journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine

Yan Wang[a,*], *Lizhuang Ma*[a,b], *Ping Liu*[c]

[a] *Department of Computer Science & Engineering, Shanghai Jiao Tong University, Shanghai, China*
[b] *Center of Traditional Chinese Medicine Information Science and Technology, Shanghai University of Traditional Chinese Medicine, Shanghai, China*
[c] *Institute of Liver Diseases, Shanghai University of Traditional Chinese Medicine, Shanghai, China*

ARTICLE INFO

Article history:
Received 2 September 2008
Received in revised form
31 January 2009
Accepted 13 March 2009

Keywords:
Traditional Chinese medicine
Syndrome prediction
Feature selection
TCMSP

ABSTRACT

Traditional Chinese medicine (TCM) treatment is one of the safe and effective methods for liver cirrhosis. In the process of its treatment, a very important step, syndrome prediction is generally performed by physicians at present, which actually hinders the application prospects of TCM. Based on the data mining algorithm, a novel method called TCMSP (traditional Chinese medicine syndrome prediction) is proposed, which consists of two phases. In the first phase, based on an improved information gain method in multi-view, the critical features are filtered from the original features. In the second phase, the class label of a new case is predicted automatically based on accuracy-weighted majority voting. The proposed method is evaluated by the liver cirrhosis dataset, 20 critical features are selected from original 105 features and the corresponding syndromes of 138 new cases are identified respectively. The critical features are in sound agreement with those used by the physicians in making their clinical decisions. Finally, this new method is also demonstrated on three standard datasets (SPECT Heart, Lung Cancer and Iris) and the results are compared with some other methods. The experimental results show that TCMSP method performs well in the field of TCM diagnosis.

1. Introduction

1.1. *The concept of TCM*

Traditional Chinese medicine, the essential of the experience of the Chinese laboring people in struggling against disease for thousands of years, is invaluable for its rich practical knowledge and a unique integrated theoretical system established since ancient times [1]. It is one of the most important complementary and alternative medicines used increasingly in the world [2]. Syndrome enables the doctor to determine the stage that the disease developed and the location of the disease [3]. Syndrome differentiation is the method of recognizing and diagnosing diseases or body imbalances by analyzing patient information based on TCM theories and the doctor's experiences [4].

TCM has been clinically observed to have dramatic performance in treating many chronic and systematic diseases such as the treatment of liver cirrhosis [5–7]. However, the lack of objective diagnosis standards hinders its wide acceptance. One cannot apply this prescriptive methodology in a professional standard until or unless one has mastered the syndrome differentiation process.

In an attempt to achieve effective and objective standard of syndrome prediction, many researchers have used data mining approach to construct the classifier for TCM dataset [8].

* *Corresponding author*. Tel.: +86 2134204586.
E-mail address: wangyan8383@sjtu.edu.cn (Y. Wang).
0169-2607/$ – see front matter.
doi:10.1016/j.cmpb.2009.03.004

1.2. *The state of the art of data mining techniques in TCM diagonosis*

Among various data mining techniques, classification analysis is widely adopted for supporting medical diagnostic decisions. Medical diagnosis is considered as a classification problem: a record represents a given patient's case, predictor features are all patients' data and the class label is the diagnosis. Subsequently, the built classification model is essential and used to predict appropriate classes for novel and uncategorized cases [9].

Medical data, such as liver cirrhosis data, often contain irrelevant features and noise. Feature selection is frequently adopted to identify and remove the irrelevant and redundant information as much as possible. Fewer features means less data should to be collected, as we know; collecting data is never an easy job in medical applications because of time-consuming and costly work. The selection of appropriate subset of the available features can yield a compact and easily interpretable representation of the target concept, model the target task adequately, and improve the classification accuracy especially in medical region [10].

In recent years, many research efforts have been devoted to syndrome prediction in TCM [11]. Wang et al. [12] constructed a self-learning expert system for TCM diagnosis. In their system, a novel hybrid learning algorithm GBPS* based on BayesNet was proposed to discover the dependence and independence relationships among symptoms and essential symptoms. Qu et al. [13] used Decision tree method to self-extract diagnostic rules from 290 patients related to blood stasis syndrome. In result, 35 features from 52 features were selected to build classification model and five diagnostic rules were induced from the model. Wang et al. [14] proposed a DFP-growth feature selection algorithm. Based on the algorithm, 24 features were extracted from 212 attributes and the corresponding association rules were obtained for 1500 children pneumonia cases. Zhang et al. [15] proposed hierarchical latent class (HLC) models to discover latent structures of TCM dataset and applied HLC models to discover the latent variables in TCM diagnosis of kidney deficiency syndromes. In their study, 2600 cases were investigated to collect 67 symptoms related to kidney deficiency syndromes. The diagnosis based on the model made conclusions consistent with those by experts. Shi et al. [16] improved BP neural network to predict syndrome in TCM diagnosis. The results of simulation showed that neural network could not only learn experts' experiences, but also have the capability for applying the learned knowledge to more general situations. Qin et al. [17] applied Rough set (RS) method in the diagnosis of rheumatoid arthritis. The results showed that the diagnostic accuracy of Rough set for rheumatoid arthritis was greatly higher than that of fuzzy set. Zheng et al. [18] presented a brand-new traditional Chinese medicine *sizheng* integrated recorder and aided syndrome differentiator (TCM-SIRD) method to implement an auxiliary diagnostic tools for traditional Chinese medicine *sizheng*.

Although many researches in TCM diagnosis have been made during past decades, there are still some problems left and deserved discussing in syndrome prediction area. Firstly, most datasets used by previous methods only contain TCM symptoms and signs, while many objective indicators in Western medicine are ignored. Actually, traditional Chinese medicine and Western medicine unite in essence [19–21] and should be combined to predict diagnosis.

Secondly, the course of previous syndrome prediction is based on the single classifier; any single classifier cannot reflect the inherently nonlinear, ambiguous, and complex [22] characteristic of TCM diagnosis, so it is difficult to improve the classification accuracy just using one single classifier.

Thirdly, various feature selection methods have been used in syndrome differentiation with different datasets, but the rules of how to select the optimal feature subset deserves further research.

In this paper, 268 liver cirrhosis cases with three different syndromes, which are stasis-heat smoldering syndrome, damp-heat smoldering syndrome and liver–kidney yin deficiency syndrome, are collected. To make our syndrome prediction more objectively and more thoroughly, a novel method called TCMSP is proposed, whose dataset combines the symptoms and signs of TCM with the Western medicine indicators as feature set. The critical features pertaining to the syndrome are selected by improved information gain method in multi-view (i.e. TCM view and Western medicine view) instead of in the single view; finally, syndrome diagnosis is predicted by the majority voting.

The rest of this paper is organized as follows: Section 2 describes the ideas how to select critical features and how to build classification model for syndrome prediction. The experimental results based on TCMSP method are shown in Section 3. Some discussions about TCMSP method are presented in Section 4. Finally, conclusion is given in Section 5.

2. Materials and methods

2.1. *Description of dataset*

Based on the criteria for case-included and case-excluded, 268 cases with three different liver cirrhosis syndromes (i.e. stasis-heat smoldering syndrome, damp-heat smoldering syndrome and liver–kidney yin deficiency syndrome) have been offered by Shanghai University of Traditional Chinese Medicine to constitute the sample dataset [21]. The dataset includes 85 cases with stasis-heat smoldering syndrome, 103 cases with damp-heat smoldering syndrome and 80 cases with liver–kidney yin deficiency syndrome. Besides, Shanghai University of Traditional Chinese Medicine offers another 138 cases whose syndromes have not been identified. We will use the proposed method to identify the syndromes of these 138 cases in this paper.

Each case includes 106 recorded features, which are regarded as the basic symptoms required by physicians to identify the liver cirrhosis syndrome in clinic. Among all features, 67 are symptoms, signs of TCM, the other 38 are lab-measured indexes and the last one is the syndrome label. A list of these features except the syndrome label are shown in Table 1.

The features are encoded using the following rules:

(i) Symptoms of TCM are encoded using the four-value ordinal scales measured by the severity degree: with 1