African Journal of Pharmacy and Pharmacology Vol. 6(26), pp. 1971-1976, 15 July, 2012
Available online at http://www.academicjournals.org/AJPP
DOI: 10.5897/AJPP12.493
ISSN 1996-0816 © 2012 Academic Journals

Full Length Research Paper

Statistical analysis of medical experiment data for discovering groups of correlated symptoms

Chenghe Shi[1#], Qingqiong Deng[2#], Peng Lu[3#], Minquan Zhou[2] and Gang Xiong[4]*

[1]Department of Traditional Chinese Medicine, Peking University Third Hospital, Beijing 100191, P. R. China.
[2]College of Information Science and Technology, Beijing Normal University, No. 19 Xin-Jie-Kou-Wai Street, Beijing 100875, P.R China.
[3]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China.
[4]Department of Physics, Beijing Normal University, No. 19 Xin-Jie-Kou-Wai Street, Beijing 100875, P. R. China.

Accepted 3 May, 2012

Three ways of statistical analysis, that is, Pearson correlation (PC), Spearman correlation (SC) and mutual information (MI), are applied on medical experiment data to obtain correlation matrix. When the results of this study's analysis were combined with those of professional analysis, it was found that the method based on MI may be the best way for discovering groups of correlated symptoms.

Key words: Association, symptom combination, coronary heart disease.

INTRODUCTION

Coronary heart disease (CHD) remains the single leading cause of death for adults worldwide (Lloyd-Jones et al., 1999). Effective prevention and therapy for CHD poses a major challenge to the entire medical community. Traditional Chinese medicine (TCM) has fought against CHD and its related diseases for more than 1000 years, and has accumulated thousands of prescriptions as well as clinical literatures (Chen and Jia, 2010; Guo et al., 2009; Liu et al., 2010). Therefore, more and more patients all over the world take TCM as a complementary and alternative avenue to treat CHD.

TCM is a medical system with at least 3000 years's uninterrupted clinical practice and it has the advantage of collecting macroscopic information (including symptoms, tongue and pulse recognition) of a patient for diagnosis, while syndrome is the core of diagnosis and target of herbal remedy in TCM. Nowadays, syndrome in TCM has been always studied in the context of a specific disease or biomedical condition and several literatures have demonstrated that syndromes are significantly associated with diseases (Chen and Jia, 2010, Guo et.al., 2009). TCM is taken by most people in China as a complementary therapeutic alternative since herbal remedies have the advantage over Western medicine in that it has less side effects and are less costly. TCM has always been regarded as a key component in 5000 years of Chinese civilization history. In ancient times before modern medicine was born, people all over the world mainly benefit from three traditional medicines, among which only TCM is still alive today; while Chaldaic and ancient Hindu medicines only have extremely rare documents as evidence that they ever existed in history. TCM, whose core is syndrome, is on the way to modernization. It is aiming to be accepted as a science, like Western medicine (Bohigas et al., 1984; Seba, 2003; Zhong et al., 1998; Zhong and Geisel, 1999).

Unstable angina (UA) is a type of CHD. It describes a biomedical condition that is intermediate between myocardial infarction (MI) and the more chronic state of stable angina. UA is now a heavy burden on the society and families in both industrialized and developing countries. So UA presents a better example and context for investigating diagnosis method and biological basis of syndromes in TCM.

The syndrome is the basic pathological unit and the key concept in TCM theory since herbal remedy is prescribed according to syndrome or syndromes a patient catches Li et.al., 2007).

Therefore, identification and determination of syndrome (s) in CHD patient become significantly important for TCM

*Corresponding author. E-mail: phgxiong@bnu.edu.cn.

[#]These authors contributed equally to this work.

Table 1. Clustering and validation results.

No.	Patterns	Syndrome diagnosed by experts	Corresponding treatment	Number of cases of patterns	The maximum Number of syndrome	Sensitivity of each pattern (%)
1	Chest pain, chest tightness, short breath, palpitation, hypodynamia, spontaneous perspiration	Qi deficiency and blood stasis	Tong Xin Luo capsule	560	500	89.29
2	Xerostomia, dizziness, amnesia; vértigo, tinnitus	Deficiency of Yin	Jisheng Shenqi pills	240	225	93.75
3	Sighing, depression, short and yellow urine, low speaking voice	Qi stagnation	Composite Salvia Dropping Pill	150	135	90
4	Inappetency, abdominal distension, stomach discomfort, eructation	Deficiency of spleen	Ren sen Jian Pi pill	108	96	88.89
5	Dry and hot face, swelling in the costal regions, light color in lips and nails	Yu re syndrome	Sijunzi decoction	54	48	88.89
6	Night sweat, feverishness in palms and soles;	Deficiency of Yin	Pingwei powder	124	116	93.55
7	Fear of cold and cold limbs, insomnia; lumbago	Deficiency of Yang (Subtype)	Sijunzi decoction	206	192	93.20

physicians. Nevertheless, there are few documents dedicated to this issue.

In this paper, we carried out a clinical epidemiology survey and we proposed a novel unsupervised data mining model, in which we treat mutual information (MI) as an association measure of two variables. In our effort, we tried to discover syndromes in CHD data and clinically verify these syndromes to test the performance of our model without supervision. Based on revised MI, we proposed an unsupervised pattern discovery algorithm to self-organize allocate significantly associated symptoms to patterns. By using diagnostic data, each pattern is verified to have clinical meaning.

MATERIALS AND METHODS

Study population

Syndrome is diagnosed according to symptom combinations. As shown in Table 1, we choose 72 symptoms that are closely related to CHD. The pulse information of every patient was not included for its bad consistency during the process of survey. In the survey, the data set was recruited from six clinical centers located in six provinces from the same demographic area and at the same time from October, 2005 to March, 2006, where a total of 1069 patients who suffer from CHD were surveyed. Eligible patients in this paper were defined as with UA based on diagnosis criteria of UA established in 2002 by ACC (American College of Cardiology) and AHA (American Heart Association), that is, chest pain at rest and transient S-T segment changes, without significant increases in creatine kinase and creatine kinase MB fraction (Zhong et al., 1998).

The criterion for enrollment was admission within 48 h after the onset of chest pain. Moreover, the exclusion criteria were composed of four conditions: (1) Besides UA, a patient also suffers from other cordis diseases such as acute myocardial infarction, myocarditis and cardiac nerve functional disease; (2) A patient with angina caused by other diseases, for example, rheumatic fever, syphilis, congenital coronary anomalies, hypertrophic cardiomyopathy and cardiac mitral stenosis; (3) Besides UA, a patient also suffers from stroke, diabetes, nephritis, renal failure, pulmonary infection, urinary tract infection,

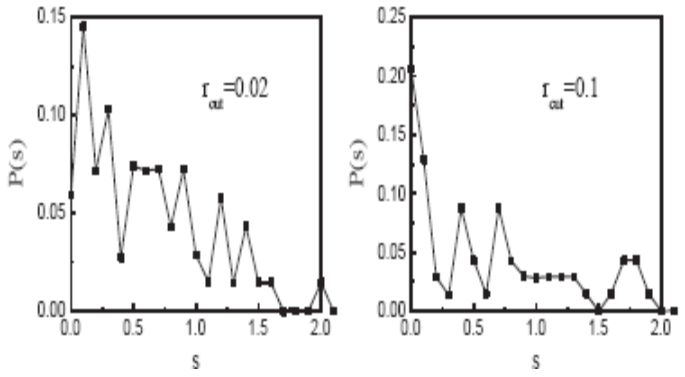

Figure 1. Level-spacing distribution function P(s) of MI matrix with cut-off value rcut =0.02 and rcut =0.1.

rheumatism, osteoarthritis and serious diseases caused by liver, renal, haematogenous system and incretion system; (4) A woman patient in gestation or lactation.

Every case is with 72 symptoms, together with the basic information of each subject. The frequencies of 72 symptoms are shown in Table 1; each variable (symptom) has four categories, that is, none, light, middle and severe, represented by 0, 1, 2, 3, respectively. The latter three categories of each variable means that the symptom has appeared and then separated into light, middle and severe by clinical doctors, who are strictly and uniformly trained to reach a high consistency.

All subjects gave informed consent and were approved by the Medical Ethics and Human Clinical Trial Committee at Guanganmen Hospital.

Data mining methods

It is essential to develop powerful computational methods to extract as much information as possible from medical experiment data.

In this paper, the data we dealt with are 70 symptoms of 1,070 patients. What we did was to find a way based on correlation analysis to category these symptoms into groups. We used three widely used statistical methods: Pearson correlation (PC), Spearman correlation (SC) and MI. The standard PC coefficient between two symptoms *xi* and *xj* is defined as:

$$PC(x_i, x_j) = \frac{1}{N}\sum_{k=1}^{N}\frac{x_{i,k} - M_{x_i}}{\sigma_{x_i}}\frac{x_{j,k} - M_{x_j}}{\sigma_{x_j}}. \quad (1)$$

The standard SC coefficient between two symptoms *xi* and *xj* is defined as:

$$SC(x_i, x_j) = \frac{1}{N}\sum_{k=1}^{N}\frac{R(x_{i,k}) - M_{R(x_i)}}{\sigma_{o(x_i)}}\frac{R(x_{j,k}) - M_{R(x_j)}}{\sigma_{R(x_j)}}. \quad (2)$$

with *R*(*xi,k*) the rank of *xi,k*. The MI coefficient between two symptoms *xi* and *xj* is defined as:

$$MI(x_i, x_j) = H(x_i) + H(x_j) - H(x_i, x_j) \quad (3)$$

Where *H*(*xi*) and *H*(*xj*) are information entropies of distributions of each symptom, and *H*(*xi,xj*) is information entropy of joint distribution of the two symptoms. Using PC, SC and MI coefficients, we can construct three matrices of rank 70 from experiment data, respectively. The non-diagonal elements of these matrices give information of correlations among the 70 symptoms.

We shall analyze these matrices in the following way similar with Zhong et al. (1998). We set a cut-off value *rcut* that non-diagonal elements with its absolute value smaller than *rcut* are neglected and set to be zero. At the starting point *rcut* = 0, we expect strong correlations among the 70 symptoms. By tuning *rcut* from zero to the maximum value, the correlation induced by non-diagonal elements is expected to decrease and the 70 symptoms should be separated into uncorrelated groups. If we consider the symptoms as sites forming a random network and the non-diagonal elements as existence probability of bonds connecting those sites of that random network, the procedure of decreasing non-diagonal elements is similar with the bond-percolation transition of a random network. At the beginning *rcut* = 0 the network forms a wholly percolated cluster. With increasing *rcut*, the network begins to split into separate parts and each part corresponds to a group of correlated symptoms.

The stated percolation transition can be studied by considering eigenvalue spacing distribution of the real symmetrical matrices obtained from the network of symptoms. According to the random matrix theory (RMT) (Wigner, 1967), eigenvalue spacing distributions of real symmetrical random matrices fall into two universal types depending on the strength of eigenvalue correlation induced by non-diagonal elements (Hofstetter and Schreiber, 1993). Strong correlation between eigenvalues leads to the so-called Gaussian orthogonal ensemble (GOE) while weak correlation between eigenvalues leads to Poisson ensemble (PE) (Plerou et.al., 1999). The nearest neighbor spacing distribution (NNSD) of eigenvalues in the case of GOE is close to the Wigner-Dyson distribution (Wigner, 1967).

$$P_{GOE}(s) \approx \frac{1}{2}\pi s \exp -\pi s^2/4 \quad (4)$$

While the distribution in the case of PE is the Poisson distribution

$$P_{PE}(s) = \exp(-s). \quad (5)$$

It should be noted that the obtained result of RMT holds only for statistical results of large number of eigenvalue spacings, while in the case, we study the rank of obtained matrix and it is 70 and not large. Thus statistical fluctuation is expected strong in NNSD. However, we still decide to have a try. We obtain the eigenvalues *Ei* of the matrices and line them in order. Then we unfold *Ei* by dividing them with their average value $M_E = \frac{E_N - E_1}{N}$ and obtain the normalized eigenvalues *ei* = *Ei/ME*. Thus, the eigenvalue spacings are obtained as *si* = *ei*+1-*ei*.

RESULTS AND DISCUSSION

Figure 1 shows numerical NNSD *P* (*s*) obtained from the MI correlation matrix with cut-off value *rcut* = 0, *rcut* =0.02 and *rcut* =0.1, respectively. We can see that *P* (*s*) with *rcut* =0 and *rcut* =0.02 show a peak around *s* =0.1 which is similar with the behavior of PGOE.

However, *P* (*s*) with *rcut* =0.1 is monotonically decreasing with increasing *s* similar to PPE. Figures 2 and 3 are NNSD obtained from correlation matrices of PE and SP. One can see that with increasing *rcut*, they also show

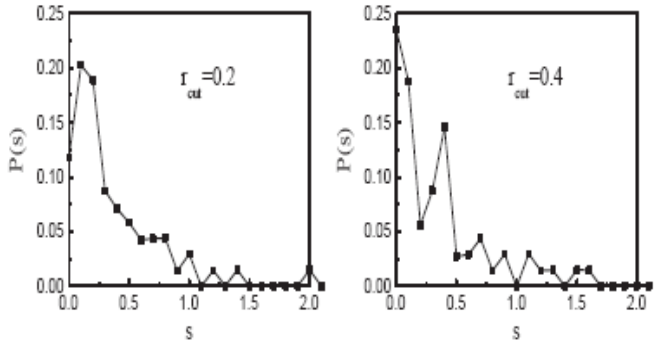

Figure 2. Level-spacing distribution function P(s) of SC matrix with cut-off value rcut =0.2 and rcut =0.4.

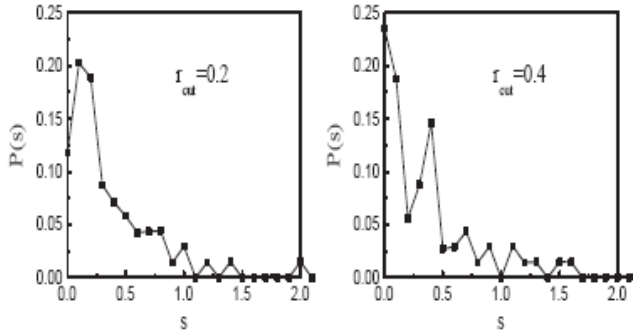

Figure 3. Level-spacing distribution function P(s) of SC matrix with cut-off value *rcut* =0.2 and *rcut* =0.4.

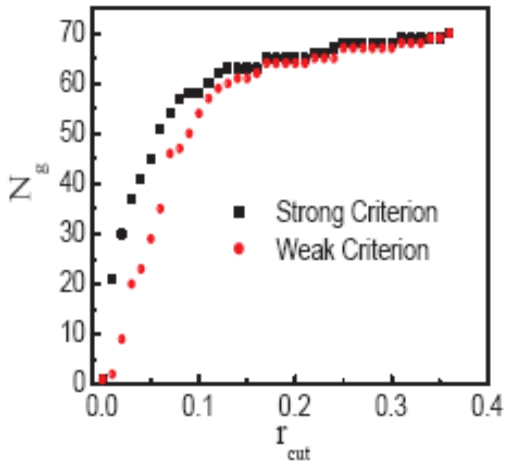

Figure 4. Number of group *Ng* versus cut-off value *rcut* by MI analysis. Black and red dots correspond to strong and weak criterion, respectively.

similar behavior as that of MI. Therefore, although the number of eigenvalues in our case is not large enough to reach the thermodynamic limit, one is still able to see a transition from a typical GOE behavior to a typical PE with increasing *rcut*. This means that with increasing *rcut* the

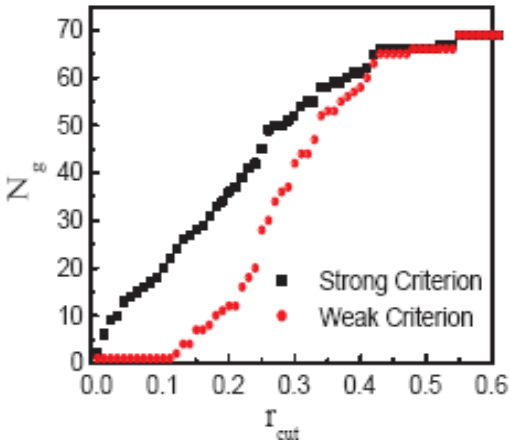

Figure 5. Number of group *Ng* versus cut-off value *rcut* by SC analysis. Black and red dots correspond to strong and weak criterion, respectively.

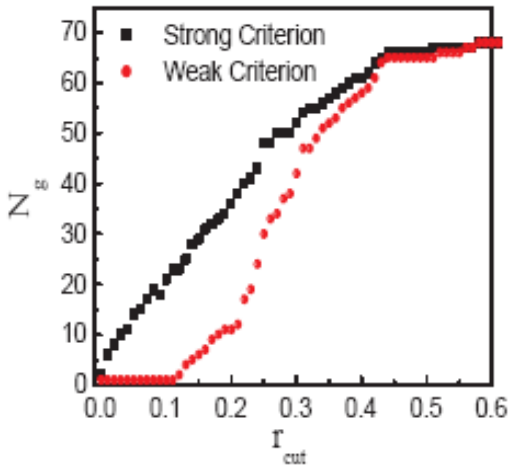

Figure 6. Number of group *Ng* versus cut-off value *rcut* by PC analysis. Black and red dots correspond to strong and weak criterion, respectively.

finite network of correlations among the 70 symptoms also encounters a percolation transition, turning from a connected network into separate groups, which is just what we expect to see.

Considering the structure of the correlation network of the symptoms, since the number of symptoms is not too large, we can directly consider the number of correlated groups the symptoms forms with increasing *rcut*. We used two different standards to determine whether a symptom shall be added into a group. One is a weak criterion that the symptom is added into a group if at least one symptom in that group has non-zero correlation coefficient with it. The other is a strong criterion that the symptom is added into a group only if every symptom in that group has non-zero correlation coefficient with it. Figures 4, 5 and 6 show the curve of group number *Ng* versus *rcup*.

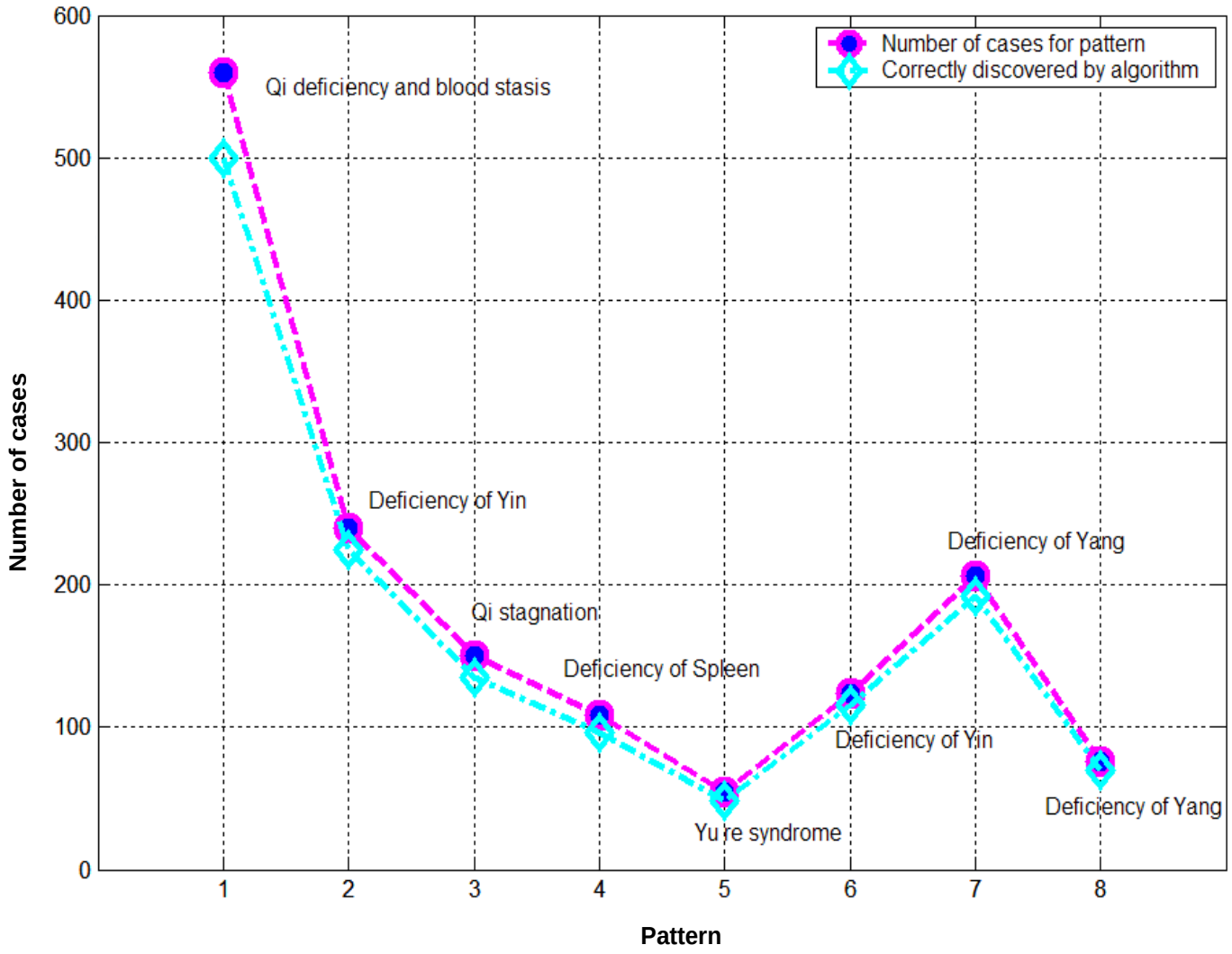

Figure 7. X-axis is pattern; Y-axis is corresponding number of cases in the data. For purple circle point, the data is symptoms data, while for green diamond point, the data is syndrome data. The percentage around each point indicates the sensitivity of the corresponding pattern. The average of them is sensitivity of the algorithm -96.48%.

One can see that the curves of the two criterions by MI are quite similar, while the curves of the weak criterion by PE and SP show that the network is not split into groups at smaller value of *rcut*. Therefore, we may conclude that MI analysis is easier to obtain the form of uncorrelated groups.

Finally we evaluate clinical meaning of the clusters by two avenues.

The first avenue is using clinical theory of TCM and expert domain knowledge to estimate pattern results. As shown in Figure 7, the pattern results recognized by MI analysis (green diamond points) are highly accordant with the results diagnosed by TCM physicians (purple circle point).

The second avenue is to objectively estimate each pattern discovered by above algorithm using a supervised validation method of the following three steps.

Step 1. Each pattern S is returned to the unsupervised data, if all variables of the pattern simultaneously appear (their values are non-zero) on a patient, then serial number of the patient is recorded as shown in Table 1. All serial numbers are stored in a vector with Ls dimensions denoted as Vs.

Step 2. We track the vectors back to the syndrome data by adding up the vectors one by one to generate a new vector. The maximal number in the new vector and the corresponding syndrome are stored as shown in Table 1.

Step 3. The accuracy of a pattern is defined as the ratio of maximal number and total number, and the accuracy of the algorithm is determined by averaging the accuracy of all patterns. We found that the pattern results of MI analysis reached a high accuracy as depicted in Figure 8. The optimal rcut is 0.04. and the corresponding accuracy is 91.04%.

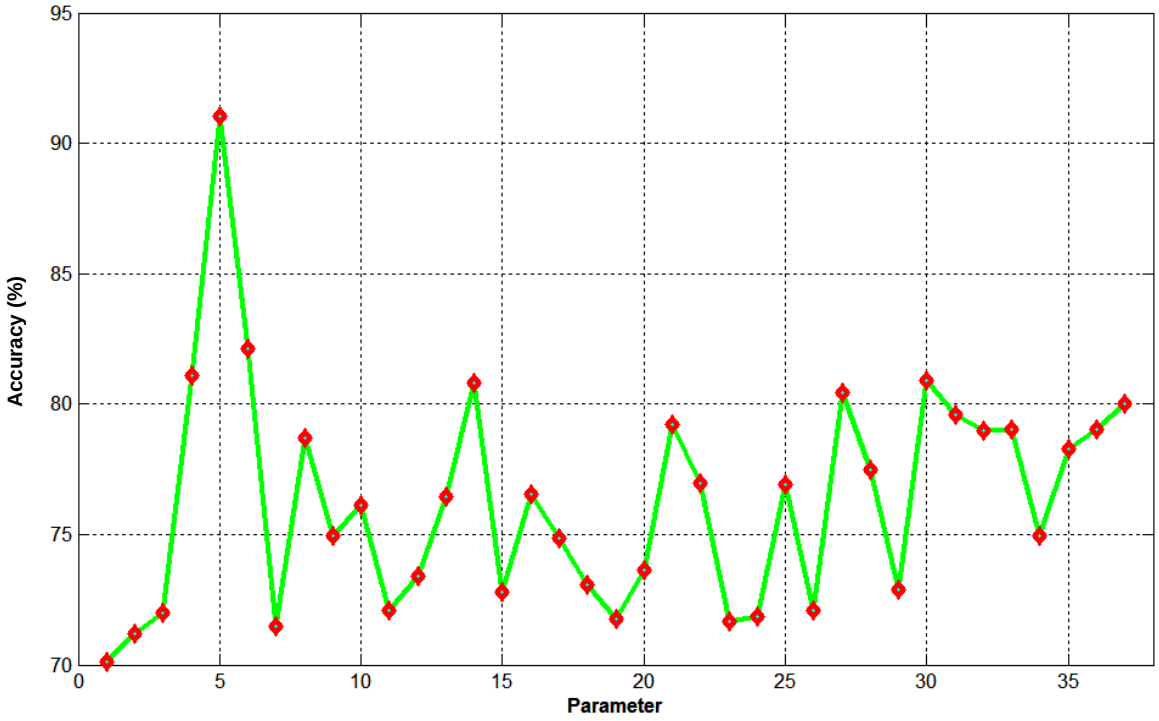

Figure 8. The optimal selection step (fourth step).

Conclusion

In summary, we present a kind of correlation based cluster algorithm to cluster symptoms in medical data. We also propose a supervised validation algorithm to objectively evaluate the clustering results. We find that MI is better fit to describe correlation between symptoms in the data. The corresponding clustering results reached an accuracy of 91.04%. The algorithm presented here paves a significant basis for discover new pattern in medical data.

ACKNOWLEDGEMENT

Authors acknowledge the support of CNSF under Grant No. 10604006 and SCC of Beijing Normal University.

REFERENCES

Bohigas O, Giannoni MJ, Schmit C (1984). Characterization of Chaotic Quantum Spectra and Universality of Level Fluctuation Laws. Phys. Rev. Lett. 52:1.

Chen JX, Jia ZH (2010). Selecting biomarkers for primary hyperlipidemia and unstable angina in the context of neuro-endocrine-immune network by feature selection methods. J. Biol. Syst. 18:605-619.

Guo SZ, Chen JX, Zhao HH, Wang W, Yi JQ, Liu L (2009). Building and evaluating an animal model for syndrome in Traditional Chinese Medicine in the context of Unstable Angina (myocardial ischemia) by supervised data mining approaches. J. Biol. Syst.17: 531-546.

Hofstetter E, Schreiber M (1993). Statistical properties of the eigenvalue spectrum of the three-dimensional Anderson Hamiltonian. Phys. Rev. B48:16979.

Lloyd-Jones DM, Larson MG, Levy D (1999). Lifetime risk of developing coronary artery disease. Lancet, 353: 89-92.

Li S, Zhang ZQ, Wu LJ, Zhang XG, Li YD, Wang YY (2007). Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network, IET Syst. Biol. 1(1):51-60.

Liu GP, Li GZ, Wang YL, Wang YQ (2010). Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning. BMC Complement Altern. Med. 10:37-48.

Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Stanley HE (1999). Universal and Nonuniversal Properties of Cross Correlations in Financial Time Series. Phys. Rev. Lett. 83:1471.

Seba P (2003). Random Matrix Analysis of Human EEG Data. Phys. Rev. Lett. 91:198-204.

Wigner EP (1967). Random Matrices in Physics. SIAM Rev. 9:91.

Zhong JX, Geisel T (1999). Level fluctuations in quantum systems with multifractal eigenstates. Phys. Rev. E59:4071.

Zhong JX, Grimm U, Romer RA, Schreiber M (1998). Level-Spacing. Distributions of Planar Quasiperiodic Tight-Binding Models. Phys. Rev. Lett. 80:3996.