A Traditional Chinese Medicine Prescription Recommendation method based on Mutual Information Clustering

Yao Qin*, Zherui Ma

1School of Software Engineering, Beijing University of Technology, Beijing, 100124, China

*Corresponding author’s e-mail: 867040517@qq.com

Abstract. Traditional Chinese medicine (TCM) data is the main knowledge resource of TCM, which contains a wealth of clinical experience knowledge. Machine learning has made remarkable achievements in natural language processing. As the carrier of TCM knowledge and information stored in the form of text, using machine learning method to study these TCM data can save a lot of manpower cost, improve the objectivity of TCM, promote TCM related knowledge better, and have certain guiding significance for the research of TCM human engineering experiment. This paper proposes a recommendation algorithm based on mutual information clustering. Its core idea is calculating mutual information between two symptoms, and set symptom “relatives and friends group”, after getting the symptom clustering results of mutual information, then combine the clustering results and search algorithm to achieve the effect of recommendation and filtering. Experimental results show that the proposed method is effective.

1. Introduction

TCM has a long history and is one of the oldest forms of medicine. TCM can exist for thousands of years, which proves the value of TCM in its medical form. TCM is being accepted by the general public. More and more researchers are studying Chinese medicine, and more and more Chinese medicine is used in different countries [1].

With the development of science and technology, in order to improve efficiency, in many ways, human resources will be replaced by artificial intelligence. Instead, the development of TCM will follow the trend of the times. As an important subject in the study of TCM, the study of prescriptions has a high degree of enthusiasm in the current study of TCM. The development direction and research methods of Chinese medicine have great significance for promoting the overall development of Chinese medicine and realizing its overall trend to the world. The prescription, dosage and use method of herbal medicine have great practical significance for clinical medication. The compatibility law of herbal medicine has been studied, certified and summarized by countless predecessors, which has profound connotation. From these aspects, TCM prescription is a concentrated embodiment of medical syndrome differentiation.

At present, patients are mainly treated by means of consultation. TCM doctors rely on TCM clinic practical experience for many years, according to the process of care, method, prescription and medicine, first collect the discomfort symptoms of the patients, check the tongue and pulse, which is commonly known as "looking, hearing, asking and cutting", according to which syndrome differentiation can be obtained, then through the specific steps of evidence legislation, according to the law, select the prescription, and finally complete a prescription. This not only makes the accuracy of disease diagnosis
and prescription generation depend heavily on the diagnosis experience of doctors, but also the cost of diagnosis and prescription is very high. Based on the current situation of medical prescriptions, the use of data mining technology to remove the false, retain the true, eliminate the coarse and extract the fine historical data accumulated in thousands of years of TCM can greatly promote the development of the modernization of TCM. In this paper, a method of prescription recommendation based on mutual information clustering is proposed, it can effectively assist the medical staff to dispense medicine according to the patient’s condition and save the medical manpower to the maximum extent.

2. Related Works

Under the background of the rapid development of intelligent medicine, intelligent diagnosis and treatment of TCM has made great progress by combining traditional TCM syndrome differentiation theory with modern TCM diagnosis technology, statistical learning, machine learning and other data analysis methods and technologies [2]. The commonly used recommendation algorithms mainly include content-based recommendations [6], collaborative filtering recommendation [7], matrix decomposition [8], K-means [9]. Here are the recommendations for application in the field of TCM.

In 2016, Sheng et al. propose a new asymmetric probabilistic model for the joint analysis of symptoms, diseases, and herbs in patient records to discover and extract latent TCM knowledge [10].

In 2017, Ji et al. found the relationship between symptoms and herbal medicines through latent semantic analysis based on the hypothesis that the internal relationship between symptoms and herbal medicines recorded in medical records reflected the underlying pathogenesis [3]. After that, Zhang Ying and others of East China Normal University used the symptom and prescription information in the medical case to mine the relationship between TCM symptoms and drugs by using the implicit model, modeling the hidden pathogenesis in the field of TCM, recommending several drugs according to a certain probability, and combining to get the prescription [4].

In 2018, Li et al. of Peking University designed an end-to-end recommendation method based on the neural network machine translation model with symptoms as input and prescriptions as output by using the principle that doctors in TCM usually prescribe prescriptions according to the description of relevant disease symptoms [5]. However, the model used in this method has black box property, which can be poorly explained.

On the whole, there are few researches on TCM prescription recommendation. The main focus of the research is to mine the relationship between symptoms and prescriptions, determine drugs through symptoms, and then combine drugs to get prescriptions. However, this mode of inferring drugs from symptoms to form prescriptions is not in line with the core theory of TCM diagnosis and treatment.

3. Methods

The method of prescription recommendation based on mutual information clustering can be divided into three steps.

3.1. Mutual information computing

Information entropy is used to quantify information and describe the uncertainty of information source. The average amount of information excluding redundancy is called information entropy. The entropy of this information source is the expectation of the self-information of the random events generated by the information source. The larger the information entropy is, the more chaotic the system is and the greater the uncertainty is. Definition of information entropy $H$ is given in Equation 1:

$$H=H[-\log p_i]=\sum_i p_i \log p_i$$

(1)

The joint information entropy of two variables $X$ and $Y$ is given in Equation 2:

$$H(X,Y)=-\sum_x \sum_y P(x,y) \log P(x,y)$$

(2)
where \( x \) and \( y \) are specific values of \( X \) and \( Y \), and the corresponding \( P(x, y) \) is the joint probability of these values appearing together. If \( P(x, y)=0 \), then it is defined as 0. The same form can be extended to more than two variables.

### 3.2. Set symptom "relatives and friends group"

Here we calculate the mutual information between the two variables, thus forming a \( N \times N \) matrix, which is recorded as \( T = (\Delta \mu(i, j)) \). The element on the diagonal represents the mutual information between the variable itself and itself, which is set to 0. The general method is to select a whole threshold to determine whether it is relevant, but the method of selecting threshold is too subjective and too absolute. Therefore, we take a more reasonable method to select a specific variable \( i \), and take out the first \( Z(1 \leq Z \leq N - 1) \) variable with the largest value in the \( \text{Set}(i) = \{\Delta \mu(i, j), j = 1,2, ..., j \neq i, ..., N\} \), and form a set with \( Z \) variables, which is recorded as \( D(i) \). \( Z \) is usually very small compared with \( N \), so this set can be called the "family and friends group" of variable \( i \), because each of its variables is closely related to \( i \).

### 3.3. Clustering algorithm based on "relatives and friends group"

For each variable, we take their respective "friends and relatives group" \( D(i) \) \( (i = 1, 2, ..., N) \). If two variables are in each other's relatives and friends, then we think the two variables are strongly correlated.

The formal description is that the variables \( i \) and \( j \) are strongly correlated if and only if \( i \in D(j) \) and \( j \in D(i) \).

Only strong correlation can be gathered together, once analogy, three variables together, if and only if any two variables are strongly correlated. Because \( Z \) is finite, the algorithm must converge. The number of classes is automatically determined by the algorithm, which is a function of the number of variables \( N \) and the number of "relatives and friends" \( Z \).

Therefore, if a class is aggregated, it must meet three conditions, that is, a pattern must meet three conditions to be a class:

- The number of symptoms in the model is more than two, because most of the clinical syndromes need at least three or four diagnostic information to be diagnosed.
- Any two symptoms in the pattern must be strongly correlated.
- There is no element \( C \), which is added to the class to make the second hold, that is, the maximum number of elements in the class.

#### Table 1. Example of TCM prescription data

| Name of prescription | Add loss Chaihu Tang |
|----------------------|----------------------|
| Indications          | Postpartum by water appropriate break, feel strange disease, hand and foot traction convulsion, biting teeth faint. |
| Effect               | Two solutions to the exterior and the interior. |
| Drug composition     | bupleurum 8, scutellaria 4 and a half, ginseng 3, banxia 3, plaster 4, anemarrhena 2, astragalus 5, licorice 4 (roasted). |
| Prescription         | bupleurum, scutellaria, ginseng, pinellia, gypsum, anemarrhena, astragalus, liquorice. |
| Usage and dosage     | For each serving, add 5 pieces of ginger, 4 dates, 1.5 cup of water, fry to 1 cup, warm, no time limit. |
| Prescription sources | Under the volume of life preservation. |

### 4. Experiment

#### 4.1. The Source of Data Set

The TCM gynaecology medical data involved in this paper is divided into two parts: the first part is TCM data set, including 11115 structured TCM prescription data from TCM websites such as "TCM
family” and "TCM classic" to form TCM prescription data set, and 6984 structured TCM drug data from Chinese Materia Medica; The second part is the external data source, including the synonym forest of Harbin Institute of technology, the dictionary of Chinese medicine terms and the gynaecology of Chinese medicine. An example of TCM prescription data is shown in Table 1.

4.2. Experimental steps

4.2.1. Calculate the mutual information and determine the "relatives and friends group" of each symptom.
According to Equation 2, the mutual information value between the two symptom information is calculated for 6285 symptom information variables in the data set of TCM gynaecology prescription, as shown in table 2.

| Combination                        | Mutual information |
|------------------------------------|--------------------|
| <unable to eat, dry sickness>      | 0.021964           |
| <irregular menstruation, under the white belt> | 0.010256 |
| <headache, typhoid fever>         | 0.014642           |
| <dry sickness, nausea>            | 0.019781           |

From the first variable, according to the value of mutual information from large to small, select the first N variables and record them. For example, we select the first 5 symptoms and record them. These 5 symptoms are the "relatives and friends group" of this variable. Table 3 is the relatives and friends group of some symptoms.

| Symptom                   | Relatives and Friends                                                                 |
|---------------------------|---------------------------------------------------------------------------------------|
| unable to eat             | retch, postpartum, pregnancy, emaciated body, cold and hot                            |
| under the white belt      | irregular menstruation, poor diet, cold in shade, infertility, blood avalanche        |
| moving tire               | pregnancy, abdominal pain, menorrhagia, inability to eat, retching                    |
| dry mouth                 | diet, postpartum, whole body strong heat, pregnancy, mouth pain                       |

4.2.2. Find out the combination of symptom information with strong correlation.
If one variable appears in the "relatives and friends group" of another variable, the two variables are related, but not necessarily strongly related. According to the subsection 3.3, the two variables should appear in their own "relatives and friends group", which is the strong correlation between them. For example, the symptoms of "spitting blood" include "epistaxis" in the "relatives and friends group" and "spitting blood" in the five most relevant symptom information of "epistaxis", so "spitting blood" and "epistaxis "are strongly related. According to this principle, we can find all the two strong correlation symptom information combinations, as shown in table 4 is part of the symptom information combinations.

4.2.3. Find out the combination mode of three variables.
On the basis of strong correlation between two and two, we can easily find the combination pattern of three and three correlations. If three symptoms can be put in one pattern, then we must satisfy that two combinations are strongly correlated, that is to say, the combination pattern of three symptom information is that three two combinations are strongly correlated.

4.2.4. The convergence of the algorithm and the classes that meet the requirements are given.
After three variables are combined, four, five and more combinations can be obtained by analogy. The algorithm will converge quickly and self-organize into several classes. The above is the process of
unsupervised mutual information clustering, from which we can see that the algorithm is self-organized, without any prior knowledge, from the relationship between two variables to the relationship of multiple variables. For example, the number of strong correlation combinations is 220, 32 and 3.

Table 4. The combination of two strong correlation symptoms

| combination          | symptoms                          |
|----------------------|-----------------------------------|
| hematemesis, bleeding| hair head, hair sideburns         |
| fear, crying         | hair more, hair sideburns         |
| rheumatic beriberi, cold | frightened fetus, convulsion     |
| dry mouth, bitter taste | qi drench, drench              |
| under the white belt, blood collapse | long diarrhea, long malaria |

4.2.5. Design the algorithm of prescription recommendation.

After getting the strong correlation combination of each symptom, the intelligent prescription recommendation technology based on mutual information clustering algorithm carries out the prescription recommendation according to the symptom set input by the user. For the user symptom set, we first use each symptom of the symptom set as a condition and use the and statement of the Whoosh search engine to write these conditions as the Whoosh search statement to search in the prescription database, that is, only when the prescription in the database matches all the symptoms, it will meet the conditions and return the most relevant prescription as the recommendation; if there is no prescription in the database that meets the conditions, then further relax the search conditions. Through the mutual information clustering algorithm based on "relatives and friends group", we cluster symptoms according to mutual information, gather the most relevant symptoms together, and get the strong correlation combination of gynaecological symptoms in TCM, and further get the corresponding relationship between symptoms and symptom types according to the strong correlation combination. We extract the symptom strong correlation group from the symptom set of the user, recommend the user according to the prescription corresponding to the strong correlation group; if there is no strong correlation group in the symptom set of the user, the search requirements will be minimized, and the symptom set of the symptom set will be searched by using the or statement to write the Whoosh search statement and return the recommended prescription.

Table 5. Comparison of three recommended method experiment results

|       | F=3   | F=5   | F=10  | F=15  |
|-------|-------|-------|-------|-------|
| K-means| Cov   | 0.4311| 0.3923| 0.3815| 0.3824|
|       | Acc   | 0.4854| 0.4668| 0.4576| 0.4512|
| Whoosh | Cov   | 0.7212| 0.6988| 0.6874| 0.6893|
|       | Acc   | 0.8236| 0.8024| 0.8064| 0.8074|
| Mutual Information Clustering | Cov | 0.8134| 0.7893| 0.7766| 0.7797|
|       | Acc   | 0.9512| 0.9113| 0.9135| 0.9173|

4.3. Experimental result

We set two indicators to define the performance of the model: coverage and accuracy. The coverage rate refers to the proportion of drugs in the recommended formula that can cover the real recommended drugs. The larger the value is, the better the recommended formula is. For each input symptom, if the coverage rate is greater than a certain threshold value, we think this recommendation is effective, otherwise it is ineffective. The recommended accuracy rate is calculated.

We randomly divide TCM prescription data into training set and test set, and the ratio of the training set to the validation set is about 8: 2, and also designed three methods to compare the recommendation performance. One is to use the Whoosh framework to perform simple search prescription recommendation, that is, any input is retrieved with or connector, and the default BM25F scoring
mechanism is used for TOP-N prescription recommendation. We use this method as the benchmark; the other is based on K-means clustering algorithm and Mutual information clustering algorithm result. The experimental results are shown in Table 5. F represents the number of relatives and friends, Cov represents coverage rate, Acc represents accuracy.

It can be seen from the results in the table 5 that when the number of relatives and friends is equal to 3, the three methods all achieve the best results. The recommended coverage rate of prescription obtained by using mutual information clustering method is 81.34%, and the accuracy rate is 95.12%. Compared with the traditional K-means clustering algorithm and the simple Whoosh search algorithm, the mutual information clustering algorithm proposed in this paper is greatly improved, so it is very effective to recommend prescriptions.

5. Conclusions
In this study, we propose an algorithm of TCM prescription recommendation based on mutual information clustering, from the experimental results, we found that the symptom strong correlation combination mined by mutual information similarity indeed has its real connotation. In the future research, more attributes of the prescription, such as the efficacy of the prescription, may be considered to improve the accuracy of the recommended prescription.

References
[1] Hao, P., Jiang, F., Cheng, J., Ma, L., Zhang, Y. (2017) Traditional Chinese medicine for cardiovascular disease. Journal of the American College of Cardiology, 69(24):2952-2966.
[2] Meng, F., Wan, F. (2010) Information construction and development of traditional Chinese medicine in China. Chinese Journal of Information on Traditional Chinese Medicine, 17(11):3-6.
[3] Ji, W., Zhang, Y., Wang, X. (2017) Latent semantic diagnosis in traditional Chinese medicine. World Wide Web, 20(5):1071-208.
[4] Zhang, Y., Ji, W., Zhou, Y. (2017) Online diagnosis and treatment system of traditional Chinese medicine based on cryptic model. Journal of Computer Applications, 37(S1):303-307.
[5] Li, W., Yang, Z., Sun, X. (2018) Exploration on generating traditional Chinese medicine prescription from symptoms with an end-to-end method. https://arxiv.org/abs/1801.09030.
[6] Thiyagarajan, R., Thangavel, K., Rathipriya, R. (2013). Recommendation of web pages using weighted k-means clustering. International Journal of Computer Applications, 86(14).
[7] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2016). Bag of tricks for efficient text classification. https://arxiv.org/abs/1607.01759.
[8] Agathokleous, M., Tsapatsoulis, N. (2017). Voting advice applications: missing value estimation using matrix factorization and collaborative filtering. In:2013 Artificial Intelligence Applications and Innovations. Berlin.pp.20-29.
[9] Himel, M., Uddin, M., Hossain, M., Jang, Y. (2017). Weight based movie recommendation system using K-means algorithm. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC). Jeju. pp. 1302-1306.
[10] Sheng, W., Huang, W., Zhang, R. (2016) A conditional probabilistic model for joint analysis of symptoms, diseases, and herbs in traditional Chinese medicine patient records. In:2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Shenzhen. pp. 411-418.