RESEARCH ARTICLE | JANUARY 10 2019

Microarray gene feature classification based on LS-SVM FREE

Zhenbin Gao

AIP Conf. Proc. 2058, 020019 (2019)
https://doi.org/10.1063/1.5085532

Microarray Gene Feature Classification based on LS-SVM

Zhenbin Gao[1, a)]

[1]*Institute of Mathematics & Applied Mathematics, School of Statistics, Xi'an University of Finance and Economics, Xi'an 710100, China.*

[a)]Corresponding author email: gaozb2700@sina.com

Abstract. DNA microarray has the characteristics of the higher dimension and redundancy, they bring into a series of the difficulties for the gene feature clasiffication. Pertaining to the two classical microarray datasets (cancer of colon set and leukemia set), firstly, the preprocess has been taken by the normalizing and the redundant data have been withdrawn; secondly, Principal Component Analysis method has been adopted to reduce the dimension of datasets and the information gene sets have been obtained; Finally, multiple classifiers have been utilized for the simulating tests, such as LS-SVM, SVM, BP, RBF, etc. They demonstrate that LS-SVM classifier has the higher accuracy for classification and show the approached method can make the correct judgment for classifying the feature of gene dataset, and provide the verifying reliance for clinical therapy further.

Key words: Microarray; Feature Classification; Reducing Dimension; Least Square Support Vector Machine (LS-SVM).

INTRODUCTION

With the development of gene expression profile technology, more and more gene expression data in human tissues have been obtained, therefore, the analysis and the modeling of those datasets have become the focus topic in the field of bioinformatics.

Numerous researchers have conducted fruitful studies in this direction [1-4]. Chiaretti studied the classification of acute leukemia microarray dataset in T-cells and applied it to clinical treatment and prediction [5]. In the clinical treatment of lung cancer, Sun made prejudgment through the dataset feature classification [6]. Devi selected information genes based on mutual information, and then used SVM classifier to classify and evaluate the microarray dataset [7]. Wang used the improved PLS-RFE algorithm to classify and select the features of multiple datasets, and the computational efficiency was improved [8]. Sharbaf used Fisher index for reducing the dimension of dataset, and then combined Cellular Learning Automata (CLA) method with the ACO algorithm to improve the classification accuracy of genetic characteristics [9]. Khan proposed a new adaptive radial basis kernel function, and conducted simulation studies on nonlinear system identification, the dataset classification and the function approximating calculation [10]. Xiao proposed a deep learning algorithm based on multi-model integration and verified three cancer datasets [11]. Li studied the classification for acute leukemia dataset, and used SVM as a classifier for subtype recognition [12]. Ma combined density clustering with the sharing neighbor method to perform clustering analysis on the dataset [13]. Han combined rough sets theory with SVM to reduce genetic characteristics and then used SVM for data classification [14]. Zhu proposed a multiple hypothesis test method for differential expression of microarray genes, which effectively reduced false positive results caused by data noise [15]. Yao studied the parameter optimization algorithm of LS-SVM feature selection [16]. Sun used the improved Lasso method to select the characteristics of information genes and eliminate redundant genes [17]. Yang proposed a nuclear least squares feature gene selection method to reduce the dimension of microarray data, and then used the extreme learning machine for training and prediction [18].

SVM is a machine learning algorithm based on the statistical learning theory and the structural risk minimization principle, it can effectively deal with the classification problem of high-dimensional samples and becomes a powerful

International Conference on Frontiers of Biological Sciences and Engineering (FBSE 2018)
AIP Conf. Proc. 2058, 020019-1–020019-6; https://doi.org/10.1063/1.5085532
Published by AIP Publishing. 978-0-7354-1786-1/$30.00

means to overcome the traditional difficulties such as “dimension disaster” and “overlearning”. However, its inequality constraint and quadratic programming algorithm causes the less computational efficiency. LS-SVM algorithm selects the hyperplane by the least square method and then introduces the square loss function. Compared with SVM method, it can transform the inequality constraint into linear equality condition and the quadratic programming problem into linear solving problem, avoids solving time and improves operation efficiency. In this paper, LS-SVM is put forward to classify the microarray datasets and qualify the advantages of the higher accuracy and better computational efficiency.

RELATED WORK

Let $O=(X,Y)$ be a dataset. $X=\{x_1,x_2,\cdots,x_N\}$ is the sample set. $Y=\{y_1,y_2,\cdots,y_N\}$ is label set. Where, $x_k\in X$ and $x_k\in R^m$; $y_k\in Y$,which associated with x_k . Suppose that the sample can be classified two class(normal and abnormal),that is

$$y_k=\begin{cases}1, & \text{if normal}\,x_k\\ -1, & \text{if abnormal}\,x_k\end{cases} \tag{1}$$

The aim is to find out a group of feature which differentiate the gene expressions of samples accurately.

LS-SVM Method

SVM model is to construct a classifier in the following form:

$$f(x)=\text{sign}[w^T\varphi(x)+b] \tag{2}$$

Where, $\varphi(x)$ is the nonlinear mapping function used in kernel machines; w is the coefficient vector; b is the bias term; $\text{sign}(\bullet)$ is sign function.

LS-SVM classification problem can be described as solving the following equation constraint optimization problem:

$$\begin{aligned}&\min_{w,b,e}J(w,e)=\frac{1}{2}w^Tw+\frac{1}{2}\gamma e^Te\\ &\text{s.t.}\quad y_k[w^T\varphi(x_k)+b]=1-e_k,k=1,2,\cdots,N\end{aligned} \tag{3}$$

Where, $e_k(k=1,2,\cdots,N)$ is the error, $e=[e_1,e_2,\cdots,e_N]^T$, $w=[w_1,w_2,\cdots,w_N]^T$, $\gamma>0$ is the penalty coefficient to adjust the error. Construct the Lagrange function as follows

$$L(w,b,e;\alpha)=J(w,e)-\sum_{k=1}^{N}\alpha_k\{y_k[w^T\varphi(x_k)+b]-1+e_k\} \tag{4}$$

Where, $\alpha_k\geq 0(k=1,2\cdots,N)$ is Lagrange multiplier. Set $L(w,b,e_k,\alpha_k)$ partial derivative is zero, and the following linear system is obtained

$$\begin{bmatrix} I & 0 & 0 & -Z^T \\ 0 & 0 & 0 & -y^T \\ 0 & 0 & \gamma I & -1_n \\ Z & y & 1_n^T & 0 \end{bmatrix} \begin{bmatrix} w \\ b \\ e \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1_n \end{bmatrix} \tag{5}$$

Where, $Z = [\varphi(x_1)y_1, \varphi(x_2)y_2, \cdots, \varphi(x_N)y_N]^T$; I is the identity matrix; $y = [y_1, y_2, \cdots, y_N]^T$ $1_N = [1,1,\cdots,1]^T$; $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_N]^T$. Eliminate the variables e, w, and according to Mereer condition:

$$\Omega_{sl} = y_s y_l \varphi^T(x_s)\varphi(x_l) = y_s y_l K(x_s, x_l), (s, l = 1,2,\cdots,N) \tag{6}$$

Where, $K(x_s, x_l)$ is the radial basis kernel function .The equation (5) can be described as follows?

$$\begin{bmatrix} 0 & -y^T \\ y & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1_N \end{bmatrix} \tag{7}$$

Where, $\Omega = [\Omega_{sl}]_{N \times N}$. Set $A = \Omega + \gamma^{-1} I$, since it is a symmetric semi-positive definite matrix, and the above equation has the solution. LS-SVM classifier can be obtained as:

$$f(x) = \text{sign}[\alpha_k y_k K(x, x_k) + b], k = 1,2,\cdots,N \tag{8}$$

Where, α_k and b are the solution of the equation (7); Take kernel function $K(x, x_k)$ of the equation (8) as the radial basis function form:

$$K(x, x_k) = \exp\left\{ -\frac{\|x - x_k\|^2}{\sigma^2} \right\}, (k = 1,2,\cdots,N) \tag{9}$$

Using microarray technology to measure gene expression level is often characterized by strong noise and big fluctuation, and there are many related variables behind the massive data that cannot be directly observed. The following steps are used to realize the characteristic classification of microarray data:

Step1: Data preprocess. Mean - centered and normalized data sets are selected;

Step2: Extract the principal component. Calculate the correlation coefficient matrix of the sample data ,it can be processed through the dimensional reduction using PCA method, and the feature information gene can be obtained;

Step3: Classification model training. LS-SVM and other classifiers can be used to classify the dataset after feature extraction;

Step4: Test the classification model. The test samples will be substituted into the classification model, and the performance of each classifier can be evaluated by the different test method.

SIMULATION RESULTS

Experiment Data and Development Environment

Two open microarray datasets were used to evaluate the performance of the proposed algorithm. the detailed description of the datasets are shown in table 1.The colon cancer data set consists of 62 samples and is divided into two categories: normal (22 cases) and colon cancer (40 cases) [1]; The leukemia data set included 128 samples, which

were classified into two different types of tumors: T cell ALL (a total of 33 cases) and B cell ALL (a total of 95 cases) [5, 19].

TABLE 1. Experimental data and description.

Dataset	Number of Genes	Sample Size	Number of Classification
Colon Cancer	2000	62 (22/40)	2
Leukemia	12625	128 (33/95)	2

Experimental environment in this paper: Intel CPU 2.53GHz processor, 2GB PC, Windows Xp operating system, Matlab2014b development environment.

Analysis of Experimental Results

Experiment 1. Classification of Dataset of Colon Cancer

For the data set of colon cancer, the data were preprocessed .After the feature information gene was extracted, the normal and tumor samples were randomly assigned to the training set and test set at a ratio of nearly 2:1.There were 40 samples in the training set (including normal samples 14 and tumor samples 26), and 22 samples in the test set (normal samples 8 and tumor samples 14). Then, the characteristic information genes of the top 10 were extracted by PCA method (see table 2).

TABLE 2. Characteristic genes selected from data of colon cancer

Serial Number	Name of Gene	Serial Number	Name of Gene
1	X53799	6	R80427
2	M29273	7	X75208
3	U21914	8	D29808
4	L00352	9	M59807
5	X90858	10	D13627

LS-SVM and other classifiers were used to classify the selected characteristic genes. The classification accuracy results of each classifier are shown in table 3.

TABLE 3. Experimental results of selected characteristic gene sets

Classifier	Number of Gene	LOOCV Accuracy Rate	IV Accuracy Rate
PNN	10	0.625	0.6818
RBF	10	0.675	0.6818
BP	10	0.65	0.6363
SVM RBF_ker	10	0.675	0.7272
SVM Lin_ker	10	0.600	0.6363
SVM poly_ker	10	0.625	0.7272
LS - SVM RBF_ker	10	0.975	0.6818

Note: PNN stands for probabilistic neural network; RBF is a radial basis neural network; BP is a back propagation neural network; RBF_ ker, Lin_ ker and poly_ ker are radial basis kernel function, linear kernel function and polynomial kernel function, respectively; LOOCV stands for Leave One-Out Cross Validation; IV stands for Independent Validation.

Experiment 2. Classification of Leukemia Dataset

After dataset of leukemia preprocessing, the characteristic information genes of the top 10 ingredients were obtained by PCA method, as shown in table 4.

TABLE 4. Characteristic genes selected from the leukemia dataset

Serial Number	Name of Gene	Serial Number	Name of Gene
1	X110_at	6	X1221_at
2	X1098_at	7	X1210_s_at
3	X1189_at	8	X1086_at
4	X1106_at	9	X1173_g_at
5	X1178_at	10	X1199_at

Two types of samples in the data set are allocated to the training set and the test set. There are 65 samples in the training set (48 samples of T cells and 17 samples of B cells), and 63 samples in the test set (47 samples of T cells and 16 samples of B cells). The selected characteristic genes were classified. The results are shown in table 5.

TABLE 5. Experimental results of selected characteristic gene sets

Classifier	Number of Genes	LOOCV Accuracy Rate	IV Accuracy Rate
PNN	10	0.9692	0.7460
RBF	10	0.8923	0.8254
BP	10	0.7384	0.7460
SVM RBF_ker	10	0.7384	0.8254
SVM lin_ker	10	0.7384	0.7460
LS - SVM RBF_ker	10	1.0000	0.9365

As can be seen from table 3 and table 5, in LOOCV test, the accuracy of LS-SVM classifier was the highest at 97.5% (table 3) and 100% (table 5), respectively. In IV test, the accuracy of LS-SVM classifier of the characteristic gene set of leukemia was the highest at 93. 65% (table 5), while the results of the dataset of colon cancer showed little difference from those of other classifiers.

CONCLUSIONS

Microarray data plays an important role in the diagnosis of diseases. However, the high dimension and redundancy makes it extremely difficult to further explore the knowledge contained in microarray data. In order to obtain better classification prediction results, the datasets were averaged and normalized before the classification. PCA method was used for the dimension reduction, the feature information genes of two datasets were extracted and classified using different classifiers (PNN, RBF, BP, SVM and LS-SVM, etc).It can be seen from the results of experimental methods that the accuracy of LS-SVM is higher than that of other classifiers, the reason comes from its linear equality condition and linear solving matrix equation, which can provide a relatively reliable judgment basis for medical clinical practice.

REFERENCES

1. U. Alon, N. Barkai, D.A. Notterrman, K. Gish, S. Ybarra, D. Mack and A.J. Levine, Proc. Natl. Acad. Sci. USA, 96,pp.6745-6750(1999).
2. T.S. Furey, et al, Bioinformatics, 16(10), pp.906-914(2000).
3. T. R. Golub, D. K. Slonim, P. Tamayo, et al, Science, 1999, 286, pp. 531-537(1999).
4. I. Guyon, J. Weston, S. Barnhill, et al, Machine Learning, 46, pp. 38 9-422(2002).
5. S. Chiaretti, Li Xiaochun, R. Gentleman, A. Vitale, et al, Blood, 103(7), pp.2771-2778(2004).
6. Sun Zhifu, Yang ping. Cancer Composites, Biomarkers & Prevention, 15(11), pp.2063-2068(2006).
7. A.V. C. Devi, D. Devaraj, M. Venkatesulu, Procedia Computer Science, 47, pp. 13-21(2015).
8. Aiguo Wang, Ning An, Guilin Chen, Lian Li,G. Alterovitz, Computers in Biology and Medicine, 62, pp. 14-24 (2015).
9. F. V. Sharbaf, S. Mosafer, M.H. Moattar, Genomics, 107, pp231-238(2016).
10. S. Khan, I. Naseem, R. Togneri, M. Bennamoun, Circuits Syst. Signal Process, 36, pp1639-1653(2017).
11. Yawen Xiao, Jun Wu,Zongli Lin,Xiaodong Zhao, Computer Methods and Programs in Biomedicine, 153, pp.1-9(2018).

12. Yingxin Li, Quanjin Liu, Xiaogang Ruan, Chinese Journal of Biomedical Engineering, 24 (2), pp. 240-244 (2005) (in Chinese).
13. Yu Ma, Li Chen, Liqi Ou, Computer Engineering and Application, 5,pp.176-178(2006) (in Chinese)
14. Li Han, Yunsong Qi, Jun Wang, Science Technology and Engineering, 9 (1) ,pp.152-155(2009)(in Chinese)
15. Qinping Zhu, Xiaohan Hu, Yunsong Qi, Science Technology and Engineering, 10 (27), pp.6675 (2010) (in Chinese).
16. Quanzhu Yao, Jie Cai, Computer Engineering and Application, 46 (1), pp. 134-136,229(2010) (in Chinese).
17. Gang Sun, Jing Zhang, Journal of Chinese Computer Systems, 36 (6), pp. 1209-1213(2015) (in Chinese).
18. Qin Yang, Hongwei Dong, Yanna Xue, Transducer and Microsystem Technologies, 35 (5), pp. 146-148, 153(2016) (in Chinese).
19. L. Torgo, Hongcheng Li, Daolun Chen, Liming Wu, Data Mining and R Language, (Mechanical Industry Press,Beijing,2016),pp.162-174 (in Chinese).