A Data Mining Model to Predict and Analyze the Events Related to Coronary Heart Disease using Decision Trees with Particle Swarm Optimization for Feature Selection

A. Sheik Abdullah
PG Scholar
Department of Computer Science Engineering
Kongu Engineering College

ABSTRACT

Coronary Heart Disease (CHD) is a most common type of coronary disease which has no clear origin and a significant basis for premature death. Data mining has become an essential methodology for applications in medical informatics and discovering various types of diseases and syndromes. Mining valuable information and providing systematic decision-making for the diagnosis and treatment of disease from the entire database progressively becomes necessary. Classification in data mining performs an important role in data analysis and prediction. The objective of this work is to build a data mining model to be used by physicians and also to associate the risk factors related to heart disease. Data mining model has been developed using PSO – C4.5 algorithm. The proposed model yields reduced set of features using the feature selection algorithm along with improved prediction accuracy. Thereby the developed model can be successfully used in predicting other metabolic syndromes.

General Terms

Data mining, Data Classification, Feature Selection.

Keywords

Coronary Heart Disease, Decision Trees, Particle Swarm Optimization.

1. INTRODUCTION

Data mining is concerned with systematized mining of hidden predictive information from large sort of database and repositories. From the point of medical science and bioinformatics, data mining has been involved in discovering different sorts of life threatening diseases and syndromes. There by. The burden of the healthcare industry can be reduced by the models and methods used up in data mining and medical informatics disciplines.

Coronary Heart Disease (CHD) is the most important disease affecting the heart and predominant cause of death across the world. While the indications and signs of coronary heart disease are distinguished in the advanced state of disease, most persons with coronary heart disease show no indication of disease for decades as the disease develops in advance the first inception of symptoms, habitually a "quick" nervous breakdown, finally arises. Later on, the blood flow to The heart muscle reduces according to current trends in India, in future half a healthy 40-year-old men and one in three healthy 40-year-old women will progress CHD in the future. Even though various sorts of diagnosis are in practice analysing and predicting the cause and probable reason for the occurrence of the disease are still lagging. But the methods and the technologies incorporated with data mining can be indulged in discovering and identifying the probable reasons in predicting diseases and syndromes.

Some of the work has been progressed with the implementation of data mining algorithm such as K-NN, ID3, Naïve bayes, and Apriori algorithm. In CHD, if the risk factors are foreseen in earlier stage two kinds of difficulties can be solved. First, several medical treatments such as angioplasty, coronary artery bypass and heart transplant can be avoided. Second, the cost associated with the disease can be reduced. This paper focuses on the creation of a data mining model using Particle Swarm Optimization with Decision Tree classification algorithm for evaluating and predicting various events related to CHD with reduced set of features.

2. LITERATURE SURVEY

Tsien et al [1] in their study indicated that classification trees, which have certain advantages over logistic regression, with patients having Myocardial Infarction (MI). The results have shown that the existence of MI has been noticed in male than in female. Age, Blood Pressure, Smoking has found to be the important risk factor in the patients with MI. The methodology can be extended by considering other events such as Angina, Angioplasty and Coronary Artery Bypass Graft surgery.

Hlimonenko et al [2] in their study assessed the Pulse Wave Velocity (PWV) and Augmentation Index in different arteries in patients with severe CHD. Signal measurements were obtained from 28 subjects. It was found that aortic PWV was significantly increased in the CHD group when compared with that in the control group ($P<0.01$). This study shows the strong association of aortic stiffness and atherosclerosis. The major finding of this study conveys that patients with coronary artery disease have increased aortic PWV when compared with control subjects. Control group subjects had significantly less arterial stiffness (as reflected by both a lower PWV and a lower Augmentation Index). Not significant but still apparent difference between pulse wave velocity in peripheral arteries in patients with CHD compared with control group indicates that changes of elastic properties of the walls of peripheral arteries are less obvious or peripheral arteries undergo changes on the latest stages. To obtain significant difference greater number of patients should be investigated.

The euro aspire III survey [3] was carried out in 76 centers from selected geographic areas in 22 countries. The main objective of euro aspire III were to determine whether the Joint European Guidelines on CVD prevention are being followed in patients with CHD and whether the practice of preventive cardiology in patients with established coronary disease in Euro aspire III has improved, by comparison with those centers that took part in Euro aspire I and II [4][5]. Consecutive patients, with a clinical diagnosis of CHD, were identified and then followed up, interviewed and examined. With the existence of coronary events 13,935 medical records were reviewed and 8966 patients were interviewed (27% women). In an interview, 17% of patients smoked cigarettes, 35% were obese, 56% had blood pressure beyond the normal level, 51% had serum total cholesterol, 10% had fasting plasma glucose less than 6.1 mmol/I. This survey shows that large proportions of coronary patients do not achieve the lifestyle, risk factor and therapeutic targets for cardiovascular disease prevention.

The results of the Euro aspire III survey show that large proportions of coronary patients in Europe do not achieve the lifestyle, risk factor and therapeutic targets set by the Joint European Societies' guidelines on CVD prevention in clinical practice. There is considerable variation between European countries in patients' lifestyle, risk factor prevalence's and use of cardio protective medication. Information on risk factor history and measurements in the discharge documents is incomplete. There is strong scientific evidence that lifestyle modifications in relation to tobacco smoking, diet and physical activity can reduce the risk of recurrent cardiovascular events in patients with established coronary disease and improve survival. It was concluded that wide variations exist between 15 countries in the risk factor prevalence's and the use of cardio protective drug therapies. Also, there is still considerable potential throughout Europe to raise standards of preventive care in order to reduce the risk of recurrent disease and death in patients.

Karaolis et al [6] developed a data mining system for the assessment of heart related risk factors using Association analysis based on Apriori algorithm. A total of 369 cases have been collected from a hospital and three sorts of events have been investigated such as Myocardial Infarction (MI), Percutaneous Coronary Intervention (PCI) and Coronary Artery Graft Bypass Surgery (CABG). The models are evaluated using 13 attributes and the results shown that smoking is one of the main risk factor that directly affect the coronary heart disease for all of the events. The work can be extended by some of the additional risk factors related to coronary heart disease with other such data mining techniques.

Karlberg and Elo [7] calculated the burden of Ischemic Heart Disease (IHD) and coronary risk factors in a defined population using data from all public providers of healthcare. Calculation of the actual burden of disease in the population showed that when hospital discharge data were combined with the outpatient data, there were no or slight difference in the age-specific rates of Acute Myocardial Infarction (AMI), while the rates of angina were between two-fold and four-fold higher, and unspecified IHD was between three-fold and ten-fold higher in individuals aged greater than 50 years compared with using hospital discharge data alone. These findings suggest that hospital discharge data should be combined with outpatient care data to provide a more comprehensive estimate of the burden of IHD and its risk factors.

Kunc [8] presented simulation results which can be used for evaluation of patients with coronary heart disease, congestive heart failure, end-stage renal disease in Slovenia. At the same time also year treatment costs were calculated regarding each of observed diseases. The presented results enable the estimation of potential savings resulting from more intensive chronic diseases treatment.

Ping et al [9] made a study with the clinical cases of famous doctors' diagnosing and treating on Coronary Heart Disease (CHD). This study explored the clinical structural relationships between clinic information sub-system of symptom, syndrome and medicine. The point-wise mutual information method of information theory is applied for data mining and relevant association analysis of symptom-syndrome and syndrome-medicine has been discovered. The effects related to CHD such as spleen and stomach problems are identified with specific sort of records. Traditional Chinese Medicine (TCM) has made great contributions to the health care of Chinese nations over thousands of years. CHD is in the category of chest stuffiness in TCM which is proposed by Zhang and Wang [10]. But TCM can't able to resolve the clinical problems associated with CHD. So the interesting and clinically meaningful regularities of the Chinese nation are made with the cases by segregating it into three forms such as symptom, syndrome and drug; then, the algorithm is applied with mutual information parameters, to discover knowledge and seek the regularities in the correlation of the three subsystems so as to explore a new approach for solving the clinic problems of TCM.

3. MATERIALS AND METHODS

3.1 Data Collection

Data records of various CHD patients have been collected from the UCI Machine Learning Repository which corresponds to Cleveland clinical foundation. The collected dataset contains 303 records and corresponds to one of the following event such as Angina, Acute Myocardial Infarction (AMI), Percutaneous Coronary Intervention (PCI), Percutaneous Tran luminal Coronary Angioplasty (PTCA), and Coronary Artery Bypass Graft (CABG). The retrieved record contains 13 attributes as described in table 1 with respect to each patient.

Table 1. Description of Risk factors in Benchmark Dataset

S.no	Risk factor	Description
1.	Age	Age in years
2.	Sex	Sex (1 = male; 0 = female)
3.	CP Type	Chest pain type
4.	Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
5.	Cholesterol	Serum cholesterol in mg/dl

6.	Fasting Blood Sugar	Fasting blood sugar > 120 mg/dl
7.	Restecg	Resting electrocardiographic results
8.	Thalach	Maximum heart rate achieved
9.	Exang	Exercise induced angina
10.	Old peak	ST depression induced by exercise relative to rest
11.	Slope	The slope of the peak exercise ST segment
12.	Ca	Number of major vessels (0-3) colored by fluoroscopy
13.	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect

In addition to the benchmark dataset, records of various CHD patients with various events have been collected from a hospital. The collected dataset contains 306 records and corresponds to one of the following event such as Angina, Acute Myocardial Infarction (AMI), Percutaneous Coronary Intervention (PCI), and Coronary Artery Bypass Graft (CABG). The retrieved record contains 24 attributes as described in table 2 with respect to each patient.

Table 2. Description of Risk factor in Real dataset

S.no	Risk factor	Description
1.	Age	Age in years
2.	Sex	Sex (1 = male; 0 = female)
3.	CP Type	Chest pain type
4.	Systolic BP	Systolic blood pressure in mm/Hg
5.	Diastolic BP	Diastolic blood pressure in mm/Hg
6.	Serum cholesterol	Serum cholesterol in mg/dl
7.	Fasting Blood Sugar	Fasting blood sugar > 120 mg/dl
8.	Restecg	Resting electrocardiographic results
9.	Weight	Weight of the corresponding patient in Kikograms
10.	Waist Circumference	Male >40 inches and female>35 inches
11.	Smoking	If yes = 1 and no = 0
12.	Hypertension	Hypertension or high blood pressure is a cardiac chronic medical condition
13.	Hypercholesterolemia	Hypercholesterolemia is the presence of high levels of cholesterol in the blood
14.	Previous Angina	Previous angina occurs suddenly, often at rest or with minimal degrees of exertion
15.	Prior Stroke	This can be due to ischemia (lack of blood flow) caused by blockage (thrombosis, arterial embolism), or a hemorrhage (leakage of blood)
16.	Antero Lateral	Determines the existence of disease in the regions of the heart
17.	Antero Septal	Determines the existence of disease in the regions of the heart
18.	Infero Septal	Determines the existence of disease in the regions of the heart
19.	Infero Lateral	Determines the existence of disease in the regions of the heart
20.	Septo Anterio	Determines the existence of disease in the regions of the heart
21.	Diabetes	If present value = 1 else value = 0
22.	Obesity	If present value = 1 else value = 0
23.	Family History	If present value = 1 else value = 0
24.	Pericardial Effusion	fluid around the heart

3.2 Data Pre-processing

The collected data is cleaned to remove noise and inconsistency. Duplications among the data are extracted, data relevant to the analysis task are retrieved from the database, and finally data are transformed or consolidated into appropriate form for mining.

After all the above preprocessing steps, the number of cases for each sort of event such as AMI, PTCA,CABG should be structured accordingly, which is suitable for the evaluation of risk factors related to CHD.

3.3 Data Classification

Particle Swarm Optimization (PSO) is a robust and effective optimization technique based on the movement and intelligence of swarms [11]. PSO relates the idea of social relations to problem solving. It involves a number of particles that establish a swarm moving around the exploration space looking for the best solution. In this optimization technique each of the particles is considered as a point in an N-dimensional space which regulates its flying with its own involvement and also the experience gained from the other particles in the search space. Each particle retains a path of its coordinates in the search space which has been related with the best solution (fitness) that has reached so far by that particle. This value is called personal best, **pbest**. Additional best value that is traced by the particle which is the value acquired so far by any of the particle in the locality of that particle. This value is called **gbest**. The fitness value for each of the particle has been evaluated by means of J48 classification algorithm at each of the iteration step for each particle in the search space until the stagnation point has been attained. The selected feature set is represented in the form of,

f_1	f_2	f_3	f_4	…….	………	f_n

PSEUDOCODE PSO – J48

```
begin
    for i=1 to number of particles
    Randomly initialize particle position and velocity
    end for
    do
        for i=1 to number of particles
        Calculate fitness value of particle swarm
         By J48 ()
            if fitness value is better than pbest
             In history set current value as new pbest
            end if
        end for
        Choose the particle with best fitness value of
      All particles as gbest
        for i=1 to number of particles
          The new velocity of particle i at iteration k
          is calculated as,
          $V_i^{k+1} = wV_i^k + c_1 rand_1(\ldots) x (pbest_i-P_i^k) +$
          $c_2 rand_2(\ldots) x (gbest-P_i^k)$
          The position of the particle i at iteration k
          is updated as,
          $P_i^{k+1} = P_i^{k} + V_i^{k+1}$
        end for
     while (stopping criterion is not met)
end
```

The modification of the particle's Velocity equation is described as follows,

$$V_i^{k+1} = W V_i^k + c_1 rand_1(\ldots) x (pbest_i - P_i^k) + c_2 rand_2(\ldots) x (gbest_i - P_i^k)$$

Where,

V_i^k : initial velocity of the particle I at iteration k

P_i^k : current position of the particle I at iteration k

W : weighting factor

$rand_1$: uniformly distributed random number between 0 & 1

$rand_2$: uniformly distributed random number between 0 & 1

C_1 : cognition learning factor

C_2 : social learning factor

$pbest_i$: pbest of agent i

$gbest_i$: gbest of the group

The updation of the particle position equation is described as,

$$P_i^{k+1} = P_i^{k} + V_i^{k+1}$$

Where,

P_i^{k} : current position of the particle I at iteration k

V_i^{k+1} : The updated velocity of the particle I at iteration k

```
begin
    for d=1 to number of training observations and
    its class values
            for a=1 to number of candidate attributes
                Select a splitting criterion
            end for
    end for
    Create a node $N_d$
    if all observations in the training dataset have
    the same class output        value C, then

       return $N_d$ as a leaf node 52abelled with C.
       if attribute list = {∅}, then
  a.   return $N_d$ as a leaf node 52abelled with majority class
       output value.
  b.   Apply selected splitting criterion
  c.   Label node $N_d$ with the splitting criterion attribute.
  d.   Remove the splitting criterion attribute from the
       attribute list.
            for each value i in the splitting

               $D_i$ = no. Of  Observations in
            training dataset satisfying attribute value i.
               if $D_i$ is empty  then
                  attach a leaf node labeled
                 with majority class
                  output value to node $N_d$.
              else
                  attach the node returned by
    decision tree to node $N_d$.
               end if
            end for
         return node $N_d$
       end if
    end if
    for i=1 to number of training tuples (N)
    if $class_i$ = predicted $class_i$ of testing data then
        if $class_i$ = class label of positive tuples then
            TP= TP+1
        else  if $class_i$ = class label of negative tuples then
            TN=TN+1
        end if
    end if
    end for
    fitness value = (TP+TN / N)
end
```

4. PERFORMANCE METRICS

The following are the performance metrics used up for the evaluation of the model,

1. Accuracy - It refers to the total number of records that are correctly classified by the classifier.

2. True Positive Rate (TP) : It corresponds to the number of positive examples that have been correctly predicted by the classification model.
3. False Positive Rate (FP) : It corresponds to the number of negative examples that have been wrongly predicted by the classification model.
4. True Negative Rate (%TN) : It corresponds to the number of negative examples that have been correctly by the classification model.
5. False Negative Rate (%FN) : It corresponds to the number of positive examples that have been predicted as negative by the classification model.
6. Precision – It refers to what percentage of positive prediction were correct
7. Recall – It refers to what percentage of positive cases were caught.
8. Kappa Statistics - A measure of the degree of non-random arrangement between same categorical values of a variable.

5. PARAMETER SETTINGS

In our experiment, the parameters for the feature selection algorithm are set as: population size = 10, weighting factor w= 0.9, cognition learning factor = 1.4, social learning factor = 1.4 and the number of generations is based upon the stopping criterion.

6. EXPERIMENTAL RESULTS & DISCUSSION

To measure the effectiveness of the approach experiments have been conducted using the UCI machine learning dataset and the real dataset. The proposed work mainly concerns with the development of a data mining model with the PSO – J48 algorithm. The interpretation of feature selection mechanism for the benchmark and real dataset has resulted in reduced set of features.

For the benchmark dataset [12], all the 13 features has been considered for evaluation using PSO – J48 algorithm and when the stopping criterion is met the risk factors has been reduced to 9 such as age, sex, chest pain type, serum cholesterol, fasting blood sugar, restecg, thalach, ca and thal with an improved accuracy of about 60.74%. The following figure 1 depicts the number of iterations performed using the PSO – J48 algorithm with respect to accuracy obtained for the UCI machine learning dataset.

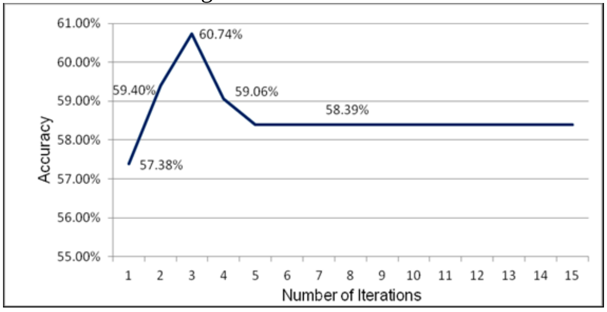

Fig 1: Iterations performed for UCI Repository Dataset

While considering the real dataset, initially 24 features has been considered for evaluation using the PSO – J48 algorithm when the stopping criterion is met the risk factors has been reduced to 20 such as age, sex, chest pain type, serum cholesterol, fasting blood sugar, restecg, weight, smoking, hypercholesterolemia, previous angina, prior stroke, antero lateral, antero septal, infero septal, infero lateral, septo anterio, obesity, family history, and pericardial effusion with an improved accuracy of about 55.23%. The following figure 2 depicts the number of iterations performed using the PSO – J48 algorithm with respect to accuracy obtained for the real dataset.

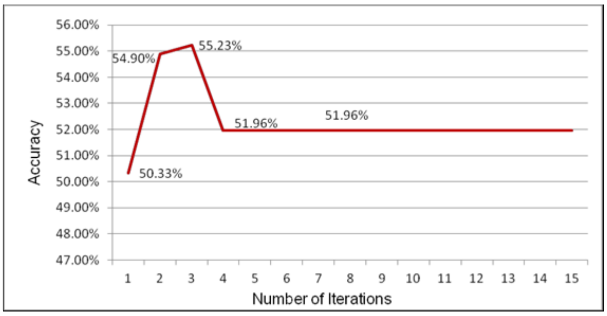

Fig 2: Iterations Performed for the Real dataset

The effectiveness of PSO optimization algorithm has been compared with other such optimization techniques such as stepwise forward selection, stepwise backward elimination and the combination of the both techniques. The working of PSO algorithm is analogy to the flock of bird and it is an evolutionary based parallel search algorithm. In this algorithm each particle or agent searches based on its own experience and its neighbor’s experience. It is a derivative free technique and less sensitive to the nature of the objective function and the parameters involved. Thereby it generates high quality solutions with diminutive calculation time and constant convergence characteristics.

Stepwise forward selection and backward elimination algorithm are the heuristic approaches as given by Han and Kamber [13]. In this algorithm, the features once selected cannot be discarded or re-selected. Forward selection initiates with an empty set of features and progressively adds features. Backward elimination starts with all the features and progressively eliminates some. Both of the algorithms generate nested subsets of features. From the point of statistical convolution, they are alike. When a stopping condition is met, forward selection is computationally more proficient because the trainings are executed on smaller subsets of features.

Hence forward selection works well for small set of features, but not for larger ones. On the other hand, backward elimination executes well for larger subsets of data but not for smaller ones.

For the benchmark dataset, the evaluation results with PSO – J48 with respect to other such optimization algorithm has been discussed in table 3 in accordance with the accuracy obtained and the other statistical measures. The prediction accuracy of stepwise forward selection and the combination of both forward selection and backward elimination is of same which is of 60.40% but stepwise backward elimination provides only 57.72% which is lower than the other such algorithms and also 10 features has been selected. Meanwhile, PSO – J48 provides a predicted accuracy of about 60.74% which is due to its evolutionary search in the entire population.

Table 3. Comparison of algorithms for benchmark dataset

Algorithm	Accuracy	Precision	Recall	Kappa Statistics	Features selected
PSO – J48	60.74%	34.64%	33.91%	0.302	9
Stepwise forward selection	60.40%	37.74%	35.38%	0.334	4
Stepwise backward elimination	57.72%	32.27%	32.17%	0.308	10
Combination of both forward selection and backward elimination	60.40%	37.85%	35.38%	0.332	4

For the real dataset, the prediction accuracy and the number of features selected for the stepwise forward selection and the combination of both forward selection and backward elimination algorithms are found to be same which is of 50.98% and 4. The stepwise backward elimination provides a predicted accuracy of about 54.90% with 21 features selected. But for the PSO – J48 optimization algorithm the prediction accuracy has found to be 55.23% with an increase in precision and recall rates as discussed in table 4.

Hence for both the benchmark and real dataset the PSO – J48 algorithm performs better than that of other such optimization algorithm due to its entire high-dimensional problem space. PSO won't use the gradient of the problem being optimized, so it does not have need of the optimization problem be degree of difference as required by classic optimization methods.

Table 4. Comparison of algorithms for Real dataset

Algorithm	Accuracy	Precision	Recall	Kappa Statistics	Features selected
PSO – J48	55.23%	55.61%	61.17%	0.372	20
Stepwise forward selection	50.98%	50.12%	45.65%	0.247	3
Stepwise backward elimination	54.90%	56.84%	57.12%	0.345	21
Combination of both forward selection and backward elimination	50.98%	50.12%	45.65%	0.247	3

7. CONCLUSION & FUTURE WORK

Data mining plays an important role in the identification and prediction of various sort of metabolic syndromes and hence various sorts of diseases can be discovered. The proposed work is mainly concerned with the development of a data mining model with the PSO – J48 algorithm. The developed model have the functionalities such as evaluating the risk factors related to CHD, predicting the occurrence of various events related to each patient record. The experimental results have shown that the prediction accuracy using PSO – J48 provides an improved result and reduced number of features than that of the other optimization algorithm. Thereby the proposed model can be greatly deployed in evaluating and predicting various other diseases and syndromes in medical informatics.

8. REFERENCES

[1] C.L. Tsien, H.S.F. Fraser, W.J. Long and R.L. Kennedy (2001) "Using classification trees and logistic regression methods to diagnose myocardial infarction" in Proc. 9th World Congr., Inf., vol. 52, pp. 483-497.

[2] I. Hlimonenko, K. Meigas, M. Viigimaa, K. Temitski, (2007) "Assessment of Pulse Wave Velocity and Augmentation Index in arteries in patients with severe coronary heart disease", International Conference of the IEEE EMBS Cite Internationale, Lyon, France, pp. 1703-1706.

[3] Euro aspire study group, (2009) "Euro aspire III: A survey on the lifestyle, risk factors and use of cardio protective drug therapies in coronary patients from 22 European countries," Eur. J. Cardiovascular. Prev., Vol. 16, No.2, pp. 121-137.

[4] EUROASPIRE Study Group, "EUROASPIRE. A European Society of Cardiology survey of secondary prevention of coronary heart disease" Principal results. Eur Heart J. 18: 1569–1582.

[5] EUROASPIRE Study Group, (2001) "Lifestyle and risk factor management and use of drug therapies in coronary patients from 15 countries" Principal results from EUROASPIRE II. Euro Heart Survey Programme. Eur Heart J. 22:554–572.

[6] M. Karaolis, J.A. Moutiris, L. Pattichs (2010) "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE Transactions on IT in Biomedicine, vol. 14, No. 3.

[7] I.H. Karlberg and S.L. Elo (2009) "Validity and utilization of epidemiological data: A study of Ischaemic heart disease and coronary risk factors in a local population", Public Health – Elsevier health/journals/pub.

[8] J. Kunc, S. Drinovec, Rucigaj, and A. Mrhar, (2010) "Simulation analysis of coronary heart disease, congestive heart failure and end-stage renal disease economic burden." Mathematics and computers in simulation.

[9] L. Ping, G. Ronglin, Z. Xuezhong, W. Sa, (2010) "Empirical Study on Treatment of Coronary Heart Disease with Famous Doctors' Method of Regulating Spleen and Stomach based on Simplified Point-wise Mutual Information", IEEE International Conference on

Bioinformatics and Biomedicine Workshops, pp. 617-619.

[10] Z.J Zhang and S.H Wang (2007) “Synopsis of Prescriptions of the Golden Chamber” Beijing: People’s Medical Publishing House, pp.27–29.

[11] R.F.Abdel-Kader,(2010) “Genetically Improved PSO Algorithm for Efficient Data Clustering”, in Proc IEEE Int., Conf., (ICMLC) pp. 71-75.

[12] http://archive.ics.uci.edu/ml.

[13] Han J and Kamber M. (2006) “Data mining: concepts and techniques” Second edition’ University of Illinois at Urbana-Champaign.