一种用于中药最优配方挖掘的3-阶段选举筛选算法*

向正贵+

(清华大学 计算机科学与技术系 智能技术与系统国家重点实验室,北京 100084)

A 3-Stage Voting Algorithm for Mining Optimal Ingredient Pattern of Traditional Chinese Medicine

XIANG Zheng-Gui+

(State Key Laboratory of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: 86-10-62782266 ext 8405, E-mail: xzhgui00@mails.tsinghua.edu.cn

Received 2002-06-28; Accepted 2003-03-03

Xiang ZG. A 3-stage voting algorithm for mining optimal ingredient pattern of traditional Chinese medicine. *Journal of Software*, 2003,14(11):1882~1890.

http://www.jos.org.cn/1000-9825/14/1882.htm

Abstract: In this paper a voting algorithm is presented to mine the optimal ingredient pattern of Chinese medicine. First this paper visualizes the data about the curative effect of different prescriptions of Chinese medicine ingredients on several cardio-indexes. Then the features about the curative effect from the curves generated by visualizing the data are extracted. Finally, a voting algorithm is adopted in which votes are held in three stages: the preliminary vote generates the features of each test group and cardio-index by each sample; the metaphase vote obtains the value of the curative effect by each feature of the given group and index; the final vote mines the optimal ingredient pattern. Experimental results show that the method is effective for the kind of problems and instructive to clinic medicine experiment. Its potential applications include the development of new medicine, optimal combination investment of venture capital and gene series analysis etc.

Key words: ingredient pattern mining; Chinese medicine; voting algorithm; feature extraction

摘 要: 提出了一种用于挖掘中药最优配方的三阶段选举筛选算法.首先对不同配方的中药新药的临床疗效数据进行可视化.然后,从可视化后的疗效数据曲线中提取若干疗效特征.最后,采用三阶段选举算法筛选新药配方.第一阶段是初选,每一实验样本投票产生指定配方指定指标的特征值.第二阶段是中选,指定配方指定指标的特征值投票产生指定配方指定指标的药效值.第三阶段是终选,每一指标的药效值投票产生指定配方的综合药效值.通过权衡所有配方的综合药效值,就可以找到最优中药配方.实验结果显示该方法对于这类问题是有效的,能较好挖掘出最优中药新药配方,并对新药的临床实验具有指导意义.该方法的潜在应用包括新药开发、风险投资最优组合和基因序列分析等.

关键词: 成分模式挖掘;中药配方;选举算法;特征提取

*XIANG Zheng-Gui was born in 1977. He is a master graduate at the State Key Laboratory of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University. His current research interests include data mining, computer vision and information process in biology and medicine.

中图法分类号: TP18 文献标识码: A

1 Introduction

A formula of a traditional Chinese medicine consists of corresponding herbs with a suitable dosage, which includes a principal herb producing the leading effects, an assistant herb increasing the effects of principal herb, an adjuvant herb treating accompanying symptoms or eliminating drastic actions of the principal and assistant herbs, and a dispatcher herb leading the other herbs to the affected site or coordinating the effects of various ingredients. On bases of definite diagnosis and therapy, syndrome differentiation, and the principle of herb cooperation and proper preparation form, a formula has so cooperative effects of herbs that it can increase the curative effects through additive and synergetic effects or decrease the toxic and side effects of a single herb through mutual restraint and antagonism[1]. The compatibility of a single herb includes synergism, assisting, detoxification or restraint, antagonism, increasing toxicity[2]. By adjusting compatibility of a formula through the modification of assistant and adjuvant herbs, preparation forms and dose, it is more suitable to complicated diseases. For example, minor purgative decoction with rhubarb 12g as principal herb, immature bitter orange 9g as assistant herb, magnolia 6g as adjuvant and dispatcher herb, has an effect of purging heat and easing defecation and is indicated for excess syndrome marked by tidal fever, delirium, constipation and abdominal pain etc., while magnolia three ingredients decoction with magnolia 24g as principal herb, immature bitter orange 15g as assistant herb, rhubarb 12g as adjuvant and dispatcher herb, has an effect of activating vigor and easing defecation and is used for abdominal distention[1].

However, the modification of a dose or ingredients of a herb in a formula of Chinese medicine is myriad, so it is important to find an efficient method for mining the optimal ingredient pattern of a given formula or predicting a new formula. Some work has been done toward this goal. For example, Zaptron System, Inc. has carried out the research about the optimization of new drug exploration from Chinese herbs[4].

This paper presents a method based on a feature extracting and voting algorithm for mining the optimal ingredient patterns of formula of Chinese medicine. Its potential applications include the development of Chinese medicine or compound of Western medicine, optimal combination investment of venture capital and gene series analysis etc.

The first contribution of this paper is the method for extracting the features about the curative effect of different prescriptions of Chinese herbs. In Ref.[5], a distribution algorithm was presented to select feature subset for medicine data. In Ref.[6], the co-location pattern discovery process was used to mine the subset of features. Bojarezuk *et al*. presented a constrained-syntax genetic programming method for mining classification rule from medical data[7]. Lavrak introduced data mining techniques in medicine[8]. In this paper, we extract feature based on some rules about curative effect. We first preprocess the data by sub-sampling, visualizing, normalizing and smoothing etc. Then we construct the data model in the feature space to mine the optimal prescription pattern evaluated by test sample set. Finally, we represent the mined pattern by some visual technologies.

The second contribution is to present a 3-stage voting algorithm for selecting the optimal prescription from all the candidates. In the voting algorithm, votes are held in three stages. In the first stage, each sample is considered to be a voter while the feature vector of each cardio-index is as the ballots of the voter. In the second stage, the test groups of Chinese medicine for each cardio-index are viewed as candidates and the features of each index are as a voter. In the third stage, the test groups of Chinese medicine are viewed as candidates and each cardio-index is as a voter.

This paper is organized as follows. In Section 2, we model the data about the curative effect of formula of

Chinese medicine. The method for mining the ingredient pattern is represented in Section 3. The results of experiments are given in Section 4. Finally, we discuss the conclusions of the method in Section 5.

2 Data

2.1 Data acquisition

Our data come from Tianjin University of Traditional Chinese Medicine. The herbs of the formula for curing a cardiopathy are composed of Red Sage Root and Notoginseng with the proportions: 10/6, 10/3, 10/1, 1/1, 10/0, 0/10, 1/10. As a comparison to the proportions, the western medicine named "Eliminating-cardiodynia" and the model of compound and pseudo-surgery are designed to test the cardio-index of all samples. The samples tested are dogs with some conditions controlled. The cardio-index tested can be classified as five kinds. The first kind denotes the intuitionistic index about anemia of cardiac muscle, including ΣST and *NST* based on cardiograph of epicardium, morphology of anemia in left-cavity of heart and whole heart, *CKMB* and *CTNL* based on biochemical index of anemia. The second kind is such key cardio-index as blood stream of coronary artery and oxygen wastage of cardiac muscle. The third kind indicates the index about kinetics of blood stream such as cardiotach, average pressure of artery, peak value and maximum raise ratio of inner pressure in left cardio-cavity, pressure of dilatation in left cardio-cavity and maximum decline ratio of inner pressure in left cardio-cavity, blood output and index of heart, resistance outside blood vessel and work index of left cardio-cavity. The fourth kind includes CO_2, endothelium element. The fifth kind is such freeradicals as *SOD* and *MDA*. The data are sampled from 10 test groups including seven proportion groups and three comparison groups mentioned above. In each group, cardio-index are sampled at most in seven points of time for each sample. The seven points of time are respectively the point before the dog is ill, the 30^{th} minute after ill, the 30^{th}, 60^{th}, 90^{th}, 120^{th}, 180^{th} minute after the dog took the Chinese medicine.

2.2 Data model

Data variable. In order to describe more easily the data above, we define four variables to denote the different dimensions of the data: the variable g indicates the test groups with the range of value in [1,10], the variable S indicates the samples, the variable x indicates the cardio-index with the range of value in [1,25] and the variable t indicates the points of time with the range of value in [1,7]. Here, g=1 denotes the model of compound; g=2 the western medicine "Eliminating-cardiodynia"; g=3 pseudo-surgery and g=4,…,10 the proportions of Red Sage Root to Notoginseng 10/6, 10/3, 10/1, 1/1, 10/0, 0/10, 1/10 respectively. As to variable x, x=1,…,8 indicate respectively the anterior eight kinds of cardio-index while x=11,…,24 indicate correspondingly the posterior fourteen kinds of index.

Data function. Based on above defined variables, we further define the function $d(g,s,x,t)$ to express the data sampled by the given test group g for each sample s about cardio-index x in the point of time t.

Rule function. For each cardio-index, there is a rule to evaluate the data of the curative effect. For example, the rule for the cardiograph of epicardium is that the larger is the value of the data, the more serious is the degree of anemia of cardiac muscle; the rule for blood stream of coronary artery is that the higher the value, the better the state of the heart; and the rule of average pressure of artery is that the result is good if the effect of drugs on it is small. We define a function $r(x)$ as follows to express the rule of curative effect about cardio-index x:

$$r(x)=\begin{cases}-1, & \text{if the smaller the data, the better the state of the heart}\\ 0, & \text{if the smaller the change of the data, the better the state of the heart}\\ 1, & \text{if the larger the data, the better the state of the heart}\end{cases} \quad (1)$$

Based on the above formula, we get the rule values about all the cardio-indexes, as shown in Table 1 where x

denotes cardio-index while $r(x)$ indicates the rule value of x.

Table 1 The value of rule about each cardio-index

x	1	2	3	4	5	6	7	8	11	12	13
$r(x)$	−1	−1	−1	−1	−1	−1	1	−1	0	0	1
x	14	15	16	17	18	19	20	21	22	23	24
$r(x)$	1	−1	1	1	1	−1	−1	1	−1	1	−1

Sampling vector. Because of the points sampled are different for each cardio-index, we design a vector $\vec{S}(x)$ for each index x to express its states of sampling in seven points of time:

$$\begin{aligned}\vec{S}(x) &= \{b_1(x), b_2(x), b_3(x), b_4(x), b_5(x), b_6(x), b_7(x)\} \\ &\triangleq b_1(x)b_2(x)b_3(x)b_4(x)b_5(x)b_6(x)b_7(x)\end{aligned} \tag{2}$$

where $b_t(x)$ (i=1,…,7) denotes whether the index x is sampled in the point of time t=i (i=1,…,7), as given in expression (3):

$$b_i(x) = \begin{cases} 0, \text{ if not sampled} \\ 1, \text{ if sampled} \end{cases} \quad (i = 1,..., 7) \tag{3}$$

And we use $n_t(x)$ to denote the sampling number of the index x:

$$n_t(x) = \sum_{i=1}^{7} b_i(x) \tag{4}$$

The sampling vector and the sampling number of all indexes are shown in Table 2.

Table 2 The sampling vector and number about each cardio-index

x	1,2,7,11~17	3,4,6	8	18~20	5,21~24
$\vec{S}(x)$	1111111	1000000	1110100	0111111	0110000
$n_t(x)$	7	1	4	6	2

3 Method

3.1 Feature extraction

To describe and cluster the data validly, we design a feature vector as follows:

$$\vec{F}(g,s,x) = \{f_{10}, f_{11}, f_{20}, f_{21}, f_{30}, f_{31}, f_{40}, f_{41}, f_{50}, f_{51}, f_{60}, f_{61}, f_{70}\} \tag{5}$$

where f_{10}, f_{20}, f_{30}, f_{40} and f_{50} denote respectively the change of the curative effect in the 30th minute, during the periods from the 30th to the 60th minute, from the 60th to the 90th minute, from the 90th to the 120th minute and from the 120th to the 180th minute after the dog took the Chinese medicine and f_{11}, f_{21}, f_{31}, f_{41} and f_{51} denote respectively the change ratio of the curative effect during the above five periods. f_{60} and f_{61} indicate respectively the change and change ratio of the curative effect during the period from the beginning time when the dog is ill to the 180th minute after it took the new medicine. f_{70} denotes the fluctuant ratio of the curative effect during the period from the 30th minute to the 180th minute after it took the Chinese medicine. For different cardio-indexes, because the sampling vector is different, the expressions for the same feature are also different in the following formulae.

$$f_{i0}(x) = \begin{cases} -\left|d(g,s,x,i+2) - d(g,s,x,i+1)\right|, r(x) = 0 \wedge b_{i+2}(x) = 1 \wedge b_{i+1}(x) = 1 \\ r(x) * [d(g,s,x,i+2) - d(g,s,x,i+1)], r(x) \neq 0 \wedge b_{i+2}(x) = 1 \wedge b_{i+1}(x) = 1 \\ r(x) * d(g,s,x,1), r(x) \neq 0 \wedge n_t(x) = 1 \\ 0, \text{Other} \end{cases} \quad (i = 1,...5) \tag{6}$$

$$f_{i1}(x)=\begin{cases}\dfrac{d(g,s,x,i+2)-d(g,s,x,i+1)}{d(g,s,x,1)-d(g,s,x,2)}, d(g,s,x,1)\ge 0\wedge b_{i+2}(x)=1\wedge b_{i+1}(x)=1 & (i=1,...,5)\\ 0, \text{Other}\end{cases} \tag{7}$$

$$f_{60}(x)=\begin{cases}-\left|d(g,s,x,7)-d(g,s,x,2)\right|, r(x)=0\\ r(x)*[d(g,s,x,t)-d(g,s,x,2)], r(x)\ne 0\wedge b_t(x)=1\wedge b_j(x)=0 \quad (j=t+1,...,7)\\ 0, \text{Other}\end{cases} \tag{8}$$

$$f_{61}(x)=\begin{cases}1-\dfrac{d(g,s,x,7)-d(g,s,x,2)}{d(g,s,x,1)-d(g,s,x,2)}, r(x)=0\\ \dfrac{d(g,s,x,t)-d(g,s,x,2)}{d(g,s,x,1)-d(g,s,x,2)}, r(x)\ne 0\wedge b_t(x)=1\wedge b_j(x)=0 \quad (j=t+1,...,7)\end{cases} \tag{9}$$

$$f_{70}(x)=\begin{cases}\sum_{t=3}^{T}|\,d(g,s,x,t)-\mathrm{d}(g,s,x,t-1)\,|, r(x)=0\wedge b_T(x)=1\wedge b_j(x)=0 \quad (j=T+1,...,7)\\ \dfrac{1}{2}-\dfrac{1}{8}\sum_{t=3}^{6}\mathrm{sgn}([d(g,s,x,t)-d(g,s,x,t-1)]*[d(g,s,x,t+1)-d(g,s,x,t)]),\\ \qquad\qquad r(x)\ne 0\wedge b_t(x)=1\wedge b_j(x)=0 \quad (j=t+1,...,7)\\ 0, \text{Other}\end{cases} \tag{10}$$

To express intuitionally the formulae mentioned above, we take $\sum ST$ (the cardiograph of epicardium) for an example. Based on the visualization technologies[3], we get the graph shown Fig.1, from which we can calculate the feature functions about curative effect of the comparison group of western medicine on $\sum ST$ of the 5th sample, see Table 3.

Table 3 The feature functions of ΣST (x=1, g=2, s=5)

Name	$f_{10}(1)$	$f_{11}(1)$	$f_{20}(1)$	$f_{21}(1)$	$f_{30}(1)$	$f_{31}(1)$
Expression	|AB|	|AB|/|OA|	|BC|	|BC|/|OA|	|CD|	|CD|/|OA|
Name	$f_{40}(1)$	$f_{41}(1)$	$f_{50}(1)$	$f_{51}(1)$	$f_{60}(1)$	$f_{61}(1)$
Expression	−|DE|	−|DE|/|OA|	|EF|	|EF|/|OA|	|AF|	|AF|/|OA|

According to $f_{10}(1)>0$, $f_{20}(1)>0$, $f_{30}(1)>0$, $f_{40}(1)<0$, $f_{50}(1)>0$, there is one turning point (5,100) in this curve. So $f_{70}(1)=1/2-(1+1-1-1)/8=1/2$.

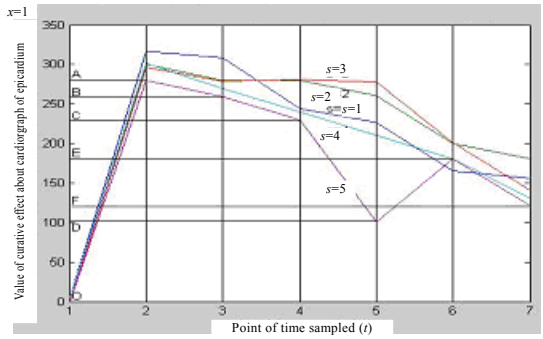

Fig.1 The curve about the comparison group of western medicine (g=2)

3.2 Feature normalize

Given a cardio-index x of a given sample s, we define a matrix $M(s,x)$ to describe its feature vectors of all test groups.

$$M(s,x)=\{\vec{F}^T(1,s,x),\vec{F}^T(2,s,x),...,\vec{F}^T(10,s,x)\}=\{m_{i,j},i\in[1,13],j\in[1,10]\} \tag{11}$$

where $\vec{F}^T(g,s,x)(g=1,...,10)$ is the transpose of $\vec{F}(g,s,x)$.

For each feature i, we get the maximum and minimum of all test groups $m_{\max}(i)$ and $m_{\min}(i)$, and then obtain the normalized matrix $M'(s,x)$:

$$m_{\max}(i)=\max\{m_{i,1},...,m_{i,10}\},i\in[1,13] \tag{12}$$

$$m_{\min}(i)=\min\{m_{i,1},...,m_{i,10}\},i\in[1,13] \tag{13}$$

$$M'(s,x)=\left\{m'_{i,j}=\begin{cases}\dfrac{m_{i,j}-m_{\min}(i)}{m_{\max}(i)-m_{\min}(i)}, m_{\max}(i)>m_{\min}(i)\\ 1, m_{\max}(i)=m_{\min}(i)\end{cases},i\in[1,13],j\in[1,10]\right\} \tag{14}$$

3.3 Voting algorithm

The voting algorithm is a well-known and widely used method for achieving fault-tolerance in distributed systems[9,11,12]. To select the optimal prescription from all the candidates, we present a different voting algorithm. In this algorithm, votes are held in three stages.

The first stage (Preliminary Vote):

Candidates: each feature f_i of each test group g for the formula of the Chinese medicine for each cardio-index x

Voter: each sample s tested

Weight of voter: $w_1(s)$ s.t. $\sum_s w_1(s)=1$, where s is the number of the sample

Ballot: the value of feature $f_i(g,x)$

Input: the feature vectors based on the sample for each test group g to each index x: $\{f_i(g,s=1,x),\ldots\}$

Output: the results voted by all samples for each feature f_i of each test group g

Method:

```
1)      for each cardio-index x{
2)      for each test group g{
3)          for each feature f_i {
4)              for each sample s
5)                  vote with its value of feature f_i(g,s,x);
6)              get the ballot  f_i(g,x)=Σ_s w_1(s)*f_i(g,s,x);
7)          }
8)      }
9)      for each feature voted with the ballot f_i(g,x)
10)         normalize f_i(g,x);
11)     store f_i(g,x) to database for each group and each index;
12)     }
```

The second stage (Metaphase Vote):

Candidates: each test group g_i for each cardio-index x

Voter: each feature $f\in\{f_{10},f_{11},f_{20},f_{21},f_{30},f_{31},f_{40},f_{41},f_{50},f_{51},f_{60},f_{61},f_{70}\}$

Weight of voter: $w_2(i)$ s.t.$\sum_i w_2(i)=1$, where i is the number of the feature

Ballot: the value of the curative effect $f'_i(g,x)$ after normalized

Input: the ballots $\{f_i(g,x),i[1,13]\}$ gained in the first stage of all features for each index x and each test group g

Output: the value $v(g,x)$ and grade $c(g,x)$ of the curative effect voted by all features for each index x and each test group g

Method:

1) get $f'_i(g,x)$ by normalizing $f_i(g,x)$ with Eq.(14)
2) for each cardio-index x {
3) for each test group g {
4) for each feature f_i
5) vote with its ballot $f'_i(g,x)$ of the first stage;
6) get the ballot $v(g,x)=\sum_i w_2(i)*f'_i(g,x)$;
7) get the grade $c(g,x)=round(v(g,x)/0.2+1)$;
8) }
9) store $v(g,x)$ to database for each index and each group;
10) }

The third stage (Final Vote):

Candidates: each test group g_i

Voter: each cardio-index x

Weight of voter: $w_3(x)$ s.t.$\sum_x w_3(x)=1$, where x is the number of the cardio-index

Ballot: the value of curative effect for each cardio-index x and test group g_i

Input: the value $v(g,x)$ generated in the second stage for each index x and each test group g

Output: the final result $u(g)$ of each test group g and the number of the optimal test group

Method:

1) for each test group g {
2) for each cardio-index x
3) vote with its ballot $v(g,x)$ of the second stage;
4) get the total ballot of group g: $u(g)=\sum_x w_3(x)*v(g,x)$;
5) normalize result;
6) }
7) get the optimal pattern of ingredient proportions;
8) hypothesis testing for the optimal pattern;
9) clustering analysis based on the distance between indexes;

4 Experiments

We first get the results of the preliminary vote about each cardio-index and test group as shown in Table 4. Then we get the results of the metaphase vote about each index in Table 5. Finally, we get the results of the final vote in Table 6.

Table 4 The result of the preliminary vote (x=1, g=1)

The Preliminary Vote: samples as voters and features as candidates (x=1, g=1)											
	SampleNo	f10	f11	f20	f21	f30	f31	f40	f41	f50	f51	f70
	010101	.86	.0032	-1.75	-.0065	-2.09	-.0077	-1.61	-.0059	3.84	.0142	0
▶	010102	-1.17	-.0046	-5.88	-.0232	7.6	.03	-3.56	-.0141	-4.16	-.0164	.5
	010103	-.84	-.0029	-1.06	-.0037	-8.52	-.0294	-.04	-.0001	8.22	.0283	.25
	010104	-1.58	-.0059	-1.82	-.0067	-9.98	-.037	2.76	.0102	-4.5	-.0167	.25
	010105	-4.61	-.0164	-1.54	-.0055	-.84	-.003	-8.24	-.0293	19.48	.0693	.25
	Vote	-1.468	-.0054	-2.41	-.0088	-2.766	-.0101	-2.138	-.0078	4.576	.0168	.25

Table 5 The result of the metaphase vote (x=1)

The Metaphase Vote: features as voters and groups as candidates											
	Criterion	FeatureNo	Model	WestDrug	Surgery	Prop10_6	Prop10_3	Prop10_1	Prop1_1	Prop10_0	Prop0_10	Prop1_10
▶	01	f10	.1444	.5797	.1745	.7548	1	0	.0562	.1737	.0798	.1678
	01	f11	.1552	.5645	.1859	.801	1	0	.0654	.1853	.0807	.1791
	01	f20	.1885	1	.2617	.2459	.1725	.6258	0	.63	.3082	.256
	01	f21	.2018	1	.2789	.2614	.1868	.6842	0	.6684	.3333	.2728
	01	f30	0	.754	.0922	.4779	1	.8145	.0108	.8951	.226	.1146
	01	f31	0	.7201	.0942	.5075	1	.8741	.0103	.916	.2444	.1175
	01	f40	0	1	.0413	.6524	.4612	.234	.1926	.6106	.1056	.0566
	01	f41	0	1	.0445	.7347	.486	.2635	.2088	.6583	.1204	.061
	01	f50	.1266	1	.0131	.7676	.8434	.8792	.3045	.4988	0	.1112
	01	f51	.1383	.9985	.0154	.8676	.8881	1	.3321	.5391	0	.1222
	01	f70	.5	0	0	1	1	.5	.5	.5	1	.5
	01	vote	.1212	.7181	.1001	.5892	.6698	.4896	.1401	.5229	.2082	.1632

Table 6 The result of the final vote

The Final Vote: cardio-index as voters and group as candidates											
	Criterion	Weight	Model	WestDrug	Surgery	Prop10_6	Prop10_3	Prop10_1	Prop1_1	Prop10_0	Prop0_10	Prop1_10
▶	01	.085	.1212	.7181	.1001	.5892	.6698	.4896	.1401	.5229	.2082	.1632
	02	.085	.0833	.8333	0	.2978	.3283	.2448	.2091	.3138	.0833	.2305
	03	.085	0	.0459	.0833	.0492	.0463	.0373	.0356	.0358	.0327	.0415
	04	.085	0	.0863	.1429	.0752	.0748	.0542	.0575	.0545	.0368	.0608
	05	.08	0	.0684	.0833	.073	.0548	.0705	.0566	.0666	.0392	.0631
	06	.08	0	.0717	.0833	.0647	.0559	.065	.0739	.0715	.052	.0392
	07	.07	.1582	.3421	.3245	.4849	.5449	.289	.3944	.3087	.4472	.431
	08	.07	0	.1893	.0837	.0996	.0751	.1653	.1175	.0228	.1392	.049
	11	.055	.4152	.4718	.6953	.5424	.5608	.4928	.5355	.5401	.4486	.3759
	12	.055	.6322	.417	.579	.5569	.4688	.506	.6349	.5241	.6017	.5327
	13	.055	.5805	.4259	.2531	.3649	.3253	.3159	.5425	.3584	.5056	.491
	14	.055	.0447	.4818	.5723	.6299	.5866	.4774	.5814	.4974	.3722	.5389
	15	.03	.1855	.5786	.6313	.6194	.6294	.5612	.5842	.5933	.5369	.4731
	16	.03	.2539	.4243	.4905	.4736	.5285	.5095	.553	.4479	.3942	.5176
	17	.03	.3285	.5335	.3013	.5688	.4413	.3762	.5272	.3889	.5803	.4142
	18	.03	.5385	.6755	.6009	.6765	.6634	.6243	.6931	.6944	.6762	.656
	19	.005	.438	.6957	.554	.6979	.7266	.619	.6246	.6417	.7639	.6428
	20	.005	.6617	.7009	.6677	.5907	.7412	.6802	.6582	.6935	.5767	.6643
	21	.0025	.0833	.1522	.1286	.1467	.1087	.1667	.1159	.1558	.1594	.1177
	22	.0025	.0833	.148	.1667	.1006	.1306	.0882	.1304	.1363	.1575	.0889
	23	.0025	.0833	.1595	.1432	.0934	.1667	.1555	.1315	.1178	.1222	.127
	24	.0025	.0833	.1327	.1445	.1407	.1207	.122	.1524	.1667	.1489	.1569
	Vote	1	4.1493	9.1285	6.3352	8.2718	8.2676	7.0343	7.213	7.2625	6.4719	6.5038

Through experiment, we have mined the optimal ingredient pattern for the formula of the traditional Chinese medicine with the proportion 10/6 of Red Sage Root to Notoginseng, as shown in the last row of Table 6. From Table 6, we can also get the similar relation between the cardio-indexes about the curative effect of all test groups. The Euclidean distance between the cardio-indexes with the subscripts i and j is shown in the following Eq.(15) and the similar coefficient of the two cardio-indexes is calculated by the following Eq.(16).

$$D_{ij} = sqrt(\sum_{g=1}^{10}[v(g,i)-v(g,j)]^2) \tag{15}$$

$$R_{ij} = 1 - \frac{D_{ij} - \min\{D_{ij}\}}{\max\{D_{ij}\} - \min\{D_{ij}\}} \tag{16}$$

If the two cardio-indexes with the subscripts i and j meet the condition R_{ij}=1, we consider them the same on the curative effect. By calculation, the cardio-indexes with the subscripts 3,4,5,6 are the same and so are the cardio-indexes with the subscripts 21,22,23,24. We can only use one index to replace them respectively. From the medical view, the similar relation can be understood easily because the former similar relation denotes the similar effect of the indexes about anemia while the latter similar relation denotes the similar effect of the material indexes.

To evaluate the ingredient pattern mined by the method, we use some test samples to compare the curative effect of all test groups based on hypothesis testing and expert knowledge. Through evaluation, the percentage of

correctness reaches 100%.

5 Conclusions

In this paper, we present a method based on the feature extraction and voting algorithm for mining the optimal ingredient pattern about formula of traditional Chinese medicine. In this method, we first visualize the data with the curves from which we can extract some features from data about curative effect of formula of Chinese herbs in different sampling points. Then, we calculate all features for each test group and each cardio-index of each sample. Finally, we adopt the voting algorithm to mine the optimal ingredient pattern of Chinese herbs in which vote is held in three stages. In the first stage, a preliminary vote is held for each sample to vote the features of the given cardio-index and test group. In the second stage, a metaphase vote is held for each feature to vote the test group of the given cardio-index. In the third stage, a final vote is held for each cardio-index to vote the test group. From the result of the final vote, we can get the optimal pattern. Through experiment, the method is validated with the percentage of correctness 100% for mining the ingredient pattern.

Acknowledgement: We thank Professors Jiaxin Wang and Yannan Zhao at our laboratory and Professor Guoan Luo at the Chinese Medicine Laboratory of Tsinghua University for giving me the chance to investigate this problem and Professor Hongcai Shang at the Tianjin University of Traditional Chinese Medicine for giving the data.

References

[1] Li QY, Liu ZW, Jiang YN. Formulas of Traditional Chinese Medicine. Academy Press, 1998.

[2] He XD, Zou JH. The Chinese Materia Medica. Academy Press, 1998.

[3] Buono P, Costabile MF, Lisi FA. Supporting data analysis through visualizations. In: Proceedings of the International Workshop on Visual Data Mining, 2001. In conjunction with ECML/PKDD 2001−−the 2nd European Conference on Machine Learning (ECML'01) and the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD2001), 2001. IEEE Computer Society, 2001.

[4] Zaptron System, Inc. Optimization of New Drug Exploration from Herbs. http://www.zaptron.com/masterminer/solutions/bio/herb.htm.

[5] Sierra B, Lazkano E, Inza I, Merino M, Larranaga P, Quiroga J. Prototype election and feature subset selection by estimation of distribution algorithms (a case study in the survival of cirrhotic patients treated with TIPS). Artificial Intelligence in Medicine 2001. Lecture Notes in Computer Science 2101, 2001. 20~29.

[6] Shekhar S, Huang Y. Discovering spatial co-location patterns--a summary of results. In: Proceedings of the Seventh International Symposium on Spatial and Temporal Databases (SSTD), 2001. Lecture Notes in Computer Science 2121, 2001. 236~256.

[7] Bojarezuk CC, Lopes HS, Freitas AA. Data mining with constrained-syntax genetic programming−−applications in medical data set. In: Proceedings of the Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP 2001). IEEE Engineering in Medicine & Biology Magazine 2001.

[8] Lavrac N. Selected techniques for data mining in medicine. Artificial Intelligence in Medicine, 1999,16(1):3~23..

[9] Hardekopf B, Kwiat K, Upadhyaya S. A decentralized voting algorithm for increasing dependability in distributed systems. Joint Meeting of the 5th World Multi- Conference on Systemic, Cybernetics and Informatics (SCI2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS2001). 2001. http://www.iiis.org/sci/.

[10] Han JW, Kamber M. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, Inc., 2001.

[11] Conitzer V, Sandholm T. Vote elicitation: Complexity and strategy-proofness. Department of Computer Science, Carnegie Mellon University, National Conference on Artificial Intelligence (AAAI). 2002.

[12] James G. Majority vote classifiers: Theory and applications [Ph.D. Thesis]. Stanford University, 1998.