Applied Mathematics & Information Sciences 5 (2) (2011), 39S-44S
– An International Journal

Efficiently Mining Positive Correlation Rules

Zhongmei Zhou
Department of Computer Science & Engineering, Zhangzhou Normal University, China

Email Address: zzm@zju.edu.cn

Received Jun 8, 2010; Revised Jan 3, 2011

One of the main tasks of knowledge discovery in Traditional Chinese Medicine is discovering novel paired or grouped drugs from the Chinese Medical Formula database. Novel paired or grouped drugs, which are special combinations of two or more kinds of drugs, have strong efficacy in clinical care. We use positive correlation rule mining to analyze the large number of complex correlation relationships among various kinds of drugs in order effectively to find novel paired or grouped drugs in the Chinese Medical Formula database. We firstly give some new related definitions and then develop an efficient algorithm for discovering all positive correlation rules based on frequency patterns from a large database. Experimental results on the Chinese Medical Formula database and the mushroom database show that all techniques developed in this paper are feasible.

Keywords: Positive correlation rule, frequent pattern, paired drug.

1 Introduction

Confronted with the increasing popularity of TCM (Traditional Chinese Medicine) and the huge volume of the TCM database, such as the CMF (Chinese Medical Formula) database, the TCM literature database and the TCM clinical database, there is an urgent need to explore these resources effectively by the techniques of knowledge discovery in a database (KDD) [5]. Since novel paired or grouped drugs, which are special combinations of two or more drugs, have extraordinary efficacy in clinical care, discovering novel paired or grouped drugs from the CMF database becomes one of the hot topics in TCM formula researches. In the CMF database there are many complex correlation relationships among various kinds of drugs. Therefore we use positive correlation rule mining to learn these correlation relationships in order to extract novel paired or grouped drugs from the CMF database effectively.

The most frequently employed method for correlation mining is that of two-dimensional contingency table analysis of categorical data using the chi-square statistic as a measure of significance. Sanjeev and Zytkow [4] analyze contingency tables to discover students who are poorly prepared for university level course and at risk of dropping out. Brin et al. [1] analyzed contingency tables to generate correlated patterns that identify statistical dependence in both the presence and absence of items in itemsets.

H. Liu et al. [2] used contingency tables to discover unexpected and interesting patterns that have a low level of support and a high level of correlation. Although in terms of the concept of the chi-square the low chi-squared value (less than cutoff, e.g. 3.84 at the 95% significance level [3]) efficiently indicates that all of XY, $X\overline{Y}$, $\overline{X}Y$, $\overline{XY}$ are independent, the high chi-squared value does not show that all XY, $X\overline{Y}$, $\overline{X}Y$, $\overline{XY}$ are dependent. Hence X and Y might be not dependent even if the chi-squared value is considerably higher. Interest $P(XY)/P(X)P(Y)$ [1] does not have proper bounds. In this paper we use the correlation confidence in [6] as a measure of correlation significance. Some new notions are given and an efficient algorithm is developed for discovering all positive correlation rules based on frequency patterns from large database. Experimental results on the CMF database and the mushroom database show that positive correlation rule mining is quite a necessary method to discovering novel paired or grouped drugs from the CMF database.

The remainder of this paper is organized as follows: In section 2 we firstly give some new related definitions and then develop an efficient algorithm for mining all positive correlation rules based on frequency patterns from a large database. In section 3 we show our experimental results. We conclude our study in section 4.

2 Mining Positive Correlation Rules

In this section we firstly give some new related definitions and then illustrate these definitions with an example. Finally we develop an algorithm for discovering all positive correlation rules based on frequency patterns from a large database effectively.

Definition 2.1 (correlation confidence) The correlation confidence of rule $X \leftrightarrow Y$, which we denote as c-conf, is defined as follows:

$$P(XY) - P(X)P(Y) \big/ P(XY) + P(X)P(Y) \tag{2.1}$$

Definition 2.2 (a correlation rule) Let $\eta > 0$ be the given minimum c-conf. If

$$\left| P(XY) - P(X)P(Y) \big/ P(XY) + P(X)P(Y) \right| \geq \eta , \tag{2.2}$$

then rule $X \leftrightarrow Y$ is called a correlation rule.

Definition 2.3 (a positive correlation rule) Let $\eta > 0$ be the given minimum c-conf. If

$$P(XY) - P(X)P(Y) \big/ P(XY) + P(X)P(Y) \geq \eta , \tag{2.3}$$

then rule $X \leftrightarrow Y$ is called a positive correlation rule.

Definition 2.4 (a positive correlation rule based on a frequency pattern) Let X be a frequency pattern and Y be a subset of X, $Z = X - Y$. Let $\eta > 0$ be the given minimum c-

conf. If

$$P(YZ) - P(Y)P(Z) \Big/ P(YZ) + P(Y)P(Z) \geq \eta, \quad (2.4)$$

then rule $Y \leftrightarrow Z$ is called a positive correlation rule based on frequency pattern X.

From definition 2.4 we have two conclusions. (1) If frequency pattern X contains more items, then we can generate more positive correlation rules based on pattern X. (2) If a positive correlation rule $Y \leftrightarrow Z$ is based on pattern X, then the size of rule $Y \leftrightarrow Z$ is equal to the size of pattern X.

We illustrate these definitions using the following example.

Example 2.1 Let the given minimum c-conf be 0.1 and the given minimum support be 0.4. For the example database in Table 2.1, we have

$$P(AC) - P(A)P(C) \Big/ P(AC) + P(A)P(C) = 1/4,$$
$$P(CE) - P(C)P(E) \Big/ P(CE) + P(C)P(E) = 1/19.$$

Thus rule $A \leftrightarrow C$ is a positive correlation rule and rule $C \leftrightarrow E$ is not a correlation rule. Moreover rule $A \leftrightarrow C$ is a positive correlation rule based on frequency pattern AC. Since item A and C have a positive correlation relationship, we have more reason to believe that AC has a higher probability than CE to be a novel paired drug.

Table 2.1 an example database.

id	Items
10	A, B, C
20	C, D, E
30	A, C, D, E
40	D, E
50	B, D

We mine all positive correlation rules in two steps. We firstly discover all frequency patterns and then derive all positive correlation rules based on these frequency patterns.

Algorithm: *Mining positive correlation rules based on frequency patterns.*

Input: a database DB, a given minimum support ξ and a given minimum correlation confidence, i.e. minimum c-conf η.

Output: the complete set of positive correlation rules based on all frequency patterns.

C_k : Candidate patterns of size k

L_k : Frequency patterns of size k

M_k : Positive correlation rules based on frequency patterns of size k

L_1 = {frequent items}

For ($k = 1; M_k \,! \neq \phi; k++$) do begin

C_{k+1} = candidates generated from $L_k * L_k$

For each record t in database do

increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1} = candidates in C_{k+1} with minimum support

For each pattern l_{k+1} in L_{k+1} and each subset Y of pattern l_{k+1}, $Z = l_{k+1} - Y$.

If $P(YZ) - P(Y)P(Z) \Big/ P(YZ) + P(Y)P(Z) \geq \eta$,

then generate positive correlation rule $Y \leftrightarrow Z$ based on frequency pattern l_{k+1}

Insert $Y \leftrightarrow Z$ into M_{k+1} Return $\cup M_{k+1}$

3 Experimental Results

All experiments are performed on two kinds of datasets: 1. Mushroom characteristic database. 2. The CMF database, which consists of 4,643 formulas with 21689 kinds of drugs involved.

Minsup, Max_len, P_mum, Rule_num and minc_conf represent the minimum support, the maximum size of all frequency patterns, the number of all frequency patterns, the number of all positive correlation rules based on frequency patterns and the minimum c_conf respectively. Table 3.1 shows the number of positive correlation rules varied when the minimum support decreases with fixed minimum correlation confidence of 1% . From Table 3.1 we can see if a frequency pattern contains more items. Then we can generate more positive correlation rules based on this frequency pattern.

Table 3.2 and Table 3.3 have the same attributes. Table 3.2 and Table 3.3 show the number of positive correlation rules varied on the mushroom database and on the CMF database respectively as the given minimum correlation confidence c-conf increases with a fixed minimum support. From Table 3.2 we can see that, when the given minimum correlation confidence is 30% , there are only 3 positive correlation rules based on all frequency patterns on the mushroom database. However, from Table 3.3 we can see that, when the given minimum correlation confidence is 50% , there are 701 positive correlation rules based on all frequencies on the CMF database. The CMF database is very sparse. There are fewer frequency patterns even if the given minimum support is small enough. Experimental results show that there are a great many of positive correlation relationships among various kinds of drugs on the CMF database. Therefore we can conclude that positive correlation rule mining is quite a good method to find novel paired or grouped drugs from the CMF database.

Table 3.1 minimum c-conf 1%

minsup	Max_len	P_mum	Rule_ num
90	3	5	3
80	4	18	15
70	5	26	48
60	5	43	60
50	5	140	213
40	7	544	2208
30	9	2706	41628

Table 3.2 minimum support 0.4

minc_conf	Max_len	P_mum	Rule _num
5	7	544	1062
10	7	544	681
15	7	544	462
20	7	544	138
25	7	544	63
30	7	544	3
35	7	544	0

Table3.3. minimum support 0.02 (TCM_CMF database)

minc_conf	Max_len	P_mum	Rule_num
5	5	378	924
10	5	378	923
15	5	378	916
20	5	378	907
25	5	378	890
30	5	378	877
35	5	378	857
40	5	378	809
45	5	378	753
50	5	378	701

4 Conclusions

Positive correlation rules reflect positive correlation relationships and thus have more interest in discovering novel paired or grouped drugs from the CMF database. In this paper we proposed to mine all positive correlation rules based on frequency patterns from the CMF database in order to find novel paired or grouped drugs effectively. We developed an efficient algorithm for discovering all positive correlation rules based on frequency patterns from a large database. By our algorithm we can provide a large number of positive correlation rules for TCM experts to judge whether they are novel paired or grouped drugs. Experimental results show that all techniques developed in this paper are effective and efficient.

Acknowledgement

Research supported by a program of China NSF (No. 10971186) and a grant from the Education Ministry of Fujian of China (JA10202).

References

[1] S. Brin, R. Motwani and C. Silverstein, Beyond market basket: Generalizing association rules to correlations, *In Proc. ACM SIGMOD***.** (1997) 265-276.

[2] H. Liu, H. Lu, L. Feng and F. Hussain, Efficient search of reliable exceptions, *In Proc. PAKDD***.** (1999) 194-203.

[3] F.C. Mills, Statistical Methods, Pitman, London, 1955.

[4] A.P.Sanjeev and J.Zytkow, Discovering enrolment knowledge in university databases, *In Proc. ACM SIGKDD***.** (1995) 246-251.

[5] Yi Feng, Zhaohui Wu and Zhongmei Zhou, Combining an order-semisensitive text Similarity and closest fit approach to textual missing values in knowledge discovery, *KES LNAI***.** 3682 (2005) 943-949.

[6] Zhongmei Zhou, Zhaohui Wu, Chunshan Wang and Yi Feng, Mining both associated and correlated patterns, *ICCS LNCS***.** 3994 (2006) 468-475.

Zhongmei Zhou received the PhD degree in Computer Science from the College of Computer Science and Technology, Zhejiang University, in 2006. She is currently a Professor. Her research interests are in the areas of Data Mining, Distributed Systems and Database Systems.