Study on Application of Data Mining Technology in Syndrome Differentiation of TCM

Dai Zhiguo[1,] , Han Yangyang[2, a]

[1]Application technology College of Jiamusi University, Jiamusi, China.

[2]Construction Engineering College of Jiamusi University, Jiamusi, China.

[a]Corresponding Author:hanyangyangjms@163.com.

Keywords: data mining; association rule; FP-tree algorithm; TCM application

Abstract: Study on the applications of association rule mining in traditional Chinese medicine (TCM) knowledge and experience is carried out in this paper. The association rules of disease symptoms and syndrome differentiation, syndrome differentiation and prescription, disease symptoms and prescription are mined by analyzing the cases of patients with chronic gastritis, and then the mined association rules are interpreted that provide the beneficial reference for data mining technology in TCM.

Introduction

As an edge discipline, data mining emerges as the integration of artificial intelligence and database technology. It is committed to find knowledge and rules about the essence f things and development trend hidden in the data, furthermore to dig out hidden rules about things to support experts' decision-making.[1]

Data mining technology has a good application prospect in TCM. Its application in diagnosis of TCM can not only help to discriminate the complicated relationship between syndromes and symptoms with a clinical diagnostic data, summarize the rules of syndrome differentiation TCM experts exerting, simulate the diagnostic reasoning process, but also may help to find new knowledge to enrich the experiences of experts and the theory of TCM, which lay the foundation for improving the diagnostic accuracy, and promoting industrialization and internationalization.

TCM experience relies on the combination of academic and clinical thinking. Experts have accumulated large amounts of data, including signs and symptoms, clinical examination, diagnostic methods and processes of syndrome differentiation which are of multi-dimensional space-time, multiple factors, nonlinear.[2] Syndrome differentiation and treatment is essence and features of TCM that has become the dominant model in the field of TCM diagnosis and treatment. However, the traditional syndrome differentiation heavily relies on the practitioners' experiences and the diagnostic accuracy is affected by multiple factors such as the practitioner's ability and academic school. The consequent shortages of strong subjectivity and poor reproducibility have seriously hindered the promotion and development of TCM.[3] Therefore, transforming TCM from depending on the experiences to developing a quantitative precise science has become the objective requirement of TCM modernization. So, the objective research on diseases and symptoms of TCM is carried out to improve the accuracy of syndrome differentiation.

Data Preprocessing before Association Rule Mining

Symptoms Pretreatment

The symptoms pretreatment includes terminology specifications of symptoms, the logical relationship specifications among symptoms.

(1) Terminology specifications of symptoms.

This specification mainly refers to the code that is expressed differently but has same meaning. For example, similar expressions of the term "cold extremities" are as below: "physical cold and cold limbs", "physical cold and fear of cold", "cold limbs and fear of cold", "intolerance to cold and

Published by Atlantis Press, Paris, France.

cold limbs", and so on. Similar implications of the term "anorexia" are as below: "inappetence", "poor appetite", "less food", "loss of appetite" ect. "vertigo" can be expressed in "sour waist and weak limbs", "sour waist and weak knee", "lassitude in loin and legs" and other terminology associated with similar symptoms.

(2) Logical relationship specifications among symptoms

In the symptoms manifested by diseases, there are some symptoms are not standard symptoms, such as no sputum, no fever, not thirsty, no sweat, no aversion to cold, and these expressions are not counted as symptoms.

All of the symptoms were statistically classified by referring to relevant books about TCM. According to the above specifications, symptoms descriptions used for mining experiments are attained and each symptom is represented with unique figure that is shown in Table 1.

Table 1 Some Symptoms Descriptions for Mining Experiment

Code	Symptoms descriptions	Code	Symptoms descriptions
1	gastric discomfort	6	whitish tongue
2	hot drink preference	7	pulse counting
3	heartburn	8	acid regurgitation or pantothenic acid
4	emaciation	9	yellow and greasy tongue fur
5	white and thin coating of the tongue	10	pink tongue

Pretreatment for Syndrome Differentiation, Treatment and Prescription

In the mining process of clinical experiences, syndrome differentiation is very important data, thus the data pretreatment is also essential.[4] Table 2 lists a part of the data modified according to the above criteria.

Table 2 Examples of Syndrome Differentiation Data Modified

Code	Syndrome differentiation data	Code	Syndrome differentiation data
300	stomach yin deficient	304	phlegm heat resistance
301	incoordination between liver and stomach	305	stomach fails to propel downwards
302	gastric network congestion	306	dam -heat in the liver and the gall
303	spleen-qi deficiency syndrome	307	functional activity of qi being not smooth

Here Chinese medicine name were unified according to the standard of TCM database, and some of the data unified and the corresponding digital code is shown in Table 3.

Table 3 Examples of Chinese Medicine Names Unified and Corresponding Codes

Code	Medicine name	Code	Medicine name
78	angelica sinensis	83	liriope
79	white peony root	84	radix pseudostell
80	cassia twig	85	polygonatum
81	Chinese yam	86	dendrobium
82	radix adenophorea	87	bergamot

Specific Mining Steps of Association Rules Applied in Chinese Medicine Knowledge

The case data is mainly numerical after preprocessing, which is four-dimensional including the symptoms, syndrome differentiation, treatment, and prescription. The basic data format for patient cases has listed in Table 4.

Table 4 Basic Data Format for Patient Cases

Patient ID	Main symptom	Syndrome differentiation	Treatment	Prescription

According to the above data format and content, the restrictive relations between two of the four-dimensional data above are mined, mainly including:

(1) Association rule between main symptom and syndrome differentiation

$$\text{Main symptom}(X, A_i) \Rightarrow \text{Syndrome differentiation}(X, B_j)$$

(2) Association rule between main symptom and prescription

$$\text{Main symptom}(X, A_i) \Rightarrow \text{Prescription }(X, D_j)$$

(3) Association rule between syndrome differentiation and prescription

$$\text{Syndrome differentiation}(X, B_i) \Rightarrow \text{Prescription }(X, D_j)$$

This shows that this mining model is a two-dimensional association mining model. So dimensionality reduction is also required to develop a single dimension data mining model. FP-tree algorithm is used to mine the new single dimension data, after which rule screening is carried out aiming to delete the inappropriate and take appropriate rules.[5]

Example of Data Mining Course

Table 5 lists the transaction data that need to be mined.

Table 5 Transaction Database of Patients

Patient ID	Main symptom	Syndrome differentiation	Treatment	Prescription
1	1，3	10	23	4
2	2，3	12	20	5
3	1，2，3	10	23	5
4	2	13	21	5

Now assuming that association rule between main symptom and prescription is to be mined, namely main symptom(X, A_i)⇒prescription (X, D_j), then both the data of “main symptom” and “prescription” must be combined into a new single dimension data, based on which FP-tree association mining algorithm is applied. Patient data and mining record is shown in Table 6.

Table 6 Patient Data and Mining Record

Patient ID	X
1	1，3，4
2	2，3，5
3	1，2，3，5
4	2，5

The mining process has been finished, and frequent patterns are <3, 2, 5:2> and <3, 2:2>, but the 2 and 3 indicates that the basic symptoms, 5 represents the prescription. In the above two frequent pattern, only <3, 2, 5:2> meets the requirements, namely the mining association rule between basic symptoms and prescription is selected.

Therefore, a frequent set {2, 3, 5} is available, nonempty subsets are {2}, {3}, {5}, {2, 3}, {2, 5}, {3, 5} respectively. The association rules and the confidences are listed in Table 7.

Table 7 Association Rules and Confidences

Rule	Confidence
$2 \Rightarrow 3\wedge 5$	2/3=66%（{2，3，5} frequency /{2} frequency）
$3 \Rightarrow 2\wedge 5$	2/3=66%（{2，3，5} frequency /{3} frequency）
$5 \Rightarrow 2\wedge 3$	2/3=66%（{2，3，5} frequency /{5} frequency）
$2\wedge 3 \Rightarrow 5$	2/2=100%（{2，3，5} frequency /{2，3} frequency）
$2\wedge 5 \Rightarrow 3$	2/3=66% ({2，3，5} frequency /{2，5} frequency)
$3\wedge 5 \Rightarrow 2$	2/2=100% ({2，3，5} frequency /{3，5} frequency)

In the example 1, 2, 3 represent the basic symptoms, and 4, 5 represent the drug names, hence rule screening is essential. The two-dimensional data of symptoms and drug names is reserved and the corresponding rules of main symptom⇒drug name or drug name⇒main symptom are applied that are shown in Table 8.

Table 8 Association Rules Selected and Their Confidences

Rule	Confidence
$5 \Rightarrow 2\wedge 3$	2/3=66% ({2，3，5}frequency/{5} frequency)

$2 \wedge 3 \Rightarrow 5$	2/2=100% ({2，3，5} frequency /{2，3} frequency)

The "$5 \Rightarrow 2 \wedge 3$" means that the two sorts of symptoms 2 and 3 represent can be cured by 5 with 66% confidence. “$2 \wedge 3 \Rightarrow 5$”means that the symptoms 2 and 3 represent usually be treated by the drug 5 represents.

This is the basic idea of the model. According to this idea, one dimensional association rule mining of the other data can be achieved.

Summary

Association rule mining is applied to the experience mining of TCM in this paper, and a series of mining results are available after data pretreatment and FP-tree method application. Analysis of the experimental results shows that experimental results are consistent with the theory of TCM and the association rules about principle and prescription mined in this paper is valuable experiences for TCM.

References

[1] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers, 2001.

[2] Chen Ming, Zhang Shuhe. Application of Association Rules in Symptoms Diagnosis of Diseases [J]. Chinese Medicine Series, 2004, 4(5):14-16.

[3] Wang Xuewei, Qu Haibin, Wang Jie. A Quantitative Diagnostic Method Based on Data-mining Approach in TCM [J]. Journal of Beijing University of Traditional Chinese Medicine, 2005, 28(1): 177-182.

[4] Brameier M, Banzhaf W. A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining [J]. IEEE Transactions on Evolutionary Computation, 2001,5(1):17-26.

[5] Duch W, Adamczak R, Grabezewski K. A New Methodology of Extraction, Optimization and Application of Crisp and Fuzzy Logical Rules [J]. IEEE Transactions on Neural Networks,2001,12(2):277-306.

Published by Atlantis Press, Paris, France.