FORMULAE MINING FOR DIABETIC NEPHROPATHY WITH TREATMENT EFFECT IN TRADITIONAL CHINESE MEDICINE

XIAOLIN ZHU, YONGGUO LIU

School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
E-MAIL: pgzxh_xc@163.com, liuyg_cn@163.com

Abstract:

Many studies have applied association rule algorithms in mining formulae in traditional Chinese medicine (TCM) so far. However, in these studies, the treatment effect of prescriptions is not considered. In the meantime, a proper minimum support threshold is needed for these association rule algorithms, which is difficult for doctors. This paper proposes TCMM (traditional Chinese medicine miner), a novel algorithm for mining the top-K weighted association rule which does not require the users to set a support threshold in advance. In addition, TCMM can analyze TCM formulae with treatment effect. After conducting several experiments on TCM formulae for diabetic nephropathy, we verify that our algorithm can facilitate doctors to conduct new drug discovery and make better clinical medicine selection.

Keywords:

Traditional Chinese medicine; Weighted association rule; Treatment effect

1. Introduction

In China, TCM has been applied to disease treatment for thousands of years. During treatment, the doctor devises Chinese medicine formulae constituted by several herbs. Association rule mining is applied to obtain prescription patterns, helping the doctor with clinical selection of medications[1], providing reference objects for future clinical experiments, and contributing to new drug discoveries[2]. For example, in[3] the authors analyzed TCM prescription patterns for menopausal syndrome and found that salvia miltiorrhiza is the single most frequently used Chinese herb prescribed at 11.09%; the most frequently used two-medicine combination is Salvia miltiorrhiza with Dan-zhi-xiao-yao-san at 5.18%. The most prevalent three-medicine combination consists of Eclipta alba, Ligustrum lucidum, and Dan-zhi-xiao-yao-san. It is found in [4] that Hedyotis diffusa plus Scutellaria barbata constitutes the core formulae. In [2], osteoporotic treatment records retrieved from Pubmed, CNKI and other sources were collected and the formulae present in the records were analyzed. The results showed that the most common medication is Rehmannia glutinosa Libosch.

However, in the above applications, the researchers did not consider the treatment-effect. Therefore, we adopt TCMM, an efficient top-K weighted association rule algorithm, to analyze TCM formulae for diabetic nephropathy with treatment effect.

2. Background and relevant work

2.1. Mining the N-most interesting itemsets

Considering association rule algorithms, the minimum support threshold ξ must be set in advance, which is sometimes difficult for the user. To solve this problem, the N-most interesting itemsets were proposed[5].

2.2. Weighted association rules

In practical applications of association rules, the importance of different transactions and items usually needs to be considered. Therefore, weights of the corresponding transaction are assigned. In [6] the author defined a type of weighted support (*WSP*).

Definition 1. WSP(P), the weighted support of itemset P, is obtained by dividing the sum of the weights of the transactions that include P by the sum of the weights of all transactions. Its formula is computed as follows:

$$\mathrm{WSP(P)} = \frac{\sum_{p \in t_j} w_j}{\sum_{t_j \in \mathrm{WTD}} w_j} \quad (1)$$

where t_j ($j \in [1, m]$) denotes the transactions in *WTD*, and w_j is the weight of t_j.

Definition 2. Frequent weighted itemsets or patterns (*FWI*): Assume the minimum weighted support threshold is

δ. An itemset P is an FWI iff WSP (P) > δ.

Definition 3. *k-FWI*: the *FWI* whose length is *k*.

2.3. Meta-analysis and its application to TCM

Meta-analysis refers to collecting, combining, and performing statistical analyses of various research results concerning the same scientific problem. Meta-analysis has increasingly been applied to research on TCM. For example, in [7], literatures concerning TCM treatment of aspirin resistance were retrieved from databases such as PubMed and EMBASE, and a meta-analysis of 18 randomized controlled trials was conducted, leading to the conclusion that TCM was an effective and safe way of replacement therapy and collaborative treatment for aspirin resistance.

3. Method

First, 132 research articles in Chinese or English on diabetic nephropathy treatment in TCM are retrieved from databases (e.g., PubMed and CNKI) using key words, such as 'TCM' and 'diabetic nephropathy'. Then, 78 of the 132 research articles that adopt double-blind randomized trials are selected. Third, weights are assigned to the prescriptions in the article in accordance with the therapeutic data. Taking the prescriptions as transactions and the drugs included in the prescriptions as items, we form a weighted transaction database of TCM prescriptions (*TCMWTD*). A global tree is built by scanning *TCMWTD* twice. Finally, the top-*K* most frequent weighted itemsets, namely, the prescription patterns with good treatment effect, are iteratively generated from the tree.

3.1. Determination of weights

Table 1 Comparison of treatment effects between the treatment group and the control group

	effective	ineffective	sample
Treated	A	B	n_1
Control	C	D	n_2

In literatures, comparisons of clinical effects between these two groups are recorded. Table 1 shows an efficacy comparison in a TCM study literature, where n_1 and n_2 refer to the sample sizes of the treatment group and the control group, respectively. In this independent study, *A* and *C* stand for the numbers of corresponding effective cases, respectively, while *B* and *D* are the numbers of corresponding ineffective cases, respectively.

In this study, we apply the weight formula for meta-analysis in the medical field [8], defined below.

$$W = \frac{1}{\frac{1}{A} - \frac{1}{n_1} + \frac{1}{C} - \frac{1}{n_2}} \quad (2)$$

3.2. Top-*K* weighted association rules mining algorithm TCMM

The doctor is only interested in the "top" rules. Moreover, it is difficult for the doctor to determine *δ*. This paper proposes TCMM, a top-*K* weighted association rules algorithm in which, instead of setting *δ*, the user inputs more easily obtainable parameter values for *K* and *N*. *K* controls the length of the longest pattern and *N* refers to the number of each pattern with different lengths.

3.2.1. Preliminaries

Definition 4. The *N*-most frequent weighted *k*-itemsets: When all *k-FWI*s are arranged in the descending order of *WSP* and the *WSP* of the *N*th *k-FWI* is WS_N, the *k-FWI*s whose *WSP* is greater than WS_N constitute the *N*-most frequent *k*-itemsets.

Definition 5. The top-*K* frequent weighted itemsets (top-*K FWI*): all the *N*-most frequent weighted *k*-itemsets, where $1 \le k \le K$.

Problem Statement: To explore the most effective patterns of TCM formulae, the top-*K FWI* should be mined from *TCMWTD*.

3.2.2. Constructing a global TCMM-tree using a dynamically changing threshold

By scanning the transaction dataset twice, TCMM establishes a global TCMM-tree. The structure of a TCMM-tree is similar to that of an FP-tree and consists of a root node (named "root"), subtrees under the root node, and a Header Table.

Each node in the subtrees is composed of five domains: *item*-name, *weightCount*, *parent-link, child-link* and *node-link*. Each node in the HeaderTable saves information about an item and includes three domains: *item-name*, *weightSupport* and *the head of node-link*.

The process to build the global TCMM-tree is as follows: initially, *δ* is set to 0. The first scan obtains the *WSP* of all the items, the root node of the tree and the HeaderTable are established. Then, during the second scan, the items in each transaction are visited in *WSP* descending order, and *δ* is updated according to the dynamic condition of the transaction.

For example, suppose that the topmost *K+1* items of highest *WSP* are arranged in *WSP* descending order and form an itemset, $S_{max} = \{t_0, t_1, \ldots, t_k\}$. Thus, *N* subsets of S_{max}

with a length of K are most likely to become the N-most frequent weighted K-itemsets and are denoted as follows: S_0:{t_0, t_1,…,t_{k-1}}, S_1:{t_0, t_1,…, t_{k-2}, t_k},…, S_{N-1}:{t_0, t_1,…, t_{k-N}, t_{k-N+2},…,t_k}

Hence, we add a count array, W, to save the WSP of S_i. The i-th entry in W is marked as $W[i]$ and corresponds to the WSP of S_i. Its initial value is set to 0. When transaction $MTrans$ with a weight of ω is scanned, before it is inserted into global tree, the number of its items included in S_{max} α is calculated. When $\alpha < K$, this transaction should be inserted into the global tree directly, when $\alpha = K$ and $MTrans$ contains S_i, $W[i]$ is incremented by ω, and when $\alpha = K + 1$, all the entries of W are incremented by ω. Then, δ is updated according to the minimal WSP in W.

After removing the items in $MTrans$ whose WSP values are less than δ, during the call to *InsertTCMM-Tree* ($MTrans$), $MTrans$ is inserted into the prefix tree. In this process, suppose $currNode = Root$ and $i = 0$. Then, when i is less than the length of $MTrans$, the following steps must be executed:

We first determine whether $currNode$ has a child node whose item-name is equal to the i-th item of $MTrans$. If such a child node exists, that child node becomes the $currNode$ and its $weightCount$ should be incremented by ω. When no such child node exists, a new node must be constructed, whose item-name is equal to the i-th item of $MTrans$, and whose $weightCount = \omega$. Then, this new node becomes $currNode$, and i is incremented (i++). TCMM-tree is established by scanning all the transactions in this way.

3.2.3. Mining TOP-*K* weighted frequent patterns

After the global TCMM-tree is built, to mine the top-K FWI, structure $wresult$ should be formed at first. In $wresult$, K lists are used to save the FWIs of different lengths from 1 to K respectively. The i-th ($1 \le i \le K$) list $wresult_i$ saves FWIs with a length of i and includes N cells, where each cell holds an FWI and its WSP. Suppose the minimum WSP is δ_i in $wresult_i$, if the WSP of a newly explored i-FWI is greater than δ_i, it is substituted for i-FWI with the minimal WSP saved in the list. At the same time, δ_i is updated, and if $i = K$, $\delta = \delta_K$.

Then, the TCMM-growth procedure is called, and the global TCMM-tree is used as the input parameter. Starting from the item with the highest WSP in the HeaderTable, the nodes in the HeaderTable are accessed sequentially downward until the WSP values of the items in the HeaderTable become to be less than δ. Suppose the node for item 'a_i' in the HeaderTable is connected with Node L, in the tree Prefix P' is formed by the prefix pattern and item 'a_i'. Then, the algorithm evaluates whether L's $node\text{-}link$ is empty. If so, there is only one path connecting 'a_i' and root, and all the nodes on this path can be accessed via their $parent\text{-}link$. The combination of all the items of these nodes with prefix P' generates new patterns with a length of j ($1 \le j \le K$). In addition, the WSP values of these new patterns are set to match the WSP of 'a_i'. If the WSP of these new patterns with a length of j is greater than δ_j, they are saved in $wresult_j$. In contrast, if L's $node\text{-}link$ is not empty, all the paths between 'a_i' and Root must be traversed via $node\text{-}link$ and $parent\text{-}link$, where the items on one path form a conditional pattern. By calling the *Construct_TCMM-Tree* procedure, new conditional TCMM-trees, TP, can be built using these conditional patterns. Then, taking TP and P' as input parameters, the $TCMM\_growth$ procedure is called recursively. Finally, we obtain all the top-K FWIs from $wresult$.

4. Experimental results

4.1. Results of TCM prescription mining

TCMM is written in C++ and used to mine the top-3 effective prescription patterns. The results are shown in Table 2. For comparison, we use the support to measure mine the top-3 most frequent patterns without weighting of TCM prescriptions, and the results are shown in Table 3.

4.2. Discussion for TCM prescription mining

According to Formula (2), as the sample sizes and treatment-effect of randomized controlled trials in the literature increase, the weights assigned to prescriptions should also increase. As shown in Table 3, based on the mining results without considering the prescription effects, *Astragalus* ranks second and *Cornus officinalis* ranks first among the 1-Patterns. However, as shown in Table 2, after prescription weighting, *Astragalus* ranks first among the 1-Patterns. This result occurs because the literature contains studies with larger sample sizes and better effects concerning prescriptions that include *Astragalus*. For instance, in [9], which focused on *Astragalus*, the randomized controlled trial had 126 patient cases, among which 63 were in the treatment group and 63 in the control group. In addition, the number of effective cases was 60 in the treatment group but only 53 in the control group. According to Formula (2), the weight of this prescription was 263.95. Although 32 articles involved prescriptions with *Cornus officinalis*, the sample sizes of their randomized controlled trials were small and had smaller effects.

Table 2 Explored effective prescription patterns after

prescription weighting

Length	prescription patterns after prescriptions weighting	WSP (%)
1-patterns	Astragalus	(53.45)
	Cornus officinalis	(51.76)
	Rehmannia glutinosa	(43.45)
2-patterns	Astragalus, Cornus officinalis	(41.12)
	Astragalus , Rehmannia glutinosa	(36.33)
	Salvia miltiorrhiza, Astragalus	(35.13)
3-patterns	Astragalus, Ophiopogon japonicus, Radix rehmanniae	(33.55)
	Atractylodes, Codonopsis pilosula, Poria cocos	(30.47)
	Astragalus, Rosa laevigata, Semen Euryales	(29.04)

Table 3 Explored frequent prescription patterns without prescription weighting

Length	prescription patterns without prescriptions weighting	Support
1-patterns	Cornus officinalis	(32.0)
	Astragalus	(31.0)
	Rehmannia glutinosa	(26.0)
2-patterns	Poria cocos, Astragalus	(19.0)
	Atractylodes, Cornus officinalis	(18.0)
	Poria cocos, Alisma	(16.0)
3-patterns	Atractylodes, Cornus officinalis, Astragalus	(12.0)
	Astragalus, Ophiopogon japonicus, Radix rehmanniae	(10.0)
	Atractylodes, Codonopsis pilosula, Poria cocos	(8.0)

According to TCM theory [10], *Astragalus* shows effects such as tonifying the spleen, diuresis, and it can be used to tonify the kidneys, strengthen body resistance, and increase yang. In clinical TCM, *Astragalus* is one of the most common Chinese herbs in TCM prescriptions. Therefore, the high ranking of Astragalus among the 1-Patterns in Table 2 is more consistent with its use in clinical medication.

The high *WSP* of the combination of *Radix rehmanniae* with *Astragalus* and *Ophiopogon japonicus* causes that combination to rank first in the 3-Patterns in Table 2. Moreover, this combination ranks the second in Table 3. This combination has effects such as supplementing qi and nourishing yin [11]. Therefore, determining effective combinations of *Radix rehmanniae* with *Astragalus*, *Ophiopogon japonicu*s, and other medications can help doctors make selections in clinical medication.

5. Conclusions

To the best of our knowledge, this research first achieves efficacy-based mining of TCM prescriptions. To mine TCM prescription patterns effectively, we propose TCMM, a novel TOP-*K* weighted association rules algorithm, in which the doctor need only choose the length and numbers of the mining patterns. During the establishment of the global tree, TCMM carries out an estimation process for δ, which greatly enhances the algorithm's efficiency. The experimental results are consistent with TCM theory and can be used as important clinical references.

References

[1] Lin, J.-F., et al., Characteristics and prescription patterns of traditional Chinese medicine in atopic dermatitis patients: Ten-year experiences at a Medical Center in Taiwan. Complementary Therapies in Medicine, 2014. 22(1): p. 141-147.

[2] Zhang, N.D., et al., Traditional Chinese medicine formulas for the treatment of osteoporosis: Implication for antiosteoporotic drug discovery. Journal of Ethnopharmacology, 2016. 189: p. 61-80.

[3] Chen, H.Y., et al., Prescription patterns of Chinese herbal products for menopausal syndrome: analysis of a nationwide prescription database. Journal of Ethnopharmacology, 2011. 137(3): p. 1261.

[4] Yeh, Y.C., et al., Hedyotis diffusa Combined with Scutellaria barbata Are the Core Treatment of Chinese Herbal Medicine Used for Breast Cancer Patients: A Population-Based Study. Evidence-Based Complementary and Alternative Medicine, 2014, 9(3): p. 202378.

[5] Cheung, Y.L. and W.C. Fu, Mining Frequent Itemsets without Support Threshold: With and without Item Constraints. IEEE Transactions on Knowledge & Data Engineering, 2004. 16(9): p. 1052-1069.

[6] Tao, F., F. Murtagh, and M. Farid. Weighted association rule mining using weighted support and significance framework. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003. ACM. p. 661-666.

[7] Chen, H., et al., Chinese Herbal Medicine for Aspirin Resistance: A Systematic Review and Meta-Analysis. Plos One, 2016. 11(5): p. e0154897.

[8] Borenstein, M., et al., Introduction to Meta-Analysis. 2009.

[9] Zheng Zhenxiong, Zheng Yuanyu , et al., The treatment of matsuba diabetes party and Zhen Wu Tang fordiabetic nephropathy patients clinical curative effect observation. China Medicine And Pharmacy, 2015. 5(15): p. 61-63.

[10] Wang Ying，Xie Peifeng , et al., Clinical Observation of Gushen Agent in the Treatment of Diabetic Kidney Disease of Qi-Yin Deficiency and Collaterals Stasis. World Chinese Medicine, 2016. 9(11): p. 1732-1735.

[11] Zlu Cong, Zhao Jinxi, Clinical effect of Supplementing Qi and Nourishing Yin, Activating Blood and Dredging Collaterals Prescription in the treatment of diabetic nephropathy. China Medical Herald, 2017. 22(14): p. 109-112.