

# Relationship between symptoms of traditional Chinese medicine and indicator of western medicine about liver cirrhosis

Yan Wang<sup>1</sup>, Li-Zhuang Ma<sup>3</sup>, Ping Liu<sup>2</sup> & Xiao-Wei Liao<sup>1</sup>

<sup>1</sup>Department of Computer Science & Engineering, Shanghai Jiao tong University, Shanghai, China. <sup>2</sup>Institute of Liver Diseases, Shanghai University of Traditional Chinese Medicine, Shanghai, China. <sup>3</sup>Department of Computer Science & Engineering, Shanghai Jiao tong University, Center of Traditional Chinese Medicine Information Science and Technology, Shanghai University of Traditional Chinese Medicine, Shanghai, China. (wangyan8383@sjtu.edu.cn)

## ABSTRACT

Traditional Chinese medicine (TCM) is one of the safe and effective methods to treat liver cirrhosis. The practitioners of TCM assess hepatic function in term of syndrome. But the course of syndrome differentiation is subjectivity. At present most of all the researches are focused on the relationship between the syndrome and the Western medicine objective indicators such as child-pugh grade. In fact syndrome is the synthesis of signs and symptoms and collecting signs, symptoms is easy than syndrome differentiation. We try to explore the relationship between the objective Western medicine standard such as child-pugh grade, decompensation or compensation stage, active or inactive period and the signs and symptoms of TCM by using the data mining method. We use the information gain method to assess the attributes and use five typical classifiers such as logistic, Bayes-Net, NaiveBayes, RBF and C4.5 to obtain the classification accuracy. After attribute selection, we obtain the main symptoms and signs of TCM relating to the stage, period and child-pugh grade about liver cirrhosis. The experiment results show the classification accuracy is improved after filtering some symptoms and signs.

**Keywords:** Traditional Chinese medicine, Liver cirrhosis, Attribute selection, Data mining, Classification accuracy

## 1. INTRODUCTION

Liver cirrhosis is the twelfth leading cause of death by disease, killing about 26,000 people each year. Also, the cost of liver cirrhosis in terms of human suffering, hospital costs, and lost productivity is high [1]. Many efforts have been made; at last the researchers find that the treatment approach of traditional Chinese medicine (TCM) is more effective than other kinds of treatments [2-4]. Chi-

nese medicine is safe and effective because of its prescriptive methodology [5].

During the diagnostics of TCM, the diagnosis is performed based on disease entities collected by four conventional examinations: inspection, smelling, inquiry and palpation [6]. Collecting all the information, the practitioners of TCM will perform diagnosis and draw conclusions about patient's pathological conditions in term of syndromes (called zheng in Chinese). The symptoms, no matter how they are produced, are always a sign that something is out of balance in the body mind, and the goal of professional Chinese medicine is to bring the entire organism back into a state of healthy, dynamic balance. Therefore, because it is pattern discrimination which allows us to see the larger picture or the whole person, it is treatment based on pattern discrimination which allows Chinese doctors to provide safe and effective treatment without side effects [5].

Although there are many advantages in TCM, the subjectivity of the course of syndrome differentiation limits TCM's development. One cannot apply this prescriptive methodology in a professionally standard, competent way until or unless one has mastered the syndrome differentiation course. Moreover the result of syndrome differentiation isn't objective as lab parameters. A series of previous studies have shown that some correlations between the objective indicators of Western medicine and syndrome of TCM do exist [7, 8, 9].

But most of all the researches are focused on the relationship between the syndrome and the indicator [7, 8, 9]. Now that the syndrome is the synthesis of signs and symptoms and the course of syndrome differentiation is subjective. We try to explore the relationship between the objective Western medicine standard and the signs and symptoms of TCM by using the data mining method. We hope to construct the classification model which can classify a new case with the corresponding Western indicator based on his signs and symptoms of TCM.

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns.

Strong patterns, if found, will likely generalize to make accurate predictions on future data [10].

A large number of the symptoms and the signs of TCM make it difficult to estimate the parameters of a classifier model. Attribute selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. This reduces the dimensionality of the data and allows learning algorithms to operate faster and more effectively. The result is a more compact, easily interpreted representation of the target concept especially in medical region. After attribute selection we can get the key attributes that influencing the degree of liver cirrhosis. In this paper we address attribute selection problems for classify the liver cirrhosis.

## 2. MATERIALS AND METHODS

### 2.1. Dataset Construction

The sample dataset is constructed with 294 patient cases offered by Dr. Qin Zhang [9]. Qin Zhang researches on the liver cirrhosis for many years in Shanghai University of traditional Chinese medicine.

Several approaches have been introduced to assess hepatic function, such as common biochemical tests, child-pugh score, and so on [15]. The severity of cirrhosis is commonly classified with child-pugh score, decompensation or compensation stage, active or inactive period.

Child-pugh score uses bilirubin, albumin, INR, presence and severity of ascites and ncephalopathy to classify patients in child A, B or C; child A has avorable prognosis, while child C is at high risk of death.

Compensation stage belongs to child A. Decompensation stage belongs to middle and advanced liver cirrhosis. Active period shows that the patient has hepatitis clinic symptoms such as jaundice in evidence.

In this paper, in order to analyze the relationship between the number of attributes and the indicator of Western medicine about liver cirrhosis stage, we consider the following situations based on different assessing standard of liver cirrhosis. According to the decompensation stage and compensation stage, 111 cases in compensation stage, 183 cases in decompensation stage; According to the child-pugh grade, 109 cases in child A, 93 cases in child B, 92 cases in child C; According to the active period and inactive period, 212 cases in active period, 82 cases in inactive period. Considered all of the attributes besides age, sex and avoirdupois exponent could be grouped into: symptoms, signs and the results of experimental examination. The main attributes considered are: (i) forty symptoms such as lassitude and fatigue, night sweat, vexing heat in the five heart, skin itching, depression, etc. (ii) twenty-seven signs such as pale tongue, white-thick and grimy tongue fur, splenomegaly, hepatomegaly, etc. (iii) the assessing standard of liver cirrhosis (iv) age, sex and avoirdupois exponent.

In the patient records diagnosed by clinical physicians of TCM, the symptoms are in 4 grade scorings. If the

patient has no the symptom he gets 1 point. If he has these symptoms he gets 2 to 4 according the state of the illness. Signs are in 2 grade scorings. If the patient has no the sign he gets 1 point. If he has these signs he gets 2. Up to the liver cirrhosis stage, 1 represents compensation stage and 2 represents decompensation stage; child-pugh grade are in 3 grades, that is 1, 2, 3 degree; 1 represents the active period and 2 represents inactive period [9]. Besides, the datasets do not contain cases with missing values. Therefore, our research work does not require considering the situations of handling missing value.

### 2.2. Method

#### 2.2.1. Attributes Selection

During the course of data mining, in practice, adding irrelevant attributes to a dataset often “confuses” machine learning systems. The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean. However, automatic methods can also be useful. Reducing the dimensionality of the data by deleting unsuitable attributes improves the performance of learning algorithms. It also speeds them up, although this may be outweighed by the computation involved in attribute selection. More importantly, dimensionality reduction yields a more compact, more easily interpretable representation of the target concept, focusing the user’s attention on the most relevant variables [10].

We use the information gain method to assess the attributes. It evaluates attributes by measuring their information gain with respect to the class. If  $X$  and  $Y$  are random variables,  $y \in Y$  equations 1 and 2 give the entropy of  $Y$  before and after observing  $X$ .

$$H(Y) = - \sum_{y \in Y} p(y) \log_2 p(y) \quad (1)$$

$$H(Y|X) = - \sum_{x \in X} \sum_{y \in Y} p(y|x) \log_2 p(y|x) \quad (2)$$

In equation 1 and 2, where  $p(y)$  is the probability of  $y$  in  $Y$ , determined by dividing the number of tuples of  $y$  in  $Y$  by  $|Y|$ , the total number of tuples in  $Y$ ;  $p(y|x)$  is the conditional probability. A log function to the base 2 is used, because the information is encoded in bits.

The amount by which the entropy of  $Y$  decreases reflects the additional information about  $Y$  provided by  $X$  and is called the information gain. Information gain is given by [14]

$$\begin{aligned} \text{gain} &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= H(Y) + H(X) - H(X,Y) \end{aligned} \quad (3)$$

Some classification algorithms deal with nominal attributes only and cannot handle ones measured on a numeric scale. To use them on general datasets, numeric attributes must first be “discretized” into a small number of distinct ranges.

In the dataset the avoirdupois exponent is numeric so

it need discretizes first. We use the entropy-based method to discretize the attribute. The entropy is defined as equations 1 and the minimum description length (MDL) principle is used to stop the discretization procedure. The MDL principle states that the “best” attributes is the attribute that can be encoded with the least number of bits. The MDL principle can be taken as an operational definition of Occam’s Razor. More formally, if  $T$  is a theory inferred from data  $D$ , and then the total description length is given by

$$DL(T,D) = DL(T) + DL(D/T) \quad (4)$$

Equation (4) measures all description lengths in bits. After computing the information gain, we sort attributes by their individual evaluation. The procedure has low computational complexity.

### 2.2.2. Cross-Validation

Because the data is not enough, we use an important technique that is 10 times 10-fold cross-validation for accuracy estimation. In cross-validation, the data is split into ten approximately equal partitions and each in turn is used for testing and the remainder is used for training. That is, use nine-tenths for training and one-tenth for testing and repeat the procedure ten times so that, in the end, every instance has been used exactly once for testing. This is called tenfold cross-validation.

To get a reliable error estimate, we repeat the cross-validation process 10 times, and then average the results. This involves invoking the learning algorithm 100 times on datasets that are all nine-tenths the size of the original.

### 2.2.3. Classification Method

The overarching goal of classification is to build a model

that can be used for prediction. In our study, the goal of classification is to predict whether a patient is in decompensation stage and whether he is in active period and which grade is he in. We use five typical classifiers: logistic, BayesNet, NaiveBayes, RBF and C4.5 to analysis how the number of attributes affects the accuracy of classification. The five classifiers are used widely in the many areas especially in medicine area.

For example: Yanan Sun shows that naive Bayes and Bayenet have better classification capability in Chinese traditional medical clinical diagnosis model [11], Qu haibin apples the decision tree to classify 194 patient records , the results show that the decision tree method is likely a promising method to self-extract diagnostic rules from patient records of Chinese medicine [12]. By using logistic method, Hua Cong researches the development tendencies of lung disease, physiological functions of lung that are inversely proved by clinical symptoms.

Based on the classification accuracy, we filter some unimportant attributes. The main attributes will help us to research the relationship between the indicator of Western medicine and the signs, symptoms of TCM and improve the predication accuracy.

## 3. RESULTS

We compute the information gain of each attribute while choosing the decompensation or compensation stage as the classify label, the child-pugh grade as the classify label, the active or inactive period as the classify label respectively. Based on the five general classifiers such as logistic, BayesNet, NaiveBayes, RBF and C4.5, we construct five classification models.

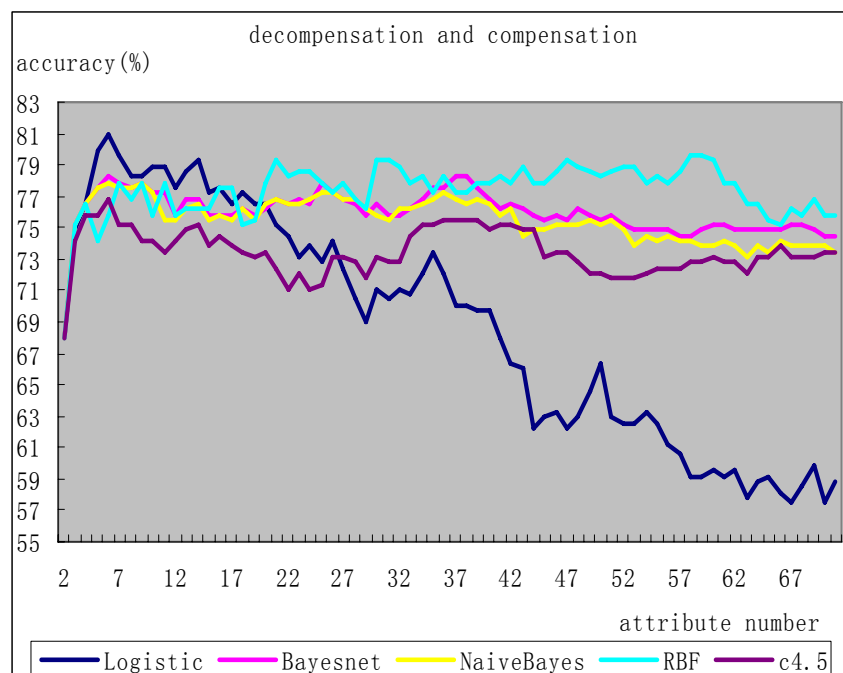


Figure 1. Decompensation or compensation stage.

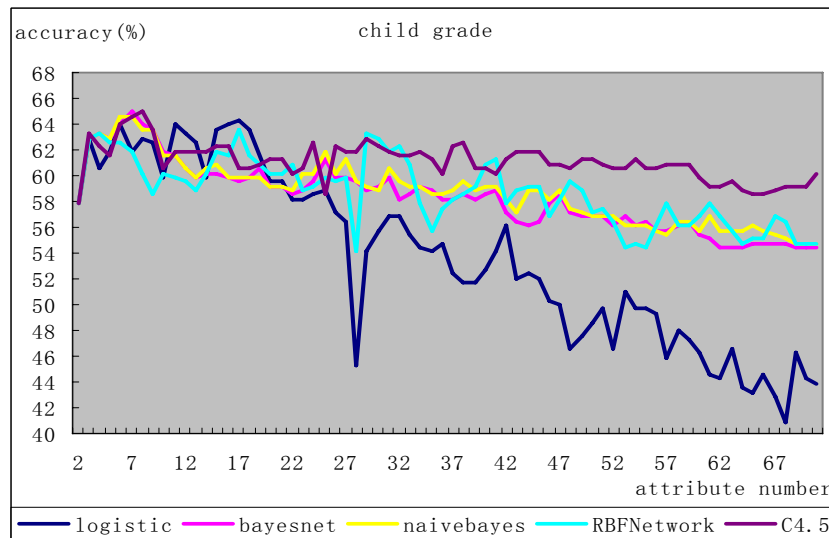


Figure 2. Child-pugh grade.

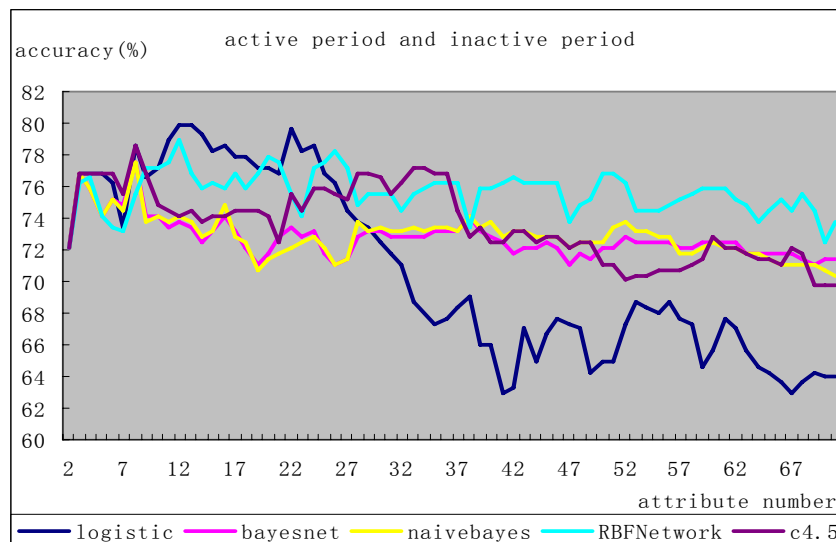


Figure 3. Active or inactive period.

Figure 1, Figure 2, and Figure 3 shows the relationship between the stage, the child-pugh grade, the period of liver cirrhosis and the number of attributes respectively. From three figures we can see that the accuracy of the classifiers will change with the number of attribute. Logistic algorithm is the most sensitive to the number of attribute. The classification capability of BayesNet is close to the NaiveBayes method, it shows the attributes are in dependently. When we analysis the relationship between the objective standard such as child-pugh grade, active or inactive period, decompensation or compensation stage and the signs, symptoms of TCM, we should filter some week attributes first.

Based on the above figures, we also know the key attributes related to the indicators of Western medicine about liver cirrhosis. Table1 shows the main signs and symptoms of TCM related to the corresponding indicators. Attributes selected are in accord with the research of a

post- doctoral of Shanghai University of traditional Chinese medicine.

After filter some weak attributes, we compare the classification accuracy with the original dataset. Table 2 shows the result.

Using the original dataset, to predicting whether the patient is in compensation stage the accuracy is only 58.8435% based on logistic model. After filtering some attributes, the prediction accuracy is over 80%. While predicting which child-pugh grade is the patient, the accuracy is 54.4218% based on BayesNet model. After filtering some attributes the accuracy is improved 10%.

From table 2, we can know if we try to predict whether the patient is in compensation stage or not, the best predictor is logistic model.

When predicting which child-pugh grade is the patient in, the best predictor is c45.

Logistic model is the suitable to predict whether the

**Table 1.** The main attributes of the Wetsern medicine indictors.

Main signs and symptoms of TCM	
Compensation or de-compensation stage	Abdominal distension, low limbs puffy swelling, yellow body, yellow eyes, and yellow urine
Child-pugh grade	Abdominal distension, scant urine, fatigued and heavy limbs, low limbs puffy swelling, yellow body, yellow eyes, and yellow urine
Active or inactive period	Lassitude and fatigue, vexing heat in the five hearts, abdominal distension, constipation, sloppy stool, fatigued and heavy limbs, distending pain in flanks, yellow body, yellow eyes, yellow urine, dim complexion

**Table 2.** The classification accuracy(%) after or before filter attributes.

	Dataset before filtering			Dataset after filtering		
	Compensation	Child-pugh	Active	Compensation	Child-pugh	Active
Logistic	58.8435	43.8776	63.9456	80.9524	64.2857	79.932
BayesNet	74.4898	54.4218	71.4286	78.2313	64.966	77.551
NaiveBayes	73.4694	54.7619	70.4082	77.8912	64.6259	77.551
RBF	75.8503	54.7619	73.8095	79.5918	63.6054	78.9116
C4.5	73.4694	60.2041	69.7279	76.8707	64.966	78.5714

patient is in active period or not. When we research on the liver cirrhosis, we can choose the best method to predict the patient situation.

#### 4. CONCLUSION

Chinese medicine becoming vibrant around the world as an important alternative source of health care, researches on modernizing TCM become to attract substantial attentions from the practitioners of TCM. But the subjective indicators of TCM limit its development.

In this paper, we express the disease severity objectively in term of the symptoms and signs of TCM. But too much attributes often confuse the practitioners of TCM. We use the data mining method to filter some unimportant attributes. From the experiment, the main signs, the symptoms about the active period, compensation stage, and child-pugh grade are showed. Also we show that the compact dataset obtain the improved classification accuracy. At present, the accuracy is only about 70%. Improving the classification accuracy is our research goal. If the accuracy is improved, we can express the TCM theory objectively.

#### ACKNOWLEDGEMENT

We gratefully acknowledge all the researchers from the Shanghai University of Traditional Chinese Medicine for the TCM databases and

discussion of TCM topics. This research is partly supported by the traditional Chinese medicine etiologic study on the theory of insufficiency and damage causing stasis and blockage in liver cirrhosis of China 973 project (No. 2006CB504801), National Science Fund of China (No. 60521002).

#### REFERENCE

- [1] <http://digestive.niddk.nih.gov/diseases/pubs/cirrhosis>.
- [2] Ping Liu. (2002) *Contemporary hepatology in traditional Chinese medicine*. People's Medical Publishing House, Beijing.
- [3] Qin Zhang, Hong Qiu, Lei Wang, *et al.* (2007) Correlation between syndromes of posthepatic cirrhosis and biological parameters: a report of 355 cases. *Journal of Chinese Integrative Medicine*, **5** (2), 130-133.
- [4] Qin Zhang, Ping Liu, HuiFen Chen, *et al.* (2003) Multi-analysis characteristics of traditional Chinese medical syndrome of hepatocirrhosis. *Chinese Journal of integrated traditional and western medicine on liver diseases*, **13** (2), 69-72.
- [5] Bob Flaws & Philippe Sionneau, (2005) *The treatment of modern western medical diseases with Chinese medicine*, Blue Poppy Press.
- [6] Xuewei Wang, Haibin Qu, Ping Liu, Yiyu Cheng. (2004) A self-learning expert system for diagnosis in traditional Chinese medicine. *Expert systems with applications*, **26**, 557-566.
- [7] Guanhua He, Lanping Zhu. (2002) Exploration on relationship between cirrhosis of liver's Traditional Chinese Medical Syndrome Differentiation Typing and Child-pugh degree, Complication. *Liaoning Journal of Traditional Chinese Medicine*, **29**(1), 12-12.
- [8] Fangshi Zhu, Ping Pu, Haihang Zhu, *et al.* (1997) Study on correlation between the syndrome type of TCM and the classification of

Child-Pugh in patients with cirrhosis. *Chinese journal of integrated traditional and western medicine on liver diseases*, **7**(4).

[9] Qin Zhang. (2005) *Study on disease-pattern-efficacy integration in posthepatic cirrhosis*. Shanghai University of traditional Chinese medicine, Shanghai.

[10] Ian H. Witten, Eibe Frank. (2006) *Data Mining Practical Machine Learning Tools and Techniques*, China Machine Press, Beijing.

[11] Yanan Sun, Shiyong Ning, Mingyu Lu, *et al.* (2006) Chinese traditional medical clinical diagnosis for coronary heart disease based on bayes classification. *Application research of computer*, **11**, 164-166.

[12] Haibin Qu, Lifeng Mao, Jie Wang. (2005) Method for self-extracting diagnostic rules of blood stasis syndrome based on decision tree. *Chinese Journal of Biomedical Engineering*, **24**(6), 709-727.

[13] Hua Cong, Qiming Zhang. (2002) Logistic regression on the diagnosis and prescription of lung disease. *Journal of Shandong University of TCM*, **26**(5), 322-327.

[14] Mark A. Hall. (1999) *correlation-based feature selection for machine learning*, University of Waikato, New Zealand.

[15] G Ercolani. (2006) predictive indices of morbidity and mortality after liver resection. *annals of surgery*, **244**(4),635-637.