Comparsion analysis of data mining models applied to clinical research in Traditional Chinese Medicine

Yufeng Zhao, Qi Xie, Liyun He, Baoyan Liu, Kun Li, Xiang Zhang, Wenjing Bai, Lin Luo, Xianghong Jing, Ruili Huo

Abstract

OBJECTIVE: To help researchers selecting appropriate data mining models to provide better evidence for the clinical practice of Traditional Chinese Medicine (TCM) diagnosis and therapy.

METHODS: Clinical issues based on data mining models were comprehensively summarized from four significant elements of the clinical studies: symptoms, symptom patterns, herbs, and efficacy. Existing problems were further generalized to determine the relevant factors of the performance of data mining models, e.g. data type, samples, parameters, variable labels. Combining these relevant factors, the TCM clinical data features were compared with regards to statistical characters and informatics properties. Data models were compared simultaneously from the view of applied conditions and suitable scopes.

RESULTS: The main application problems were the inconsistent data type and the small samples for the used data mining models, which caused the inappropriate results, even the mistake results. These features, i.e. advantages, disadvantages, satisfied data types, tasks of data mining, and the TCM issues, were summarized and compared.

CONCLUSION: By aiming at the special features of different data mining models, the clinical doctors could select the suitable data mining models to resolve the TCM problem.

© 2014 JTCM. All rights reserved.

Key words: Medicine, Chinese traditional, Biomedical research; Data mining; Model; Comparison analysis
INTRODUCTION

The distinctive and salient clinical efficacy in the diagnosis and treatment of the disease was the characteristics of Traditional Chinese Medicine (TCM), which was the research of human physiology, pathology, the disease diagnosis and prevention. As a theoretical system of medicine, TCM was also gradually formed and developed in the long-term clinical treatment practice. Simultaneously, a large number of clinical data had been accumulated in the thousands of years and the theory knowledge of TCM was also contained in these clinical data. Recently, many information methods, e.g., computer technology, machine learning technology, information technology, knowledge discovery technology and data mining technology, were mature to be widely used in the field of TCM in order to find the latent knowledge of TCM. Therefore, the rule of diagnosis and therapy based on data mining technique had become a hot topic of TCM clinical research. However, it also is a challenge for many clinical researchers to precisely use the numerous data mining models. In this paper, the TCM clinical science issues are firstly summarized from four significant elements of TCM clinical researches, i.e. symptoms, syndromes, herbs, and efficacy. Then, we represent the common data mining models and their application problems for resolving these TCM clinical issues. The common data mining models include the hybrid optimization feature selection, regression analysis, Bayesian, support vector machine, cluster, latent structure model, associated rule, decision tree, free-scale network, and topic model. That is, the application analysis of the common data mining models is investigated on the base of the feature of TCM clinical data and the characterization of data mining models. Finally, the application characterizations of these common data mining models are compared and concluded into some aspects, i.e. the advantages, the disadvantages, the satisfied data types, the tasks of data mining and the TCM issues. By being clear about the TCM clinical issues and the feature of data mining models, the precise usage of data mining model in the TCM clinical researches can be improved so as to obtain the better data analysis results and provide the reliable model reference for the clinical researchers.

SUMMARY OF TCM ISSUES BASED ON DATA MINING MODELS

By searching the Pubmed and National Knowledge Infrastructure Database, we attempted to find the trend of data mining models applied in TCM clinical researches. The search criteria were given by the following keywords: (a) Data mining; (b) Traditional Chinese Medicine; (c) clinical research. Then, we selected each relevant research. We found 745 TCM clinical studies based on data mining methods out of 1668 total studies. Data mining technology was extensively applied to resolve many recent TCM clinical issues. To maximize efficacy, these TCM clinical issues were focused on exploiting the characteristics and the relationships of symptoms, symptom patterns, and herbs. These issues can be summarized as: characteristics of symptoms; the relationship between symptoms and symptom patterns; the relationship between symptoms and herbs; the relationships between symptoms, symptom patterns, and herbs; and efficacy evaluation. Figure 1 illustrates the TCM clinical issues in detail.

Symptom elements

TCM symptom information of each patient is obtained with four diagnostic methods: inspection, auscultation and olfaction, pulse-taking and palpation, and inquiry. The basis of diagnosis in TCM is symptom information. As a matter of fact, the main foundation of TCM clinical diagnosis and treatment was the symptom information, which was the basis of data mining task. 135 researches were relevant to the symptom element in the total 1668 TCM clinical researches based on the data mining models. The analysis of symptom elements is important to TCM clinical research and includes the classification of main symptoms and secondary symptoms; the combination of TCM symptoms and Western Medicine symptoms; and the de-correlation and removal of redundant symptoms. In recent years, the relationships of diseases have been explored to capture key symptoms of diseases with some methods, e.g., simulated annealing algorithm, genetic algorithm, and greedy algorithm. Furthermore, the latent information on the clinical experience is extracted from the symptom information to propagate the experience of famous TCM doctors, which is an effective way to promote TCM infomatics. The combination of TCM symptoms and Western Medicine symptoms is a key problem of TCM disease analysis. Information combination methods, e.g., weighted average, least squares, multi-Bayesian estimation, and Dempster-Shafer evidential reasoning, is helpful to represent multi-source and multi-dimensional disease information. Feature combination is a feasible way to integrate and complement disease information. Additionally, some probabilistic models need to be satisfied with assumptions of independence and identical distributions for variables (i.e. symptoms). However, symptom variables are strongly correlated with each other from the view of clinical practice. For example, some symptoms, like fever, cough, and rhinorrhea, usually co-occur for patients with colds. Therefore, methods like AdaBoost, principal component analysis, and factor analysis are necessary to de-correlate data and remove redundant symptoms, obtaining optimal results with minimal samples. These methods can accurately convey original disease characteristics.
TCM pattern elements

Treatment based on symptom pattern differentiation is central to TCM practice. "Pattern differentiation" highlights a certain pattern of symptoms by discerning the etiology, nature, and location of a disease to recognize the relationship between the pathogenic factors and the anti-pathogenic factors. Therefore, another TCM clinical issue is the analysis of pattern elements. Research into patterns includes three aspects: pattern classification, relevant elements of patterns, and risk elements of patterns. With data mining models like cluster analysis, factor analysis, principal component analysis, and latent structural analysis, the identical group features of symptoms can be achieved by mining the inherent properties of the patients. Furthermore, pattern classification, which is validated by clinical practice, is obtained by discerning the common nature of the identical group of symptoms. Conversely, with data mining models like cluster analysis, factor analysis, principal component analysis, and latent structural analysis, the identical group features of symptoms can be achieved by mining the inherent properties of the patients. The relevant elements of patterns are captured by selecting the symptoms of patients to explore the relevant relationships between the symptoms and the pattern. The relevant elements of patterns are mined to provide a feasible method for objective TCM. On this basis, the risk factors of patterns are investigated by statistical tests. These tests have a certain practical significance for clinical TCM and provide an objective basis for TCM diagnosis and treatment.

Herb elements

TCM pattern differentiation provides the basis of the prescription of herbs for treatment. Therefore, analysis of herb elements is another issue in TCM. Studies on herbs involve three aspects: compatibility of Chinese herbal medicines, core structure and modification of Chinese herbal medicines, and indications of Chinese herbal medicines. Currently, with data mining models like association rule, scale-free network, and clustering analysis, herbal combination principles satisfied with the TCM knowledge can improve the development of new Chinese herbal medicines. Furthermore, the core structure of old Chinese herbal medicines is exploited to discover modification principles, which provide valuable experience to promote the efficacy of clinical TCM. With data mining models like text mining, association rules, Fisher discriminant, and regression analysis, the indications of Chinese herbal medicines can be studied to develop usage principles.

Efficacy elements

Currently, few data mining methods are used to investigate efficacy evaluations of TCM. The process of treatment based on symptom pattern differentiation is a complex process in which the pattern is closely linked with the therapy and efficacy. During the course of in...
tervention, there are three altered states: the state of the patient, the herb prescription at the end of treatment based on pattern differentiation, and the efficacy of the different pattern with the same prescription. Therefore, efficacy evaluation in TCM is an optimal selection problem with complex and uncertain conditions, which is particularly suitable to data mining technology. By applying data mining models, research on efficacy elements are focused on efficacy analysis, optimization plans, and evaluation indicators. The relationship between evaluation indicators and the intervention is explored to evaluate the Chinese herbal medicine with the random walk model and hidden Markov. Furthermore, the decision-making model using artificial intelligence can optimize a treatment plan and evaluate its corresponding efficacy. Optimal treatment plans are useful for TCM clinical practice to summarize effective herbal prescriptions with evidence. In addition, for efficacy evaluations in clinical TCM it is important to find the evaluation indicators, which conform to TCM clinical practice. The maximization efficacy and the evaluation indicators can be obtained with Adaboost and feature selection.

**DATA MINING MODEL PROBLEMS IN TCM CLINICAL STUDIES**

Generally speaking, the results of TCM clinical research based on data mining are not satisfactory. This is because clinical TCM studies often do not conform to data mining. For example, the patterns and the herbs usually correspond to the classification labels of data mining. However, the pattern classification and core structure of herbs do not appropriately transform into multi-class or multi-type classifications. Moreover, studies may not be acceptable because, after finding accurate correspondence between a TCM clinical problem and data mining, the data mining model may be selected incorrectly. For example, a feature of TCM clinical data may not be identical to the requirements of the data mining model. In addition, TCM clinical data may not satisfy the premise of the data mining models; for example, the premise of independent and identical distributions is necessary for many data mining models to be applied in TCM clinical analysis. Thus, the correlation of symptoms is removed by feature selection and principle component analysis. Additionally, the parameters of a data mining model are important to obtain better results of TCM clinical data analysis. Based on the features of TCM clinical data, accurate understanding is necessary to guarantee a capture of the available results of clinical analysis. With appropriate parameters, estimation methods can be correctly selected to construct an optimal data mining model.

Table 1 lists the characteristics and appropriate conditions of common data mining models. The common data mining models include the hybrid optimization feature selection, regression analysis, Bayesian, support vector machine (SVM), cluster, latent structure model, associated rule, decision tree, free-scale network, and topic model. The independent variable is defined as the reason variables, which changes (for example, sex, age, and symptom). The dependent variable is defined as the change a variable causes to the independent variable, e.g., herb and pattern.

**APPLICATIONS OF DATA MINING MODELS FOR TCM ISSUES**

Integrating the aforementioned scientific issues of TCM clinical studies, we investigated the application of various data mining models to enhance the precision of model usage. First, the feature and property of TCM clinical data must be summarized to correctly use the data mining model. Second, the characterization of the data mining model must be represented to provide objective evidence for correctly selecting the model. Finally, an example model selection must be proposed to resolve the problem of pattern classification. Common data mining models are: hybrid optimization feature selection, cluster analysis, factor analysis, principal component analysis, Bayesian analysis, support vector machines, association rules, decision trees, latent structural models, scale-free networks, and text mining.

**Technology analysis of data mining models**

The relationship model of symptom, pattern, and herb is significant for the TCM clinical research and is identical to the data mining method based on the relationship between the variables and the labels. However, a patient described by the symptoms is usually diagnosed by several patterns and is treated by some herbs. Therefore, the relationship model of symptom, pattern, and herb is more suitable to be taken as a multi-label classification. That is, the diagnosis and the herbs being equivalent to the classification label are best to characterize the relationship model of symptom, pattern, and herb. The difference between the multi-class classification and the multi-label classification is whether the classes are mutually exclusive or mutually compatible. For example, a patient described by the symptoms is diagnosed by one pattern or several patterns. The characteristics of TCM clinical data are qualitative and semi-quantitative, which are discrete or mixed. Therefore, data mining methods based on discrete data or distance measurement are more suitable to creating a relationship model of symptoms, patterns, and herbs. In addition, the premise of independent and identical distributions is necessary for many data mining models to be applied in TCM clinical analysis. Thus, the correlation of symptoms is removed by feature selection and principle component analysis. Additionally, the parameters of a data mining model are important to obtain better results of TCM clinical data analysis. Based on the features of TCM clinical data, accurate understanding is necessary to guarantee a capture of the available results of clinical analysis. With appropriate parameters, estimation methods can be correctly selected to construct an optimal data mining model.
### Table 1: Generalizations of common data mining models

<table>
<thead>
<tr>
<th>Model</th>
<th>Advantage</th>
<th>Disadvantage</th>
<th>Data type</th>
<th>Task of mining</th>
<th>TCM problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid optimization feature selection</td>
<td>Integration of global optimum and local optimum</td>
<td>Complex calculation</td>
<td>Continuous discrete</td>
<td>Characterization</td>
<td>Classification of main symptoms and secondary symptom, de-correlation and removing redundancy of symptoms</td>
</tr>
<tr>
<td>Regression analysis</td>
<td>Simple, effectiveness for multivariate analysis, sole result</td>
<td>Definite relation of independent variables and dependent variable</td>
<td>Continuous discrete</td>
<td>Classification</td>
<td>Relevant element of pattern, risk element of TCM pattern, indications of Chinese herbs</td>
</tr>
<tr>
<td>Bayesian</td>
<td>Robustness for missing data</td>
<td>Fuzzy for priori probability assignment, independence and identical distribution of independent variables</td>
<td>Continuous discrete</td>
<td>-</td>
<td>Classification of main symptoms and secondary symptom, classification of TCM pattern</td>
</tr>
<tr>
<td>SVM</td>
<td>Effectiveness for nonlinear and high-dimensional data, effectiveness for small sample, fewer parameter, global optimum</td>
<td>Low efficiency of large sample, difficulty of multi-class classification, high complexity</td>
<td>Continuous discrete</td>
<td>-</td>
<td>Classification of main symptoms and secondary symptom, classification of TCM pattern</td>
</tr>
<tr>
<td>Cluster</td>
<td>Without dependent variable, efficiency</td>
<td>Be sensitive to the distance and the initial value</td>
<td>Continuous discrete</td>
<td>-</td>
<td>Classification of main symptoms and secondary symptom, classification of TCM pattern</td>
</tr>
<tr>
<td>Latent structure model</td>
<td>Without dependent variable</td>
<td>Low efficiency, exclusive relation of independent variables</td>
<td>Discrete</td>
<td>-</td>
<td>Classification of TCM patterns</td>
</tr>
<tr>
<td>Associated rule</td>
<td>Application of indirect data mining, effectiveness for various length of variables, definition of calculated consumption</td>
<td>Redundant rules, overlook of the scarce data</td>
<td>Nominal</td>
<td>Association</td>
<td>Relevant element of pattern, risk element of TCM pattern, combined principle of Chinese herbs, core herbs of prescription, addition and subtraction of Chinese herbs</td>
</tr>
<tr>
<td>Decision tree</td>
<td>High efficiency, better understandability of generated rule</td>
<td>Be sensitive to order of variable and missing data, more error for multiple classes</td>
<td>Continuous discrete</td>
<td>-</td>
<td>De-correlation and removing redundancy of symptoms, classification of TCM pattern, relevant element of pattern, risk element of TCM pattern, combined principle of Chinese herbs, core herbs of prescription, addition and subtraction of Chinese herbs</td>
</tr>
<tr>
<td>Free-scale network</td>
<td>Without dependent variable, effectiveness for sparse data</td>
<td>High-consumption of calculation</td>
<td>Discrete</td>
<td>-</td>
<td>Relevant element of pattern, risk element of TCM pattern, combined principle of Chinese herbs, core herbs of prescription, addition and subtraction of Chinese herbs, indications of Chinese herbs</td>
</tr>
<tr>
<td>Topic model</td>
<td>Without dependent variable, many-to-many relationship, Effectiveness for sparse data</td>
<td>Difficulty of parameters choose</td>
<td>Discrete</td>
<td>-</td>
<td>Relevant element of pattern, risk element of TCM pattern, combined principle of Chinese herbs, core herbs of prescription, addition and subtraction of Chinese herbs, indications of Chinese herbs</td>
</tr>
</tbody>
</table>

Notes: SVM: support vector machine; TCM: Traditional Chinese Medicine.
CONCLUSION

The benefits of data mining models were reviewed for TCM clinical researchers so that analysis methods are more available. These models could be helpful to promote the efficacy of TCM diagnosis and treatment.

REFERENCES

Zhao YF et al. / Methodology


68 Yan DR. Four diagnostic information of the acute stage of stroke and neurological function evaluation of dynamic observation. Beijing: Beijing University of Chinese Medicine, 2004: 35-41.
