probability distribution over the space of possible Bayesian networks. These algorithms then search for that network which maximizes the posterior probability of the data provided. In general, they try to balance the complexity of the network with its fit to the data.

The main differences of our causality mining model from previous methods are simply listed as follows:

- (1) Item-based association rules and quantitative association rules can be still discovered as two special causal forms in our model. Or discovering causal rules is a generalization of mining item-based association rules and quantitative association rules. In particular, *good partition* on items as a concept is proposed.
- (2) Presenting a model to partition the quantitative items into item variables, which each item variable is a set of some quantitative items with same

property (or attribute). In particular, *good partition* on quantitative items as a concept is proposed.

- (3) Causality in databases is mined and represented of the form  $X \rightarrow Y$  with conditional probability matrix  $M_{Y|X}$ .
- (4) Our model of discovering causality in databases are focused on the causal rules of interest, which satisfy three conditions: Piattetsky-Shapiro's argument, minimum support, and minimum confidence.
- (5) Another main difference of our model is to offer a method for optimizing conditional probability matrixes of causal rules, which merges the unnecessary information in extracted causal rules. Apparently, this model can be used to optimize the knowledge in intelligent systems.

## Conclusion

Mining causality among variables in large databases is very useful for practical applications such as making decisions and planning. It is still an important and a prevailing topic in machine learning. Accordingly, some mining models for causality in databases, such as the LCD algorithm (Cooper 1997) and the CU-path algorithm (Silverstein et al., 1998), which are constraint-based causal discovery, have been proposed for mining causal relationships in market basket data. However, they in fact are of only limited efficiency on causal relationships between items. Discovering causality among variables in large databases is still pending further exploration. For this reason, we built a model in this paper of mining such causality in large databases based on partitioning. The key points of this paper are as follows:

- (1) Definition of "property tolerant" and "associated tolerant".
- (2) Presentation of a new model to partition the items into item variables for a given database, which decomposes the "bad item variables" and composes the "not-good item variables".
- (3) Proposal of a model of discovering causal rules of interest from databases based on this partition. In particular, item-based association rules and quantitative association rules are also discovered as the specific forms.
- (4) Establishment of a method of purifying the causal rules, which reduce unnecessary information in their matrices once causal rules are extracted.

## REFERENCES

- Agrawal, R., T. Imielinski, and A. Swami. 1993. Mining association rules between sets of items in large databases. *Proceedings of the AGM SIGMOD Conference on Management of Data*: 26–28 May 1993, Washington, DC, 207–216.
- Agrawal, R., and R. Srikant. 1994. Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conference*: 12–15 September 1994, Santiago de Chile, Chile, 487–499.

- Brin, S., R. Motwani, and C. Silverstein. 1997. Beyond market baskets: Generalizing association rules to correlations. *Proceedings of the ACM SIGMOD International Conference on Management of Data*: 13–15 May 1997, Tucson, AZ, 265–276.
- Chen, M., J. Han, and P. Yu. 1996. Data Mining: An overview for a database perspective. *IEEE Trans. Knowledge and Data Eng.* 8(6):866–881.
- Chen, M., J. Park, and P. Yu. 1998. Efficient data mining for path traversal patterns. *IEEE Trans. Knowledge and Data Eng.* 10(2):209–221.
- Cooper, G. 1997. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery* 1(2): 203–224.
- Fayyad, U., and P. Stolorz. 1997. Data mining and KDD: Promise and challenges. *Future Generation Computer Systems* 13:99–115.
- Frawley, W., G. Piatetsky-Shapiro, and C. Matheus. 1992. Knowledge discovery in databases: An overview. *AI Magazine* 13(3):57–70.
- Han, J., Y. Cai, and N. Cercone. 1993. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowledge and Data Eng.* 5(1):29–40.
- Han, E., G. Karypis, and V. Kumar. 1997. Scalable parallel data mining for association rules. *Proceedings of the ACM SIGMOD International Conference on Management of Data*: 13–15 May 1997, Tucson, AZ, 277–288.
- Heckerman, D., D. Geiger, and D. Chickering. 1995. Learning Bayesian networks: The combinations of knowledge and statistical data. *Machine Learning* 20:197–243.
- Miller, R., and Y. Yang. 1997. Association rules over interval data. *Proceedings of the ACM SIGMOD International Conference on Management of Data*: 13–15 May 1997, Tucson, AZ, 452–461.
- Park, J., M. Chen, and P. Yu. 1997. Using a hash-based method with transaction trimming for mining association rules. *IEEE Trans. Knowledge and Data Eng.* 9(5):813–824.
- Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers.
- Piatetsky-Shapiro, G., 1991. Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. Frawley, eds., AAAI Press/MIT Press, 229–248.
- Shintani, T., and M. Kitsuregawa. 1998. Parallel mining algorithms for generalized association rules with classification hierarchy. *Proceedings of the ACM SIGMOD International Conference on Management of Data*: 2–4 June 1998, Seattle, WA, 25–36.
- Srikant, R., and R. Agrawal. 1996. Mining quantitative association rules in large relational tables. *Proceedings of the ACM SIGMOD Conference on Management of Data*: 4–6 June 1996, Montreal, Quebec, Canada, 1–12.
- Srikant, R., and R. Agrawal. 1997. Mining generalized association rules. *Future Generation Computer Systems* 13:161–180.
- Silverstein, C., S. Brin, R. Motwani, and J. Ullman. 1998. Scalable techniques for mining causal structures. *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*: 2–4 June 1998, Seattle, WA, 51–57.