International Journal of Biomedical Data Mining

Shang et al., Int J Biomed Data Min 2017, 6:1
DOI: 10.4172/2090-4924.1000126

Review Article OMICS International

Association Rules Mining and Statistic Test Over Multiple Datasets on TCM Drug Pairs

Shang E*, Duan J, Fan X, Tang Y and Ye L

Jiangsu Key Laboratory for TCM Formulae Research, Nanjing University of Chinese Medicine, Nanjing, China

Abstract

Objective: TCM drug pair is consisted of two and only two drugs, which is the smallest drug group following special drug compatibility regulations. Formulae compatibility regulations are one of the most important problems in TCM clinical practice and modern research but still not quite resolved. TCM drug pair was a very suitable objects to discovery the complicated formulae compatibility regulations. This paper applied association rules mining to study the structural characters of TCM drug pairs find some special relationships between drugs. This study might give some help to the research on the formulae compatibility regulations.

Methods: We presented an enhanced association rules mining method to find out the property associations between two drugs in TCM drug pairs. And a binominal statistic test was introduced to get the statistical significance of rules mined. The property data from the 625 drug pairs containing 347 drugs were collected and analyzed. As most association rules mining run only in single database, the new method was proposed to find rules over multiple databases (2 in this paper standing for the two drugs in TCM drug pairs) based on a first Apriori algorithm mining. Then statistic test was applied to filter out insignificant rules furthermore.

Results: Apriori algorithm and the new method were applied to mine association rules on TCM drug pairs for comparison. The rules found by Apriori method showed false high support, part of which came from the property associations within one drug but not between the two drugs in TCM drug pairs. And Apriori method could not found the association of replicated property, such as *liver - liver* rules. The new method proposed could get the only associations between drugs even those replicated property rules. Some associations were mined with high supports and significances.

Conclusion: This paper proposed an enhanced method to perform association rules mining over multiple databases. After comparison with Apriori algorithm the new method could just obtain the associations in which each item came from different database. The method was confirmed to be quite suitable on mining over multiple databases. The statistic test was also necessary to exclude false association rules.

Keywords: Drug pair; Association rule; Apriori algorithm; Traditional Chinese medicine, Data mining

Introduction

Traditional Chinese medicine (TCM) has appeared and developed over more than 4000 years, which is derived from amounts of clinical observation and empirical evidence. A large amount of valuable experience and knowledge has been recorded and summarized. In TCM theory, some properties of TCM drugs, such as flavors, natures, channel tropisms, efficacy and compatibilities, have been defined, applied and verified in clinical practices. However, TCM theory is based on ancient Chinese philosophy frameworks such as the Theory of Yin-Yang and Five Elements. And these theory descriptions can hardly intercommunicate with modern medical science.

Data mining methods including text mining and knowledge discovery can help researchers to identify required information and discover new relationships efficiently from huge amounts of data, which may bridge the gaps between modern medicine and TCM. Both computational and experimental efforts have been made to relate the TCM components to those of modern medicine, in order to demonstrate the effectiveness and mechanisms of TCM scientifically. As the basis for data mining, some TCM databases has been developed including chemical ingredients, pharmacological information and 3D structure [1-3].

Knowledge discovery in databases (KDD) is one proper methodology to analyze and understand the underlying TCM knowledge in the form of ancient books and literatures [4]. Zhou et al. [5] presented an approach to integrate TCM literature with modern biomedical data to discover novel gene networks and functional knowledge of genes. Wang et al. [6] proposed a novel method called TCMSP (traditional Chinese medicine syndrome prediction) to select critical features for syndrome prediction for liver cirrhosis in traditional Chinese medicine. Chen et al. [7] analyzed use frequencies, the characteristics of TCM users, and the disease categories that were treated by TCM in Taiwan based on the complete datasets of TCM outpatient reimbursement claims from 1996 to 2001.

Association rule mining is one of the most important models, and is well researched in data mining area. It aims to discover all item associations (or rules) in the transaction database satisfying the user-specified minimum support (*minsup*), minimum confidence (*minconf*) or other constraints [8]. It was first introduced in 1993 [9] and then many modified methods had been proposed to enhance the efficiency

***Corresponding author:** Shang E, Jiangsu Key Laboratory for TCM Formulae Research, Nanjing University of Chinese Medicine, Nanjing 210023, China, Tel: +86 25 85811916; E-mail: sherxin@gmail.com

Received December 16, 2016; **Accepted** February 16, 2017; **Published** March 10, 2017

Citation: Shang E, Duan J, Fan X, Tang Y, Ye L (2017) Association Rules Mining and Statistic Test Over Multiple Datasets on TCM Drug Pairs. Int J Biomed Data Min 6: 126. doi: 10.4172/2090-4924.1000126

and ability of mining [10-18]. The Apriori algorithm was proposed firstly and was applied most widely for association rules mining. And then the widely applied Apriori algorithm was proposed [10]. After that various association mining techniques and algorithms has be developed to enhance mining efficiency. Techapichetvanich and Datta [11] presented a three-step visualization method for mining market basket association rules. Baralis and Psaila [12] defined a classification of association rule types, which provided a general framework for the design of association rule mining applications. Chen and Xi [13] presented a revised version of mutual information to discriminate positive and negative association. Liu et al. [14] proposed a technique to mining association rules with multiple minimum supports to reflect the natures of the items and their varied frequencies in the database. Brin et al. [15] measured significance of associations via the chi-squared test. Zhu et al. [16] conducted a case study in colorectal cancer management and the preliminary results showed that useful causal relations and decision alternatives can be extracted. Cristofor and Simovici [17] introduced a new rule of inference and defined the notion of association rules cover as a minimal set of rules that were non-redundant with respect to this new rule of inference. Han et al. [18] proposed a novel frequent pattern tree (FP-tree) structure, which was an extended prefix-tree structure for storing compressed and crucial information about frequent patterns, and developed an efficient FP-tree-based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth. However, most of algorithms focused either on the enhancement of mining efficiency or on the identification of interesting rules all based on single transaction database. There were few reports on the association rule mining over multiple databases or datasets. In this paper the association rules mining over two independent datasets was required on TCM drug pair data for which a new mining procedure was proposed.

In formulating TCM recipes, some special drug pairs have been frequently used for achieving mutual enhancement, mutual assistance, mutual restraint, mutual suppression, or mutual antagonism [19,20]. The drug pair is consisted of two drugs, following some special compatibility laws, and functions as a integrate unit. As the structure of drug pair was much simpler than TCM formulae, the study on drug pair may help to understand the complicated compatibility laws of TCM formulae. The compatibility laws in drug pairs could be illustrated partly through the associations between TCM drug properties from two drugs respectively such as flavors, nature, channel tropism and efficacy. Choong Yong Ung et al. [20] used artificial intelligence methods to examine whether there were distinguishable properties of drug pairs from those of non-drug pairs. These methods could classify drug pairs and non-drug pairs correctly, which meant drug pairs exhibited some special property patterns. So the association rules between the two drug properties should be interesting and to be mined. There were difficulties if applying popular association rule mining methods directly to perform the mining.

In this paper we proposed a strategy to mine association rules over multi-dataset on TCM drug pair data, followed two previous mining over single dataset. And then a probability test, natural probability test, was introduced to evaluate the significances of associations mined for filtering false ones.

Association Mining and Statistic Test

Definition of association rule and apriori algorism

An association rule [9-11] is a pair of the type $\mathbf{A}\Rightarrow\mathbf{B}$, where **A** and **B** is an item or item set in database. The implication of the rule is that **A** and **B** appears simultaneously in the same transaction. In mathematical words, an association rule is of the form $\mathbf{A}\Rightarrow\mathbf{B}$, where $\mathbf{A}, \mathbf{B} \Rightarrow \mathbf{I}$ and $\mathbf{A} \Rightarrow B = \varphi$. $I = \{i_1, i_2 \ldots i_n\}$ is a set of items or fields in the transaction database where i_j, $1 \le j \le n$, is an item in the database that may appear in a transaction. There are two common measures of interestingness, support (*sup*) and confidence (*conf*). The support *sup* for rule $\mathbf{A}\Rightarrow\mathbf{B}$ is the percentage of transactions that contain $\mathbf{A} \Rightarrow B$ (both A and B). This is also regarded as the probability $\mathbf{P}(\mathbf{A}\Rightarrow B)$. And $\mathbf{A}\Rightarrow\mathbf{B}$ has confidence $conf_1$ and $conf_2$ where $conf_1$ is the percentage of transactions containing **A** that also contain **B**, i.e., the conditional probability $\mathbf{P}(\mathbf{B}|\mathbf{A})$ and $conf_2$ as $\mathbf{P}(\mathbf{A}|\mathbf{B})$. Usually an interesting association rule should have both support and confidence above a user specified level.

Apriori algorism was one of the most widely applied methods in association rule mining [10]. The first pass of the algorithm simply counts item occurrences to determine the large 1-item sets. The subsequent pass was consisted of two phases. First was to generate the candidate large item sets (C_k) form k-1 item sets (L_{k-1}). Next the database was scanned to calculate the support and confidence for rules in C_k and to find frequent k-item sets (L_k) with support and confidence above user-specified minimum limitation. The phases were performed in cycle to mine all the interesting association rules.

Association rule mining over multiple datasets

A TCM drug pair is consisted of two drugs, either of which has the common properties of TCM drugs, such as flavors, natures, channel tropisms and efficacy. The structure of drug pairs, i.e., the combination mode of the two drugs, may contain some special compatibility laws. For studying their structure the objective and interesting associations should be formed with items from the two drugs respectively.

As most association mining methods run on single dataset, the two drug property datasets must be merged into one before mining. Then the constraint-based mining methods could be applied. However, this usual mining procedure could hardly find out meaningful and unrepeated rules from TCM drug pair properties. After mergence some property information might become redundancy, repeated or insufficient for constraint-based mining. A new procedure based on single dataset mining was proposed for association mining over multiple datasets. These datasets must have the same record number or row number, while might have the same or different properties or column names. The record at the same row of each dataset was considered to form a transaction. Firstly the association rules were mined on each dataset with any common method, such as Apriori, AprioriTid, FP-Tree, and so on [9-14] to form rule sets. Apriori algorism was selected in this paper for trial. Then secondly the associations between datasets were constructed and selected satisfying the user specified level based on previous rule sets.

The rule set $\mathbf{R}_i$, i={1, 2, …, n} and n as the total number of datasets, for dataset $\mathbf{D}_i$ respectively, was obtained by common Apriori method. The Cartesian product of $\mathbf{R}_i$, $\mathbf{R}_1\times\mathbf{R}_2\times\ldots\times\mathbf{R}_n$, was considered as candidate rule sets among these datasets. Then all the datasets were scanned row by row synchronously to determine the support and confidence of each candidate rule. The support (*sup*) was defined as the percentage of the rows over the all, at which the candidate rule could be satisfied in each dataset. The number of confidence (*conf*) was as many as the number n. $conf_i$ was the ratio of *sup* over the number of rows in $\mathbf{D}_i$ at which the candidate rule could be satisfied partly. Its meaning was similar with that defined in **2.1**. During mining all the candidate rule would be verified to find out those interesting and meaningful ones.

Merging for sequence-independent rules

When mining on single dataset, the association rules are unique with no relation to property sequence. That means, **A⇒B** and **B⇒A** cannot appear simultaneously. However, the rules only with different item sequence might appear when mining over multiple datasets. For example, if two datasets $\mathbf{D}_i$ and $\mathbf{D}_j$ contained same properties, **A** and **B**, the two properties might be selected in $\mathbf{R}_i$ and $\mathbf{R}_j$ respectively in the first mining. As the candidate rules over multi-dataset were from the Cartesian product of all **R** sets, which contained both the rule **A⇒B** and **B⇒A,** the two rules might satisfy the *sup* and *conf* limit and be selected out. If the arrangement of the two datasets had no special sequence, the two rules **A⇒B** and **B⇒A** should be redundant.

In our study the two drugs in a drug pair had the same importance and equal position, so the redundant rules should be eliminated. The rules only with different item sequence should be merged into one. After combination the *sup* and *conf* of the new rules would be recalculated through probability plus equation. If sup_1 and sup_2 were the support of **A⇒B** and **B⇒A,** the new *sup* after combination should be sup_1 + sup_2 – the joint probability of sup_1 and sup_2. The new confidence vector should be recalculated too according modified conditional probability equation. After combination the $conf_1$ should be the conditional probability of P({**A** in $\mathbf{D}_1$ ⇒ **B** in $\mathbf{D}_2$}⇒ {**A** in $\mathbf{D}_2$ ⇒ **B** in $\mathbf{D}_1$}|**A** in $\mathbf{D}_1$ ⇒**A** in $\mathbf{D}_2$). And conf should be P({**A** in $\mathbf{D}_1$ ⇒ **B** in $\mathbf{D}_2$}⇒ {**A** in $\mathbf{D}_2$ ⇒ **B** in $\mathbf{D}_1$} | **B** in $\mathbf{D}_1$ ⇒ **B** in $\mathbf{D}_2$). During the same dataset scan procedure the new sup and conf vector could be calculated simultaneously. The rules combination could be performed only when different dataset had the same property without special sequence.

Natural probability and statistic significance test for rules

In statistic meaning, the dataset could be considered as a sample of transactions population. Each item or property in the dataset has different population rate. For subset drown from population with some special characters, some items and rules (item sets) might show different frequencies compared with their population rates, more or less. These rules with different frequencies implied some special regulations or relations, which might be related to the subset characters. And those rules with the same frequencies with their population rates showed no subset characters and were meaningless, even if they had high supports or confidences. Therefore only the support and confidence limit could not filter out meaningful rules efficiently. Suitable statistic test should be applied for additional limitation. For example, the incidence rates of lung cancer were 61.8 for male and 24.4 for female per 100 000 standard population in Hong Kong in 2008, with the male to female ratio about 2.5:1 [20]. Normally the male to female ratio should be close to 1:1, much lower than that in lung cancer patients. It meant that there exited a special relationship between lung cancer and gender.

In our study the property dataset of drug pairs could be seen as a subset drawn from drug population. In the subset some property associations had different frequencies because of the special compatibility laws in TCM theory. In drug population some drugs with mild effects and little toxicity were applied in clinical therapy much more than those with strong effects or toxicity. These property frequencies in drug population were served as population rates of properties. In this paper the population rate, which we called *natural probability* of rules, were applied as a test standard. If the *sup* of a rule had a significant difference with its *natural probability*, the rule was regarded as meaningful rule. The difference, which was called adjusted sup, provided a corrected index to evaluate the importance of rules.

The statistical significance of rules could be tested with sup and natural probability. As there were only two statuses for a property in drug pair, the *sup* of rules should follow binominal distribution. The significance of a rule could be tested by population rate test. Usually the significance level was set at 0.05. If the probability calculated was less than 0.05, the rule was thought as significant.

Programming and data structure

All the programs were developed independently and written in Matlab script language. The software environment was Matlab R2009a for Win32 (Mathworks). The properties of drug pairs were arranged as a logical sparse matrix as input. The generation of frequent items and calculation of *sup* and *conf* were performed by logical bit-wise operation, such as AND, OR, and XOR. All the programs were available freely on request.

Results and Discussions

Data source and preparation

In TCM history there are many famous writings containing the descriptions of drug pairs, such as *Shennong Bencao Jing*, *Shanghan Zabing Lun*, *Bencao Gangmu*, and so on. All drug pairs recorded in these writings were collected exclude those duplicated, redundant, with alias names or consisted of rare drugs. There were 625 drug pairs containing 347 drugs included in the study. For each drug 49 properties were specified, including 5 natures of cold, hot, warm, cool and normal, 5 flavors of pungent, sweet, sour, bitter and salty, 12 channel tropisms and 27 effects.

The properties of a drug were described with a logical vector of 49 elements. If a drug had a special property, the value of its property vector in corresponding position was set at 1 (true), otherwise 0 (false). For example, Herba Ephedrae (Ma-Huang) with the properties of warm in nature, pungent and bitter in flavor, lung, kidney, bladder and heart in channel tropism and diaphoretic, damp-cleaning effects, had its logical vector with 1 at the 1st, 4th, 9th, 12th, 14th, 16th, 20th, 27th, 31st element and 0 at other places. The properties of all the 347 drugs were described as a logical sparse matrix with 347 rows and 49 columns. The properties of 625 drug pairs were listed in two matrices with 625 rows and 49 columns standing for the two drugs in pairs. Then the manipulation of properties could be easily performed by logical bit-wise operation.

Drug and property frequencies in TCM drug pairs

All the 625 drug pairs were consisted of 347 drugs and each drug had 49 properties. The frequencies of drugs and properties were analyzed firstly by standard Apriori method to find whether some special drugs or properties were used more frequently in drug pair compatibilities. The frequency trend of 347 drugs was based on a 625 × 347 matrix and shown in Figure 1 as follows. The largest frequency, from Rhizoma Coptidis, was 5.2% and the most drug frequencies were less than 1%. The distribution of drugs in TCM drug pairs was nearly homogeneous except several drugs such as Rhizoma Coptidis (Huang-Lian), Radix et Rhizoma Ginseng (Ren-Shen), Rhizoma Pinelliae (Ban-Xia). From Figure 1 it could be found that the composition of TCM drug pairs had nearly no preferences on different drugs.

The drug property frequencies in drug pairs were also analyzed following the same method as above on the 625 × 49 drug pair property matrix. The matrix was obtained by combination of two drug property matrices of drug pairs. The combination was performed with bit-wise

operation OR between two matrices. The frequency trend of all the 49 properties was shown in Figure 2 as follows.

Compared with drug frequencies, the distribution of drug property was more concentrated. The most frequent property was liver in channel tropisms with sup at 83.68%. Some properties in channel tropisms showed high frequencies, such as lung with sup at 79.52%, spleen at 78.88% and stomach at 68.48%. In all the properties with sup >20%, there were 7 in channel tropisms, 2 in flavors, 4 in natures and 5 in effects. The property frequencies had much larger sup than drug frequencies, however they showed slow descending trend. The property distribution could release more information than that of drugs, but not satisfying.

The property frequencies in drug profiles, not in drug pair profiles, were also calculated. The property matrix was consisted of 347 rows, which stand for 347 drugs, and 49 columns for properties. The sup of rules from the matrix could be regarded as the natural sup of drug property rules in TCM drug pairs and be prepared for significance test.

Association rules mining on drug pair property matrix

TCM drug pairs were consisted of two drugs, so the aimed property associations should be formed between drugs. Apriori method and the new procedure proposed in this paper were applied respectively for comparison.

Firstly Apriori method was used on the combined property matrix of drug pairs. After mining the top 10 frequent rules were list in Table 1.

The association liver - spleen had the highest support with 66.24%. The rules containing liver, spleen, lung and stomach had high sup, which took place 6 in top 10. Form Table 1 it could be found that the rules generated from channel tropisms appeared most frequently with high confidences. They might be thought as important or interesting rules standing for special drug pair compatibility regulations.

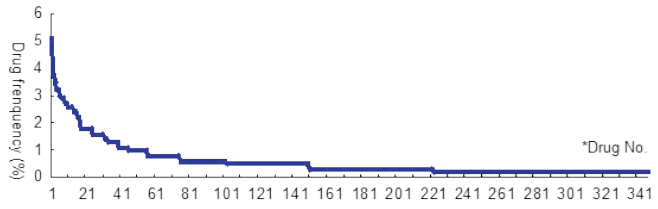

Figure 1: The drug frequency trend in TCM drug pairs.

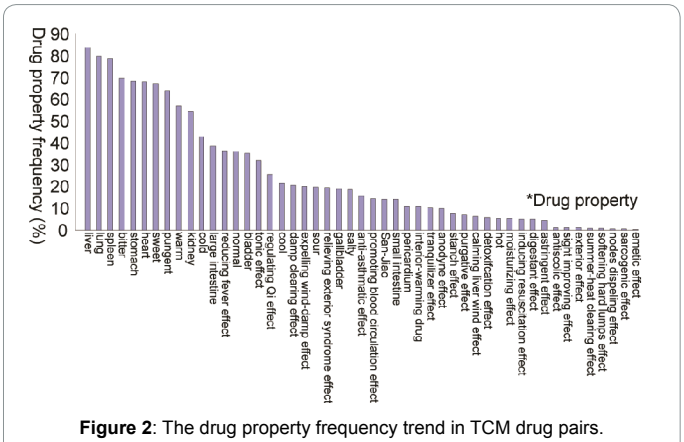

Figure 2: The drug property frequency trend in TCM drug pairs.

No.	Prop1*	Prop2	sup (%)	conf1 (%)	conf2 (%)
1	liver	spleen	66.24	79.16	83.98
2	liver	lung	65.44	78.20	82.29
3	spleen	lung	63.04	79.92	79.28
4	bitter	liver	57.60	82.76	68.83
5	pungent	liver	56.80	88.97	67.88
6	lung	stomach	56.64	71.23	82.71
7	heart	liver	56.32	82.82	67.30
8	spleen	stomach	55.68	70.59	81.31
9	liver	stomach	55.20	65.97	80.61
10	sweet	lung	54.88	81.67	69.01

Note: *- Abbreviation of Property1

Table 1: The top 10 frequent rules mined by Apriori method.

Associations between flavor properties, such as sweet, pungent and bitter, and channel tropisms were also frequent with confidences above 80%. There might be close relations between the two properties in TCM drug pairs. The highest single conf appeared in pungent - liver association with conf1 88.97% but with conf2 67.88%, which was the smallest in Top 10. In TCM drug pairs almost all the pungent drugs were coupled with drugs with liver channel tropism, but drugs with liver had many candidate partners. From the results of Apriori, some possible meaningful associations with high sup or conf might be found, but it was unknown whether these associations came from the properties of two drugs respectively or from the properties of single drug. For example Angelica sinensis (Dang-gui) had pungent and liver property. If a drug pair contained Angelica sinensis, the association pungent – liver would be satisfied no matter what drug was. Then the drug pair had false contribution to this rule. The sup and conf might be interfered by the false contribution.

The associations were mined again by the new procedure proposed in this paper. The limitation sup was set as the same as Apriori mining. Two association sets on the two property matrices of drug pairs were mined by Apriori firstly and then meaningful rules satisfying sup limitation from the Cartesian product of the two association sets were filtered. For each rule found its natural sup and adjusted sup were calculated. The Top 10 frequent rules were list in Table 2 as follows.

The contents in Table 2 were some similar with those in Table 1. The rules consisted of liver, spleen, lung and stomach channel tropisms had high sup too. The sup in Table 2 seemed smaller than that in Table 1, because the false sup contribution of associations between properties of single drug were excluded.

Since every drug had multiple properties, the associations from single drug might be mined out by Apriori method. Only by sup or conf in results we could hardly judge whether a rule generated between drugs or within properties of one drug. So Apriori reported an untruthful high sup in this study. Our new procedure could obtain target associations between drugs and was more suitable for mining on TCM drug pairs. In other way the association with two same properties, such as liver - liver, appeared in results which could not be got by Apriori method. It meant some drug pairs were consisted of drugs both in liver channel tropism. Obviously it was a meaningful association standing for a compatibility pattern in our study. This was another advantage of the new procedure in this study.

In Table 2, natural sup was the theoretical population rate of a rule and adjusted sup was the difference of sup and natural sup. As some drugs or properties were applied more frequently in TCM clinical practice than others, associations from these properties might have a high sup

No.	Prop1	Prop2	sup (%)	Natural sup (%)	Adjusted sup (%)	P_{val}^*	conf1 (%)	conf2 (%)
1	liver	lung	59.04	36.42	22.62	0	65.54	66.13
2	liver	spleen	58.72	37.57	21.15	0	68.86	67.46
3	spleen	lung	53.12	30.61	22.51	0	65.87	62.88
4	bitter	liver	51.04	28.32	22.72	0	59.74	63.04
5	liver	stomach	49.60	29.09	20.51	0	57.73	56.88
6	liver	sweet	49.12	29.29	19.83	0	58.03	56.23
7	bitter	lung	48.64	23.07	25.57	0	58.80	60.20
8	stomach	lung	48.16	23.70	24.46	0	62.06	61.93
9	bitter	spleen	47.52	23.81	23.71	0	57.67	57.34
10	liver	liver	47.36	44.70	2.66	0.084	71.50	73.09

Note: *- Significant probabilities by binominal test. Rules were judged significant when P_{val}<0.05.

Table 2: The top 10 frequent rules between two drugs sorted by descending sup.

even if the applications did not have any special relationship with each other. The parameter adjusted sup could correct this bias and show real support without the influence of different drug application frequency. The larger absolute value of adjusted sup, the closer relationship existed between the two properties. If the adjusted sup of a rule was near to 0, the rule was thought meaningless and its sup only came from normal application of the two drugs. The P_{val} from binominal test could give the confidence level for adjusted sup equal to zero. The associations sorted by descending adjusted sup were list in (Table 3 and Figure 3).

In our study the associations list in Table 3 could be thought as part of meaningful association rules in TCM drug pairs. These rules showed some special compatibility laws in drug pairs. These rules gave detailed descriptions of TCM theory for drug pairs. These rules appeared similar with those in Table 2 in some extent. The properties of bitter, liver, spleen, lung, and so on had high sup. However, the sup of rules in Table 3 was lower than that in Table 2. The highest sup in Table 3 was lower by about 11%. Compared with the rules mined by our new procedure only between two drugs of drug pairs, the abnormal high sup in Table 2 partly came from the rules between one drug's properties. These false sup could not be filtered out completely by Apriori algorithm. Our new procedure could give more real and accurate sup values.

The rule liver - liver in Table 2 has a high sup of 47.36% but low adjusted sup of 2.66%. The P_{val} was 0.084, larger than confidence level 0.05, showed no statistical significance. That meant there was no special relationship between two drugs following liver tropism in TCM drug pair. So the high sup from high drug applying frequency was meaningless for drug pair structure research. The statistic test on association rules was verified to be necessary to exclude meaningless ones after association mining.

Discussion

TCM formulae are the main application form of drugs in clinical practice. The formulae are special drug group following the instruction of compatibility laws in TCM theory. However, the theoretical descriptions of compatibility laws seemed to be understood and communicate with modern medical science with great difficulty. And there might be several to hundreds of drugs in a formula. The different drug numbers increase the complexity in formulae compatibility research. TCM drug pairs, with exact two drugs and containing compatibility laws, were quite suitable objects in the formulae compatibility research. In this paper the property associations of drug pairs were mined to demonstrate the compatibility manner.

In drug pair mining the properties in associations should come from different drugs, that is, from different datasets. So in this paper the association mining should run on multiple datasets. While most association mining methods ran on single dataset, multiple datasets must be combined into one before mining and then associations were mined by constraint mining. The combination might be done by two measures, mergence or junction. The merging of datasets was defined as the result of logical OR operation on datasets. And the junction of datasets was defined as the union with property renamed. For example, for two datasets D_1 and D_2 with the same property set P={p_1, p_2, …, p_n}, the new dataset after mergence also had property set P, while the dataset after junction would had property set of {D_1–p_1, D_1–p_2, …, D_1–p_n, D_2–p_1, D_2–p_2, …, D_2–p_n}. The properties were twice and renamed to avoid property name duplication. If D_1 and D_2 had completely different properties, the two combination measures would be the same. However, the two combination measures might cause some deficiencies in association mining. In above example if a rule p_1-p_2 had high support in D_1, the rule would also be selected with high support after datasets mergence. Then by the support and confidence it was hardly known that the rule came from D_1 and D_2 or only from D_1. The difference between Tab. 1 and Tab.2 verified this phenomenon. If p_1 appeared frequently in both D_1 and D_2, the rule p_1-p_1 between D_1 and D_2 might be meaningful. But the rule could not be recognized after mergence. By datasets junction and constraint mining the above two problems could be resolved well except large amount of computation. Most association mining method could be divided into two steps: the generation of candidate sets and objective rules selection by support, confidence or other constraint limits. Datasets junction expanded properties twice and accordingly the computation of candidate set generation was enlarged too. In above example the new dataset after junction had 2n properties and the maximum candidate set would have C_{2n}^m items if m properties were selected. For dataset mergence the maximum number was C_n^m, much less than that in dataset mergence. So the two solutions by constraint mining seemed not perfect. In this paper a new mining procedure was proposed to perform association mining on TCM drug pair data. This procedure could just obtain associations between drug pairs, which avoided the deficiency of dataset mergence and have a moderate candidate computation. The maximum candidate number was $C_n^{m1}C_n^{m2}$, whrer m1 + m2=m. In this study n=49, m=2, and m1=m2=1. Then the maximum candidate number was 1176 for datasets mergence, 4753 for junction and 2401 for the new procedure. The mining procedure was verified quite suitable for TCM drug pairs data mining with relative small computation.

The statistic test was shown to be necessary for association rules mining. There were some rules with high supports; however the high supports came from the high applying frequency of drugs but not the special relationship between drugs. The population rate test under

No.	Prop1	Prop2	sup (%)	Natural sup (%)	Adjusted sup (%)	P_{val}^*	conf1 (%)	conf2 (%)
1	bitter	lung	48.64	23.07	25.57	0	58.80	60.20
2	stomach	lung	48.16	23.70	24.46	0	62.06	61.93
3	bitter	spleen	47.52	23.81	23.71	0	57.67	57.34
4	lung	heart	45.44	21.82	23.62	0	58.44	56.24
5	spleen	heart	45.76	22.51	23.25	0	60.21	55.75
6	bitter	pungent	40.64	17.46	23.18	0	52.70	57.86
7	bitter	liver	51.04	28.32	22.72	0	59.74	63.04
8	liver	lung	59.04	36.42	22.62	0	65.54	66.13
9	spleen	lung	53.12	30.61	22.51	0	65.87	62.88
10	bitter	stomach	40.64	18.43	22.21	0	55.95	58.93

Table 3: The top 10 frequent rules sorted by descending adjusted sup.

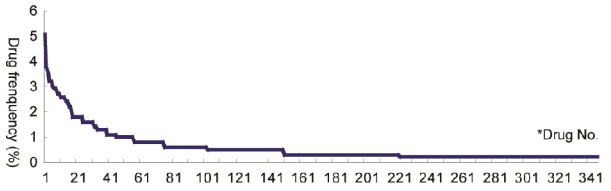

Figure 3: The top 10 frequent rules sorted by descending adjusted sup.

binominal distribution was applied in this paper, which needed an independent mining results serving as the theoretical population rate. And maybe there were some better statistic model for the mining only on the one property database. This was one of the problems to be solved in the future.

Another problem was how to analyze the association rules mined. After the new method mining and statistic test filtration, there were still many rules meaningful. The next cluster method was developing now to discovery some advanced regulations based on the mining results.

Conclusion

In this paper a new association rules mining method and a statistic test on rules were introduced and were applied in the mining of TCM drug pair information. The new method could find certain association rules over multiple databases, in which each item must come from different databases. This method was applied on the study of TCM drug pair data. The association rules with some special relationship between common drug properties were mined and tested, which stand for parts of the compatibility regulations in drug pairs and could give some help to the research on TCM formulae compatibility regulations. The method was verified to be efficient and accurate and might be applied in the formulae data mining furthermore.

Acknowledgement

Thanks are due to the support of the Major Fundamental Natural Science Project of Jiangsu collage (No. 14KJA360001).

References

1. He M, Yan X, Zhou J, Xie G (2001) Traditional Chinese medicine database and application on the web. J Chem Inf Comput Sci 41: 273-277.
2. Chen X, Zhou H, Liu YB, Wang JF, Li H, et al. (2006) Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. Br J Pharmacol 149: 1092-1103.
3. Fang Y, Huang H, Chen H, Juan H (2008) TCMGeneDIT: A database for associated traditional Chinese medicine, gene and disease information using text mining. BMC Complement Altern Med 8: 58.
4. Feng Y, Wu Z, Zhou X, Zhou Z, Fan W (2006) Knowledge discovery in traditional Chinese medicine: State of the art and perspectives. Artif Intell Med 38: 219-236.
5. Zhou X, Liu B, Wu Z, Feng Y (2007) Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks. Artif Intell Med 41: 87-104.
6. Wang Y, Ma L, Liu P (2009) Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine. Comput Methods Programs Biomed 3: 249-257.
7. Chen CT, Kung Y, Chen Y, Chou L, Chen F, et al. (2007) Use frequency of traditional Chinese medicine in Taiwan. BMC Health Services Research 7: 26.
8. Kotsiantis S, Kanellopoulos D (2006) Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering 1: 71-82.
9. Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD Conference.
10. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. Proceedings of the 20th VLDB Conference.
11. Techapichetvanich K, Datta A (2004) Visual mining of market basket association rules. ICCSA. LNCS 3046:479-488.
12. Baralis E, Psaila G (1997) Designing templates for mining association rules. Journal of Intelligent Information Systems 9: 7-32.
13. Chen J, Xi G (2009) An unsupervised partition method based on association delineated revised mutual information. BMC Bioinformatics.
14. Liu B, Hsu W, Ma Y (1999) Mining association rules with multiple minimum supports. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.
15. Brin S, Motwani R, Silverstein C (1997) Beyond market baskets: Generalizing association rules to correlations. ACM SIGMOD Record 2: 265-276.
16. Zhu A, Li J, Leong T (2003) Automated knowledge extraction for decision model construction: A data mining approach. AMIA 2003 Symposium Proceedings, pp: 758-762.
17. Cristofor L, Simovici D (2002) Generating an informative cover for association rules. Proceedings of the 2002 IEEE International Conference on Data Mining, pp: 597-600.
18. Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. ACM SIGMOD Record 2: 1-12.
19. Chan K (1995) Progress in traditional Chinese medicine. Trends Pharmacol Sci 6: 82-187.
20. http://www.chp.gov.hk/