Send Orders for Reprints to reprints@benthamscience.ae

The Open Automation and Control Systems Journal, 2014, *6*, 1975-1979

Open Access

Empirical Analysis of Genetic Evolution Algorithm Based on Multi-Layer Chromosome of Gene Expression Programming and K-Means Clustering Algorithm

Zhang Honghui* and Guo Rongyan

School of Physics and Mechanical & Electrical Engineering, Zhoukou Normal University, Zhoukou, Henan, 466001, China

Abstract: As hypertension has become one of the principal diseases affecting human health, and its prediction accuracy is a topic of concern for the medical professionals, so the computer-aided prediction model of target organ damage of primary hypertension is worthy of research. GEP model with simple chromosome, linear, compact, easily genetic manipulation can eliminate the correlation of gene expression inputs in a variety of training samples. And it has the intelligent, more flexible model structure, having wider methodological applicability, higher prediction accuracy and other characteristics. This study shows that the prediction model has a great prospect for application in the auxiliary prediction of target organ damage in primary hypertension. And classification accuracy is 87.5%, which is higher than SVM and BP neural networks.

Keywords: Feature gene, gene expression profile, K-mean clustering, multi-layer chromosome.

1. INTRODUCTION

Hypertension is a kind of systemic disease involving systemic arterial pressure as the main characteristic that is caused by polygenic inheritance, environment and interaction of multiple risk factors. Hypertension can be divided into primary hypertension and secondary hypertension, and the former accounted for more than 95% [1]. Hypertension is an independent disease and raises important risk factors of cardio-cerebro-vascular disease when some life-threatening complications occurred such as hypertensive crisis and hypertensive encephalopathy. Therefore, prevention and treatment of hypertension should not be neglected [2]. At present, there have been many methods for the disease prediction, such as Logistic regression, artificial neural networks and support vector machines (SVM). So, we need a forecasting technique and method for new applications in the field in order to further improve the accuracy.

The tumor is a highly heterogeneous disease. It is difficult to find the type of tumor based on the clinical pathological analysis. Gene Chip technology is available to people with high-flux, precise, sensitive, and rapid molecular level detection for observation of the tumor. DNA microarray experiment can get tens of thousands of gene expressions at a time. This technology provides a new research method for oncology research. Classification and Detection using gene expression profiles of tumor samples is gradually becoming an important research field of bioinformatics. Since Golub *et al*. used leukemia gene expression data as classified samples to propose gene choice algorithm based on the weight voting, many new methods of informative gene selection or feature information extraction have been proposed. Furlanello *et al*. proposed a fast feature ordering algorithm, which eliminates a large number of redundant genes with weight distribution of entropy. Chun *et al*. proposed the algorithm that obtain phenotype and gene information from gene expression data for better scalability and performance. In this paper, the signal-to-noise ratio and Bhattacharyya distance method is combined to eliminate the gene irrelevance, and the scalar feature extraction methods is used to select classification factors and determine the gene label.

This study conducts the prediction research using gene expression programming technology for the prediction of target organ damage of primary hypertension. Compared with BP neural networks and SVM prediction, the prediction model's results show that the model has a simple structure, high precision and other characteristics [3].

2. K-MEANS CLUSTERING ANALYSIS METHOD

Clustering does not require any prior domain knowledge; the basic idea is to define distance or similarity coefficient between multivariate distances or similarity coefficient to determine the multivariate relationship, which is the classification. Cluster number K and iteration time or convergence conditions must be designated before using the k-means clustering algorithm. K initial centers are appointed, and each gene is assigned to recent or "similar" center to form the class according to a certain similarity measure standard then average vector of every kind is taken as center of mass redistribution and iterative convergence until class (such as center fixed) or maximum iteration times is achieved [4].

Steps of k-means clustering analysis method:

*Address correspondence to this author at the School of Physics and Mechanical & Electrical Engineering, Zhoukou Normal University, Zhoukou, Henan, 466001, China; E-mail: zhanghonghui@zknu.edu.cn

1874-4443/14

Step 1 Enter gene expression matrix

$$U_k = \begin{Bmatrix} 1; k \\ 0; others \end{Bmatrix}$$

and class number K, initialize center V, set maximum iterations T and center convergence error tolerance δ.

Step 2 calculate distance d_{ik} between vector $X_i(i=1,\cdots,n)$

and every center $V_k(k=1,\cdots,K)$

(using Euclidean distance $d_{ik} = d(x_j, v_k)$

or Pearson related coefficient $d_{ik} = 1 - r(x_i, v_k)$, distribute X_i to the nearest center, that if

$$d_{ik} = \min\{d_{ik}\},\ X_i \in S_k,\ U_{ki} = \begin{Bmatrix} 1; k = k^* \\ 0; others \end{Bmatrix}.$$

Step 3 counter V using the newest U:

$$V_k = \frac{\sum_{i=1}^{n}(U_{ki})X_i}{\sum_{i=1}^{n}(U_{ki})} \quad (1)$$

And calculate center error:

$$E_t = \| V_t - V_{t-1} \|_2 = \sqrt{(V_t - V_{t-1})^T (V_t - V_{t-1})},$$

it is iteration time variable.

Step 4 proceed repeatedly step 2 and step 3, until $E_t < \varepsilon$ or $t = T$, output final U and V.

3. DESIGN PREDICTION MODEL AND FUNDAMENTAL OF SVM

3.1. Parameter Design

Parameter design is very important in the model design of GEP, but preference also is obtaining lacking effective laws which is usually achieved by calculation of actual conditions. So we select related parameters with the comparison of repeated experiments, and the parameter design is shown in Table **1**.

3.2. Fitness Function Design

Fitness function design directly affects the performance of the algorithm. At the same time, GEP in the early evolution usually produces some extraordinary individuals. If the fitness function is not proper, these exceptions control the selective process because the competitiveness of the individual is too prominent, which may make the algorithm too premature [5]. On the other hand, in case of GEP in late evolution, when the algorithm approaches convergence, owing to the difference of individual fitness in small population, it is unlikely to continue to evolve. And if the fitness function design is poor, the algorithm will be stagnant [6]. In order to solve this contradiction, this study uses the formula 1 as the fitness function of the algorithm.

$$f = R^2 = 1 - \frac{SSE}{\mathrm{SST}} \quad (2)$$

$$SSE = \sum_{j=1}^{m} (y_j - \hat{y_j})^2$$

is called the residual sum of squares. $\hat{y_j}$ is estimated for the variables x on the function y. The total sum of deviation square is

$$SST = \sum_{j=1}^{m} (y_j - \bar{y_j})^2,\ \bar{y_j}$$

as the average of y. The range of f is [0,1], the greater the value, the higher the degree of fitting of a model.

Table 1. Parameter Setting of GEP Algorithm.

Parameter	Detailed delineation
Evolution algebras	1000
Population size	30
Function set	+, −, ×, ／, Sqrt, Exp, Ln, Abs, Sin, Cos, log
Chromosomal structure	Genetic head length 6, Chromosome 5 genes constitute
Mutation probability	0.044
Inversion probability	0.1
IS transformation probability, RIS transformation probability	0.1, 0.1
single-point recombination probability, two-point recombination probability	0.3, 0.3
Genetic recombination probability, Genetic variation probability	0.1, 0.1
Connection functions	+

3.3. Genetic Manipulation Design of GEP

In GEP parameters, the recombination and mutation probability directly affects the astringency of the algorithm. The greater the reorganization probability is, the faster the speed of new individual is, while the greater the likelihood of destruction of genetic scale is. But, if it is too small to make the search process fast, algorithm evolution will cease to make progress. If mutation probability is too small, it is not easy to generate a new individual structure; and if it is too large, the GA becomes a pure random search algorithm. Correlated probability of genetic operators on Traditional GEP is pre-set, which can't carry on the correlation adjustment along with the algorithm running. To this characteristic with the idea of self-adapting adjustment GEP parameter, an improved method is proposed that adaptively overcome

premature phenomenon by changing crossover operators and mutation operator probability on the average fitness.

Average fitness reflects the trend of the whole population in a sense, when the maximum fitness and average fitness are close to converge, they should increase the rate of recombination and mutation. If the maximum fitness is higher than the average fitness, they should reduce the increasing rate of crossover and mutation [7, 8].

Selection of single-point and two-point recombination methods: A single-point method is randomizing a cross-location in the two parent and interchanging the part behind the crossing point of chromosome, then we can get two descendant chromosomes. The two-point method is randomizing two cross-location in the parent and interchanging the part between the crossing point of chromosome. Recombination probability P_c is defined as follows:

$$P_c = \begin{cases} P_c * \dfrac{f_{\max} - f}{f_{\max} - f_{avg}}, f \geq f_{avg} \\ P_c, f < f_{avg} \end{cases} \tag{3}$$

In which $\mathcal{f}$ is the larger one in two individuals for participation in the operation of the fitness function value, $f_{\max}$ is the largest fitness value, f_{avg} is the average fitness value.

Variation can occur in any part of the chromosome, however, the chromosomes structure must remain intact. Mutation operation in GEP can occur in the stained-bit random testing; when meeting a certain mutation probability P_m, it will become another symbol of this random one. In order to ensure the same organizational structure of GEP chromosome, the mutation in the head, variant-bit can be turned into an arbitrary sign of function set and variable set, and the mutation in the tail can only fill in the variable such as focus variable. Mutation probability P_m is defined as follows:

$$P_m = \begin{cases} P_m * \dfrac{f_{\max} - f}{f_{\max} - f_{avg}}, f \geq f_{avg} \\ P_m, f < f_{avg} \end{cases} \tag{4}$$

Assume training set,

$$(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l), x \in R^n, y \in \{+1, -1\}$$

SVM solve the optimization problem as follows:

$$\begin{aligned} &\min_{\omega,b,\xi} \frac{1}{2}(\omega \cdot \omega) + C\sum_{i=1}^{l} \xi_i \\ &s.t. \quad (\omega \cdot \Phi(x_i)) + b \geq 1 - \xi_i, \quad if \quad y_i = 1 \\ &\quad (\omega \cdot \Phi(x_j)) + b \leq -1 + \xi_j, \quad if \quad y_j = -1 \\ &\quad \xi_i \geq 0, i = 1, 2, ..., l \end{aligned}$$

In the formula, Φ is some nonlinear, C is error costs, ξ is slack variable, then its dual problem is as follows:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{5}$$

Satisfying the under constraint conditions:

$$\begin{aligned} &0 \leq \alpha_i \leq C, i = 1, 2, ..., l \\ &\sum_{i=1}^{l} \alpha_i y_i = 0 \end{aligned} \tag{6}$$

By solving quadratic optimization problem of maximization type (5) under the condition of type (6), the optimal hyperplane is constructed getting a decision function that can be used to classify the new samples

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i^0 K(x_i, x) - b^0 \right) \tag{7}$$

4. FEATURE SELECTION ALGORITHM

In the process of judging the cancer gene label, much of the "irrelevant gene" need to be removed to reduce searching scope of gene, owing to the large number of genes. In fact, some gene expression level of all samples is very close in the gene expression profiling. Neither its mean value nor its variance is an obvious difference between average person and cancerous person. It is thought that these genes are not related to the sample class, so the independent gene must be rejected.

4.1. Use Signal to Noise Ratio to Reject Independent Gene

In 1999, Golub *et al.* take the signal-to-noise ratio as the measurement weighting contribution of gene for the sample classification; *d* is signal to noise ratio of gene.

$$d = \frac{\mu_1 - \mu_2}{\sigma_1 - \sigma_2} \tag{8}$$

Bhattacharyya [6] distance is used to measure the classification information of genes; B is Bhattacharyya distance of gene.

$$B = \frac{1}{4}\frac{(\mu_1 - \mu_2)}{(\sigma_1^2 + \sigma_2^2)} + \frac{1}{2}\ln\left(\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2} \right) \tag{9}$$

Some important information may be lost by one standard, so two methods are combined in this paper. Choose appropriate threshold value to reject independent gene. According to the number of sample classification information, genes are classified to information gene and independent gene. Make S_I as genetic information collection, S_N is independent gene collection, then it is defined,

$$g \in \begin{cases} S_I, B(g) > \theta \\ S_N, B(g) \leq \theta \end{cases} \tag{10}$$

g is gene, $B(g)$ is Bhattacharyya distance or signal to noise ratio of g, θ is specified threshold.

4.2. Choose the Best Gene Center with K-Means Clustering Analysis Method

The gene scope is reduced, but direct correlation gene number of one kind of tumor is very few. Therefore, we need to carry on processing further to the gene that carries the K-average value cluster to the sample. Carry on the cluster to

start value K=K+1 in turn, starting value is selected as K=2. In computation, if K was selected too small, many primitive gene expression data would be lost; if K was selected too big, many redundancies information would be retained. Two kinds of situations can reduce classified accuracy. In order to obtain the better effect of cluster, the cluster integer K are changeable, and through many times tests, the best K value of classified effect will be found.

4.3. Establishment and Verification of Gene Label

Relative to the number of genes, the number of sample is very small, the direct application of sample for classification can cause learning problems of little sample. In fact, if features were small, the effect of categories would be better for genetic classification problem. Based on this consideration, k-means clustering analysis is used to ascertain. For classification, classification center only can't get the corresponding gene tags. Therefore, it is needed to use the classification center to get the corresponding gene tags. Get the classification center, then extract feature and choose features genes.

Features choice is that some most representative characteristics are picked out from the original features as the classification feature of the samples; the basic task is how to find out the most effective features. The most simple feature selection method is to choose those most influential characteristics for classification according to the knowledge of the expert, and the other possibility is to use mathematics method to find out the most classification of information features. If the difference of one kind of samples is called as 'inside change' and the difference of other kinds of samples is called as separate difference, separate difference of the ideal characteristics gene must be bigger. Remember separate difference as scatter and inside change as compact, use score (their ratio) to show the genes identifying ability. Score (j) represents the identification ability of gene j. The points granted is less which means that the gene and category is more related and the ability to identify is higher, the characteristics of gene j expression data separability are better. Calculate all genes score, sort up each gene according to the grades.

Suppose sample $x_{ij}, j = 1,2,....n$, feature gene center is $\bar{x}_i, i = 1,2,....k$, steps of determining gene label are:

Step 1 Classify sample genes, discriminate categories set of each gene, $S_i, i = 1,2,....k$;

Step 2 For every sample element $x_{ij} \in S_i, i = 1,2,....k$ of collection class, $S_i, i = 1,2,....k$,

Calculate Euclidean distance,

$$d_{rs}^2 = \sqrt{\sum_{i=1}^{p} (x_{ij} - \bar{x}_i)^2}$$

And related coefficient,

$$r_{ij} = \frac{\sum_{k=1}^{n} \left(x_{ij} - \bar{x}_i \right) \left(x_{kj} - \bar{x}_j \right)}{\left\{ \left[\sum_{k=1}^{n} \left(x_{ki} - \bar{x}_i \right)^2 \right] \cdot \left[\sum_{k=1}^{n} \left(x_{kj} - \bar{x}_j \right)^2 \right] \right\}^{1/2}}$$

Take the gene with the least distance as gene label.

With the support vector machines method, the classified result accuracy may be obtained by which the characteristic gene computed above is taken as the training regulations, the cancer patient sample is taken as the test collection and the sample is classified.

5. RESULTS AND DISCUSSION

5.1. GEP Computation

The best prediction discriminant was finalized by training 120 cases of sample data with GEP:

$$\begin{aligned} gep(X) &= 0.11x_1 + 0.1\sin(x_2) + 0.2x_3 \\ &+0.15\log(x_4) + 0.17\log(x_5) + 0.03\ln(0.02 \times x_6 \times x_7) \\ &+0.17\cos(x8) + 0.19\log(x9) + 0.18x10 \\ &+0.17sqrt(x_{11}) \end{aligned} \quad (11)$$

In the formula: $x_1, x_2, ..., x_{11}$ 1 represent the 11 parameters listed in Table 1 that are identified by reference[1] in 2-year actual prediction application. When the results are ≥ 0.5, the prediction has damage. And when they are <0.5, the prediction has no damage. Through the accumulation of cases and more reasonable set of parameter values, GEP excavating discriminant can provide a valuable reference. Results are shown in Table 2 that are predicted by formula (4) based on 32 cases of sample data.

Table 2. The Predicted Results

	No target organ damage(7)	Target organ damage(25)
Judge correctly number	6	22
Judge wrongly number	1	3
Percentage of judging correct	86	88
Total Percentage judging correct(%)	87.5	

In addition, this study also applied BP neural network and SVM algorithm to predict target organ damage. After the analysis of the factors, using 3-layer BP neural network, we define impact factors as 11 and the output as 1 in the application of BP neural network prediction. Referring to the Kolmogorov theorem, the number of hidden layer neurons 2× 11+1=23 is selected, which constitutes a model structure of 11-23-1. And 32 cases are selected as a training sample of network, and then the rest makes a testing sample of network. In the application of SVM regression analysis, the data sample with target organ damage is marked as -1, and without target organ damage is marked as 1. With C-SVC and ν-SVC methods respectively predicted that its radial group kernel of integral transform are selected as kernel function, C-SVC parameter is set to penalty factor C = 5000, radial basis function width σ = 10-5, ν-SVC parameter is set to ν = 0 • 5, C = 1000, σ = 10-6. Table **3** provides the accuracy comparison of GEP and other algorithms.

Table 3. The Accuracy Comparison of GEP Algorithm and Other Algorithms.

Algorithm	Judge correctly percent (%)
C-SVC	85
v-SVC	86.9
BP neural network	84
GEP	87.5

It can be seen from Table **3**: that prediction effect of GEP is better than SVM and BP neural networks.

Combine two methods to reject independent gene, when Bhattacharyya distance threshold value is $\theta = 0.1$ and signal to noise ratio threshold value i.e. $\theta > 60$,175 genes are information gene out of 2000 genes, and 1881 genes are independent gene. 134 genes after removing independent genes contain classification information in varying degrees, which is the basis for further analysis.

5.2. Simulation of K-Clustering Analysis Classification

Using Bhattacharyya distance method, the gene scope is reduced to 134 genes. Perform K-means clustering analysis method for these data. After experiment, it is found that the classification result is the best when K=2, view Fig. (**1**).

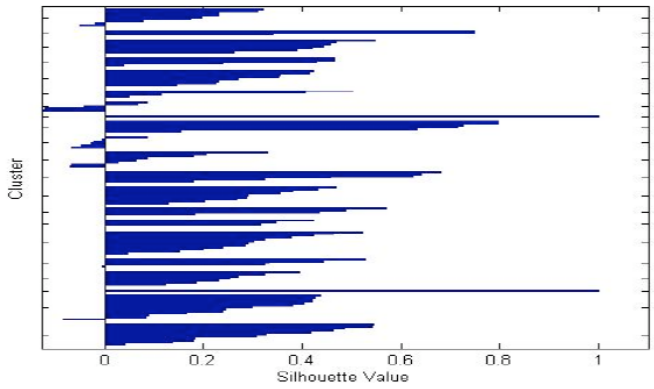

Fig. (1). Simulation figure of K-clustering analysis classification.

With fundamentals of feature selection, take feature selection to the 21 gene centers in Table, the result is shown in chart 1,C=90.

Using support vector machines method, 11 characteristic genes calculated above are taken as training set and 40 samples; of cancer patients are taken as test set to classify the samples, the accuracy of the classification results is at 87%. This suggests that the 11 characteristics genes selected contain wealth information that can represent the features of the cancer genes.

CONCLUSION

In order to predict target organ damage caused by primary hypertension, an improved algorithm is proposed in the classic GEP algorithm based on the average fitness. The algorithm changes the form of adaptive re-operator and mutation operator probabilities to overcome the premature damage. The prediction model is established to predict the sample data for 2-year target organ damage, and accuracy was 87.5%. A key problem of gene expression data classification is feature selection. In this paper, a series of data pretreatment, scalar feature extraction and K-clustering analysis are conducted; the characteristics genes have good classification ability for sample set. Support vector machines method gives 85% accuracy of the classification results. According to certain genes tags the types of cancers can be effectively judged based on the great contribution of physiology information.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] J. Wang, *Int Med [M]*, People's Medical Publishing House, 2008, pp. 243-247.

[2] Z. Sun, *Medical Statistics*, People's Medical Publishing House, 2007: 333-341.

[3] C. Ferreira, "Gene expression programming: A new adaptive algorithm for solving problems," *Complex Systems*, vol. 13, no. 2, pp. 87-129, 2001.

[4] C. Tang, T. Zhang, and J. Zuo, "Knowledge discovery on gene expression programming successive changes、outcome and direction of development," *Computer Utility*, vol. 24, no. 10, pp. 7, 2004.

[5] X. Jia, C. Tang, and J. Zuo, "Frequency excavation of function set on gene expression programming," *Computer Journal*, vol. 28, no. 8, pp. 1247, 2005.

[6] C. Ferreira, "Designing Neural Networks Using Gene Expression Programming," In: *9th Online World Conference on Soft Computing in Industrial Applications*, September 20-October 8, 2004.

[7] J.R. Koza, H. Forrest, D. Andre, and M.A. Keane, *Automatic Design of Analog Electrical Circuits Using Genetic Programming* In: *Hugh Cartwright*, (ed.). Intelligent Data Analysis in Science. Oxford: Oxford University Press. Chapter 8, 2000, pp. 172-202.

[8] J. R. Koza, M.A. Keane, J. Yu, F.H. Bennett, and W. Mydlowec, Automatic Creation of Human-Competitive Programs and Controllers by Means of Genetic Programming, *Genetic Programming and Evolvable Machines*, vol. l, no. l-2, pp. 124-164, 2000.

Received: October 16, 2014 Revised: November 24, 2014 Accepted: November 28, 2014