Symptoms selection using random forest based on Chinese medicine diagnostic cases of stomachache

Ruixiang Wang 1,a
1Department of Management, Shandong University of Traditional Chinese Medicine, No. 4655, Daxue Road, Jinan; 250355, China
a Corresponding author: xinxiguanliyu@163.com

Abstract. Stomachache is a disease characterized by pain in the upper part of the stomach in the diagnosis of Traditional Chinese Medicine (TCM). TCM has a history of thousands of years of diagnosis and treatment for stomachache and has accumulated rich experience. There are several syndromes for each disease in TCM. When the doctor faces the patient, he will judge which syndrome the patient’s condition belongs to by observing the patient's symptoms and the theory of TCM. There are eight syndromes of stomachache based on TCM theory. Due to the different experience and understandings of TCM theory, there has been a lack of clinically diagnostic criteria for stomachache. In this paper, we collected 279 cases of TCM stomachache, and adopted features selection algorithm based on random forest, to select the most important features of the cases. After research, it is found that there are eleven most important symptoms in TCM stomachache diagnosis. The results of this study can provide a basis for clinicians to diagnose stomachache, and can provide data support for the development of diagnostic criteria of TCM stomachache.

1. Introduction
Stomachache is a symptom in modern medicine. Different diseases, such as gastritis, gastric ulcer, gastroptosis, duodenal ulcer, indigestion, can cause stomachache. However, from thousands of years ago, Chinese medicine regarded stomachache as an independent disease. According to the theory of TCM, the causes of stomachache are various. For example, TCM theory believes the digestive system of the human body functions depending on the flow of a substance called "stomach qi". If the external environment is too cold, too hot or too humid, the "stomach qi" will no longer flow smoothly, and eventually the food will stay in the stomach and the stomachache will occur. If a person is depressed for a long time, it will cause a substance called “liver qi” to damage the stomach, causing stomachache. In addition, if a person is born with a weak stomach, this will lead to insufficient digestive function, which will eventually lead to stomachache. According to the cause of stomachache and the theory of pathogenesis of Chinese medicine, doctors summed up syndromes of stomachache. The doctor prescribes according to which syndrome the patient belongs to.

In clinical practice, the key symptoms in different syndromes of stomachache are controversial. In order to reconcile this controversy, we must standardize the syndromes of stomachache. The essence of TCM syndromes standardization is the standardization of TCM diagnosis. According to scientific standards, different doctors should give a completely consistent diagnosis facing the same patient. However, in the practice of Chinese medicine, the symptoms of different syndromes are often complicated, and one symptom often appears in different syndromes. Diagnostic criteria of stomachache depends on personal experience, teacher guidance and reading of the literature. Different
doctors give different syndromes diagnosis results facing a patient because of different syndromes criteria.

Standardization of TCM syndromes has been advancing. In 1994, Chinese Health Management Department announced “diagnostic criteria for syndrome differentiation of stroke in Chinese medicine”, which divided the stroke disease into eight syndromes, and assigned the symptoms different scores according to the diagnostic specificity and sensitivity. This is the initial attempt to standardize the diagnosis of TCM syndromes [1].

The diagnostic criteria for stroke described above is entirely the result of expert discussions. With the rapid development of data mining technology, Chinese medicine researchers began to try to use the data mining method to obtain the core symptoms of different syndromes.

Viewed from the perspective of artificial intelligence, syndromes of TCM are modes. The essence of TCM diagnosis of stomachache is a process of pattern recognition. The essence of pattern recognition of stomachache syndrome is to distinguish all syndromes with the least symptoms. We hypothesize that there is a set of symptoms of stomachache, and the symptoms in the set have an absolute discrimination of syndromes on the basis of the minimum redundancy. When a doctor faces the symptoms of a patient, he must follow the rules of this symptom set. If the hypothesis is true, any doctor who faces the same patient's symptom set should have the same diagnosis, which is essential for the scientific and social recognition of Chinese medicine.

The work in paper[2] obtained 7 core symptoms from 30 clinically effective symptoms by performing Bayesian network-based data mining on 484 clinical cases of possible blood stasis. The final conclusion of the article is that the clinical diagnosis of patients with blood stasis can be judged based on the presence or absence of these seven symptoms. The work in paper[3] collected 919 patients who were diagnosed with chronic gastritis. The study divided the gastritis into 8 syndromes. The cases involved a total of 113 effective clinical symptoms. The authors invited three clinical experts to conduct a syndrome differentiation of these patients. Finally, the study uses random forest and information-increasing algorithms to extract features from all diagnosed cases. The results show that the performance of random forest is significantly better than the information-increasing algorithm. The study finally obtained 15 core symptoms from 113 symptoms and sorted the core symptoms according to their importance. The work in paper[4] proposes that if a decision table can be established, the conditional attributes of the decision table are all biological indicators, and the decision attributes are syndromes, then through the feature selection algorithm, the biological index related to the syndrome can be obtained. The study attempted to establish a link between the syndromes of TCM and all clinical indicators. The study was limited to theoretical discussions and no actual data mining was performed.

So far, there are two different methods for standardization of syndromes. One way is through expert discussion. One way is through data mining of cases. In 2017, China Association of Traditional Chinese Medicine announced the “consensus opinion on diagnosis and treatment of stomachache”. This opinion divided stomachache into eight syndromes, which are “cold invasion of the stomach”, “diet injury the stomach”, “qi stagnation”, “blood stasis”, “cold-heat complex” “deficiency of stomach yin”, ” spleen-stomach damp heat”, and "weakness of spleen-stomach ". The study reached a consensus on the core symptoms of each syndrome. However, study on the syndromes standardization of stomachache based on data mining has not been carried out in China [5].

This study collected cases of stomachache from different dynasties, different doctors, and different regions, used feature selection algorithm based on random forest for symptoms selection. The feature selection algorithm is implemented on the python platform.

2. Method
The random forest algorithm was proposed by Breiman. Random forest is an ensemble learning method for classification. Feature selection algorithm based on random forest can assign different scores to attributes according to the importance of attributes. This feature selection method has been successfully applied to data mining of high-dimensional data in various fields [6].
The symptoms selection algorithm includes the construction of random forest and feature selection.

1. Random forest construction

The construction method of random forest is based on bagging and random subspace. Bagging is the process of building a forest from all random trees formed by multiple bootstrap samples. Random subspace refers to the selection of variables in the data for splitting a node. In a random forest, the Gini importance is used as a measure of which attribute should be chosen as the current node.

The Gini index at node $\tau$ is defined as the following formula:

$$i(\tau) = 1 - \frac{p_0^2}{p_0} - \frac{p_1^2}{p_1}$$  \hspace{1cm} (1)

$p_0$ and $p_1$ are the proportion of samples in all samples when the attribute $\tau$ is false or true.

Suppose a binary tree node $\tau$ divides samples into two subsets, $\tau_L$ and $\tau_R$, corresponding $p_L = \frac{n_L}{n}$, $p_R = \frac{n_R}{n}$, then:

$$\Delta i(\tau) = i(\tau) - p_L i(\tau_L) - p_R i(\tau_R)$$ \hspace{1cm} (2)

$\Delta i(\tau)$ is the basis for judging the decision tree generation [7].

2. Feature selection based on random forest

A decision tree $T_i$ classifies data out of bagging(oob) and obtains $A_i$ correct classification results. When the feature $X_i$ in the oob is randomly perturbed, the decision tree $T_i$ classifies the oob, and obtains $B_i$ correct classification results. We perform the same operation on all trees in the random forest, and the importance of the feature $X_i$ in the random forest is defined as:

$$I(X_i) = \frac{1}{n} \sum_{i=1}^{n} |A_i - B_i|$$ \hspace{1cm} (3)

Random forest judges the importance of the attributes by sorting $I(X_i)$ [8].

3. Datasets

Cases of TCM stomachache have been recorded for thousands of years. This study collected 279 cases of famous doctors from ancient and modern times. This study excluded some cases where symptoms or syndromes were unclear.

The names of the different syndromes used in this study are slightly different from the names proposed by the government. The definition of syndromes proposed by the government was determined after a group of experts conducted three rounds of the Delphi discussions. The names in the study are based on a broad collection of ancient and modern expert experience. In the clinical diagnosis of stomachache, doctors can refer to these two different syndrome patterns at the same time. This study divides stomachache into eight syndromes, which are "deficiency of spleen Yang", "diet injury the stomach", "qi stagnation", "phlegm and blood stasis", "cold-heat complex" "deficiency of stomach yin", "liver-stomach disharmony", and "weakness of spleen-stomach".

This study used a sparse matrix to store the patient's symptoms. All of the effective symptoms were conditional attributes, with a total of 80 symptoms such as "lack of appetite", "abdominal tingling", "searing in the stomach". When a patient has this symptom, the symptom attribute is defined as 1, otherwise it is defined as 0. The decision attributes of the decision table are the syndromes.

4. Results and Discussion

On the python platform, the feature selection algorithm based on random forest is applied to the TCM stomachache decision table collected in this study.

After random forest construction and feature selection, the most important symptoms and corresponding scores are as follows:
According to the theory of TCM, “abdominal tingling” and blood stasis have obvious correspondence. Therefore, patients with symptom of “abdominal tingling” are often considered to be the syndrome of "phlegm and blood stasis".

For Chinese medicine, there are two main reasons for constipation. The first reason is the dryness of the intestine caused by insufficient body fluid inside the body. The second factor is the slowing of the gastrointestinal tract caused by the lack of gas inside the human digestive system. Patients with constipation are often classified as either "deficiency of stomach yin" or "weakness of spleen-stomach".

According to the interpretation of modern diseases, the relief of stomach pain after eating is often caused by gastritis. Chinese medicine believes that this symptom is a manifestation of weakness in the spleen and stomach, so the doctor can judge whether the patient is "qi deficiency" syndrome based on whether the patient has this symptom.

It can be seen from clinical practice these symptoms do have obvious specificity and distinguishability of syndromes. From another perspective, those with a score of 0, such as “frequent micturition” and “tastelessness in the mouth and not thirsty” are either too subjective or not obvious, and are generally not discriminative in clinical practice.

Some syndromes of stomachache can be explained by modern medicine. For example, “weakness of spleen-stomach” can be explained by the lack of gastrointestinal motility in modern medicine. At the same time, some TCM syndromes of stomachache cannot be explained by science. For example, according to meridian theory, Chinese medicine believes that the liver and stomach are closely related. A person's long-term negative emotions can lead to a decline in appetite. This phenomenon is true, but it cannot be explained by modern medicine. Therefore, Using modern scientific ideas, methods and tools to study the relationship between TCM syndromes and symptoms can help to see the physiological and pathological functions of the human body from a new perspective.

At present, the state has standardized the syndromes of some diseases of TCM. Diagnostic criteria developed by government-organized experts are authoritative. However, because the knowledge and experience of experts are limited, this inevitably leads to the abandonment of some valuable ancient experience.

This research is based on ancient and modern famous medical records, which can retain the valuable experience of TCM diagnosis in history. Of course, because this study is a text-based study, the combination of the results of this study and the results of the clinician's discussion is a perfect standardization of syndromes.

The inheritance of Chinese medicine experience is an important task. In the process of inheritance of TCM, the method of data mining can be used to find the rules and knowledge that experts do not necessarily notice. Knowledge of TCM experts includes declarative knowledge and intuitive knowledge. For intuitive knowledge, it is difficult for experts to describe it in words. According to the expert's case description, finding the core elements of different syndromes through random forest is an important method to convert expert intuition knowledge into declarative knowledge. The method used in this study has a promotional value in the study of TCM diseases.
In addition to random forest, tools for feature selection include rough sets, lasso regression, and so on. In future research, a variety of methods can be considered, and then the importance of different symptoms can be compared under the guidance of clinical experts.

5. Conclusion
TCM doctors treat stomachache according to the syndrome, and the syndrome is determined according to patient’s symptoms. For a syndrome, the importance of different symptoms is different. The currently accepted method is to find discriminatory symptoms and score the importance of symptoms according to the opinion of experts. This method is highly operational, but limited by the experience of experts. This study collected 279 cases of ancient and modern clinical diagnosis of stomachache. Through the symptoms selection based on random forest, the most important symptoms were obtained on the python platform, and the results were highly consistent with the clinical diagnosis. The feature selection algorithm of random forest has a good mathematical foundation. The mining results of this algorithm can not only directly guide clinical diagnosis, but also provide data support for experts to develop diagnostic criteria for stomachache. Symptoms selection based on data-driven can avoid the limitations of clinical expert experience, and at the same time maximize the valuable experience of ancient people on disease understanding. The method of data mining based on the characteristics of TCM syndromes should be the main method for the objectification of TCM diagnosis in the future.

References
[1] Encephalopathy Emergency Research Group of Administration of Traditional Chinese Medicine. Diagnostic criteria for syndrome differentiation of stroke. Journal of Beijing University of Traditional Chinese Medicine. 17, 3(1994).
[2] Wang Xuewei, Qu Haibin, Wang Jie. A quantitative diagnostic method based on data-mining approach in TCM. Journal of Beijing University of Traditional Chinese Medicine. 28, 1(2005).
[3] XU Wei-fei, GU Wei-jie, LIU, Guo-ping, et al. Study on Feature Selection and Syndrome Classification of Excess Syndrome in Chronic Gastritis Based on Random Forest Algorithm and Multi-label Learning. 23, 8(2016).
[4] Jian-xin Chen, Hui-hui Zhao, Wei Wang. Investigative strategy for research biological basis of traditional Chinese medicine syndrome: feature selection-based data mining methods. 8, 8(2010).
[5] Chinese Society of Traditional Chinese Medicine Spleen and Stomach Disease Branch. Consensus opinion of traditional Chinese medicine diagnosis and treatment experts for stomachache. Journal of Traditional Chinese Medicine. 58, 13(2017).
[6] LEO BREIMAN. Random Forests. Machine Learning. 45, 5(2001).
[7] Bjoern H Menze, B Michael Kelm, Ralf Masuch, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 10, 1(2009).
[8] STROBL C, BOULESTEIX A L, ZEILEIS A, et al. Bias in random forest variable importance measure: Illustrations sources and a solution. BMC Bioinformatics. 25, 8(2007).