Feature Selection of Combining Relieff and Rough Set for Syndrome Classification of Chronic Gastritis in Traditional Chinese Medicine

Jianjun Yan[1,a], Qiyue Chen[1] Guoping Liu[2,b], Xiong Lu[3], Yiqin Wang[2], Rui Guo[2]

[1]East China University of Science and Technology,Shanghai 200237,China

[2]Shanghai University of Traditional Chinese Medicine, Shanghai 201203,China

[3]Technologies and Experiment Center, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

[a]jjyan@ecust.edu.cn.[b]13564133728@163.com.

Corresponding authors: Jianjun Yan and Guoping Liu

Keywords: Relieff; Rough Set; Multi-Label Learning;Syndrome Classification;TCM

Abstract. Typically, the main form of Chinese medicine is based on interrogation of the patient's condition and artificial judgement by asking related patient symptom information; which is strongly subjective and prone to errors of judgment, leading to the wrong treatment outcome. The development in computer greatly improved the level of research in medicine, and also allowed the gradually objective and systematic improvement of experience knowledge. By analyzing TCM inquiry chronic gastritis data, the Relieff & Rough Set feature selection method was presented by combining different classification algorithms with experiments and analysis; the experimental results showed that efficient feature selection method can greatly enhance the effect of the classification; therefore, Relieff & Rough Set can be used as an efficient tool for feature selection and applied in the syndrome classification of TCM.

Introduction

Chinese medicine dialectical conclusions are made based on the current nature of the disease, which plays a decisive role in its theoretical system and medical practice. Right dialectical conclusion is very important in clinical work. Syndrome research is the core of Chinese academic research, and also an important basis for Chinese medicine towards objective and scientific improvement[1]. On the road toobjective research in the traditional Chinese medicine diagnosis, many scholars have proposed their own methods and ideas: Li Li and Zhou Qi et al[2], who combined text mining technology to analyze the treatment law of Western medicine for chronic gastritis, which assisted the specification of Chinese-Western medicine combination treatment programs; Li Xin[3] studied a variety of mining algorithms and applied them in the mining of traditional Chinese medicine diagnosis and chronic gastritis treatment; Zhang Jiying and Ma Jingang et al[4], discussed the application of data mining in TCM diagnosis, syndromes, traditional Chinese medicine and other research; Xu Lei and He Jia[5] studied the application of information entropy-based decision tree in TCM of chronic gastritis; Qin Yanbin[6] focused on chronic gastritis data collection and data preprocessing technologies; Liang Jianqing and He Jiancheng[7] conducted research on digitization systems and applied them in objective inquiry in TCM diagnosis, which led to the improvement in the level of TCM academic research.

In the data analysisprocess of interrogation, the symptom data got from acquisition system has high dimensionality and sparsity; there are a lot of records of accompanied syndromes in these calibration data of syndromes, and also a large number of redundant symptoms. So for complex interrogation data, it is necessary to reduce the dimension of the original feature set by feature selectionmethod to remove irrelevant features and redundant features. In this article, the ReliefF & Rough Set method, with high efficiency ReliefF algorithms, is used with no limit on the data type; however, the limitation of this method is that no correlation between the features is considered, and therefore it can not effectively remove redundant features. Therefore, this paper uses rough

featuresetselection algorithm to eliminate data redundant features. As a result, the ReliefF & Rough Set feature selection method is used based on traditional Chinese medicine and chronic gastritis inquiry data, which will be applied ML-KNN and BSVM classification algorithm to analyze the performance of feature selection.

Introduction of experiment method

By combining Relieff and Rough Set two feature selection methods, better results of feature selection can be achieved. As one of an experimental method, when only the relieff method is used for feature selection, the first 20 features are used as a new subset of features in this article. For the ReliefF & Rough Set method, Relieff is first used to sort all features according to weights; the first 40 features are selected to form a new subset of features, and then the Rough Set was used to further screen these 40 features to remove the redundant part and get the new feature subset. By applying ML-KNN and BSVM classification algorithm, experiments and analysis of their performance are then performed.

Relieff feature selection method.

Relief algorithm is a classic filtering feature selection algorithm to solve the dichotomy problem. On this basis, Kononenko proposed the relieff method, which can solve many types of problems and regression problems. The algorithm can assess characteristics based on distinguishing ability of samples on close samples, namely the relevant features should make the same kind samples closer to each other, while further distinguish heterogeneous samples.

Relieff algorithm is shown as following, among which $class(R_s)$ representing the labels of sample R_s, $diff(X, R_i, H_j)$ representing the distance of feature X between sample R_i and H_j, $p(C)$ representing the possibility of C aim, $M_{j(c)}$ representing sample j of C aim, while j=1, 2, k. m and k are set according to the sample size.

Relieffalgorithm:
Input: training data set: D, the number of iterations: m, the nearest neighbor sample number: G
Output: the predicted characteristic weight vector: W
1) Initialization feature weight vector $W(A) = 0.0, A = 1, 2, \cdots, p$
2) From i=1:m
3) A randomly selected sample from D is recorded as R_i
4) To find the k nearest neighbor H_j with sample R_i
5）For each $C \neq class(Rs)$, to find the k nearest neighbor H_j with different kind of sample R_i
6）for A=1： p
Update each feature weight

$$W[X] = W[X] - \sum_{j=1}^{k}\left[\frac{diff(X, R_i, H_j)}{mk}\right] + \sum_{c \neq class(Rs)} \frac{\frac{p(C)}{1 - p(class(Rs))}\sum_{j=1}^{k} diff(X, R_s, M_{j(c)})}{mk}$$

7）end
8）end
After obtaining feature weight, the greater the weight of the sample indicates the stronger the distinguishing ability of the feature; in this way, thresholds can select a subset of the new features to achieve the purpose of dimension reduction.

Rough set.

In 1982, the Polish scholar Pawlak first came up with the idea of rough set theory [79] (Rough Set) based on G.Frege's thought of the border area; the Rough Set has a great advantage in dealing

with ambiguity and uncertainty problem, which is able to effectively analyze and process data, while it dose not depend on any prior knowledge to deal with problems in the data. Compared with the fuzzy set, it is more objective and has been widely used in many areas.

The algorithm is described as follows:

Input: the compromise decision information systems $DS\langle U, A\bigcup d, V, \rho\rangle$;

Output: attribute reduction sets of the decision information system;

Initialization $\mathrm{Re}\, d \leftarrow \phi, A \leftarrow A - \mathrm{Re}\, d$;

Step 1 Calculate heuristic function value ξ for each A condition attributes and sort;

Step 2 Select the smallest property of ξ, if multiple properties have equal ξ values, then select property value of the minimum properties;

Step 3 $\mathrm{Re}\, d \leftarrow \mathrm{Re}\, d \bigcup \{a_i\}, A \leftarrow A - \mathrm{Re}\, d, U \leftarrow U - POS_{\mathrm{Re}d}\{d\}$;

Step 4 If the domain U is an empty set, go to step 5, otherwise skip to step 1;

Step 5 output reduction attribute set: Red.

Experimental data

Chronic gastritis (Chronic Gastritis, CG) data samples are provided by four diagnostic information comprehensive research labs from Shanghai University of Traditional Chinese Medicine. They were collected from endoscopy rooms, wards and outpatient clinical cases during September 2008 - October 2010, from Shanghai University of Traditional Chinese Longhua Hospital, Shuguang Hospital, Yueyang Hospital and Eighth People's Hospital of Shanghai. The 8 chronic gastritis syndromes of TCM included the spleen, dampness in the resistance, spleen qi deficiency, spleen and stomach, liver qi stagnation, liver and stomach swelter, stomach Yin deficiency, and blood stasis stomach. However, due to the low frequency of stomach Yin deficiency and blood stasis stomach in the collected data set, only the left 6 syndromes are analyzed in this article. The 112 symptoms are collected for inquiry diagnosis. Ultimately 919 cases of CG are obtained in this research.

Experimental results and analysis

ReliefF & Rough Set was applied to select symptoms for six syndromes. 20symptoms was selected for Damp-heat accumulating in the spleen-stomach; 21 symptoms was selected for Dampness obstructing the spleen-stomach; 22 symptoms was selected for Spleen-stomach qi deficiency; 15 symptoms was selected for Spleen-stomach deficiency cold; 17symptoms was selected for Liver stagnation; 21 symptoms was selected for Stagnated heat in liver-stomach.

In order to verify the application effect of ReliefF & Rough Set, we focused on chronic gastritis dataset, which will be applied to ML-KNN and BSVM classification experiment; One of which does not applies any feature selection experiments, while another experiment applies ReliefF & Rough Set. We used ttest method to validate experiment data of ReliefF & Rough Set feature selection. After testing, "●" was added afterthe indicator, indicating that this index is better than the experimental data without feature selection;"○" was added afterthe indicator to show that index is worse than the experimental data without feature selection; and no label indicates that there is no significant difference in the two indicators. The results are shown in the table below:

Table 1 Results ofML-KNN and feature selection

	Average Precision↑	Coverage↓	Hamming Loss↓	One-Error↓	Ranking Loss↓
All Features	0.757±0.037	0.156±0.029	0.139±0.028	0.339±0.021	0.131±0.023
Relieff &Rough Set	0.772±0.021●	0.155±0.031	0.141±0.018	0.326±0.017●	0.126±0.021●

As can be seen, for the ML-KNN multi-label classification algorithm, Relieff& Rough Set feature selection showed significantly better results in three indicators, Average Precision, One-Error, and Ranking Loss, while for Coverage and Hamming Loss indicators, both methods showed no significant differences.

Table 2 ResultsofBSVM and feature selection

	Average Precision↑	Coverage↓	Hamming Loss↓	One-Error↓	Ranking Loss↓
All Features	0.799±0.026	0.164±0.034	0.148±0.027	0.331±0.022	0.132±0.043
Relieff &Rough Set	0.819±0.028●	0.153±0.027●	0.136±0.022●	0.342±0.012○	0.129±0.033

It can be seen that for BSVM multi-label classification algorithm, ReliefF & Rough Set feature selection results are significantly better than the results without feature selection in three indicators, Average Precision, Coverage, and Hamming Loss; for One-Error, it showed a distinct disadvantage; for Ranking Loss, the two methods did not show significant differences.

As can be seen from the above test results, for the ML-KNN and BSVM labeling algorithms, ReliefF & Rough Set has achieved good results in a number of indicators; therefore it is reasonable to believe that, ReliefF & Rough Set has a good adaptability in TCM chronic gastritis data.

Conclusions

Through experimental analysis, it can be seen that ReliefF & Rough Set has a good performance in the application of TCM inquiry data. For the outstanding feature selection method, it can help to speed up thedevelopment speed and improve their level in TCM objective inquiry. The proposed ReliefF & Rough Set method in this article is effective for chronic gastritis TCM data feature selection, which will facilitate the follow-up of further work and make contribution in TCM objective inquiry.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 81270050, 30901897, 81173199, 81302913, 81102729and 30701072.

References

[1]YuanshengWu,HuayuZhu,RuiqiangFan.The Study of Syndrome and Syndrome is the Only Way for the Modernization of Chinese Medicine.Journal of Traditional Chinese Medicine,2003, 21(8):1296-1297

[2]Li Li, QiZhou.Analyze the Treatment of Chinese Medicine and Western Medicine on Chronic Gastritis Based on Text Mining Technology. Chinese Journal of Experimental Traditional Medical Formulae,2011, 17(24): 228-231.

[3]XinLi.Research on data mining of TCM diagnosis and treatment of chronic gastritis. Nanjing University of Science and Technology.2007.

[4]JiyingZhang,JingangMa,HuiCao. Application of Data Mining Technology in Research of TCM. Journal of ShandongUniversity of TCM, 2014, 38(1): 83-85.

[5]LeiXu,JiaHe. Application of Decision Tree Based on Information Entropy in Differentiation of Chronic Gastritis in TCM. Chinese Health Statistics, 2004, 21(6): 329-331.

[6]YanbinQin. The Data Acquisition and Data Preprocessing Technology of Chronic Gastritis in TCM.Journal of Liaoning University of TCM, 2013, 15(10): 148-149.

[7]JianqingLiang,JianchengHe.Objective Research of TCM diagnosis system based on digital interrogation.China Journal of Traditional Chinese Medicine and Pharmacy,2014, 29(5): 1534-1538.