International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:5, No:3, 2011

Gene Expression Data Classification Using Discriminatively Regularized Sparse Subspace Learning

Chunming Xu

Abstract—Sparse representation which can represent high dimensional data effectively has been successfully used in computer vision and pattern recognition problems. However, it doesn't consider the label information of data samples. To overcome this limitation, we develop a novel dimensionality reduction algorithm namely dscriminatively regularized sparse subspace learning(DR-SSL) in this paper. The proposed DR-SSL algorithm can not only make use of the sparse representation to model the data, but also can effective employ the label information to guide the procedure of dimensionality reduction. In addition,the presented algorithm can effectively deal with the out-of-sample problem.The experiments on gene-expression data sets show that the proposed algorithm is an effective tool for dimensionality reduction and gene-expression data classification.

Keywords—sparse representation, dimensionality reduction, label information, sparse subspace learning, gene-expression data classification.

I. Introduction

IN recent years, with the rapid development of microarray gene-expression technology, it is now possible to simultaneously monitor the expression of all genes in the genome with a single experiment. One important application of gene expression data is the classification of cancer or other diseases, which draws a great number of researchers' attention[1]. Typically, the gene expression data sets are characterized by thousands of variables on only a few observations. It has been observed that although there are a lot of genes for each observation, the number of tissue samples ranges from tens to hundreds. In other words, there is much redundant information resided in the high-dimensional gene-expression data. To remove redundant information, dimensionality reduction technique is an effective way.

Till now, many dimensionality reduction algorithms in machine learning have been used to solve the microarray gene-expression classification problem. The popular dimensionality reduction algorithms involved in gene data analysis include principal component analysis (PCA)[2], independent component analysis (ICA) [3] and linear discriminant analysis (LDA) [4], etc. In [5], the authors used the PCA algorithm to reduce the input dimensions of gene expression data. PCA provides an efficient way to compress the gene expression data without losing much information. However, PCA can only find the second-order statistical information of the data. The authors used the ICA algorithm to model the gene expression data in [6]. Different form PCA, ICA can take into account higher-order dependencies in the data. In [7], [8], the authors used the LDA algorithm to reduce the input dimensions of gene expression data. LDA seeks the optimal transformation that maximizing the between-class scatter while at the same time minimizing the within-class scatter. LDA has been widely used in many practical applications such as face recognition due to the fact that it can extract the most discriminatory features.

However, recently studies show that just as face images, the gene expression data are also concentrated in a nonlinear subspace. In such situation, the linear subspace based dimensionality reduction methods will fail to work well. In recent years, manifold learning-based dimensionality reduction approaches such as Isomap[9], locally linear embedding(LLE)[10] and local preserving projections(LPP)[11] have attracted a lot of attention. It is believed that these methods are effective in discovering the intrinsic geometrical structure of the nonlinear data. The manifold learning-based dimensionality reduction approaches have also been used to solve the microarray gene-expression classification problem[12], [13]. Although manifold learning-based dimensionality reduction approaches are very effective, they are not easily applied in certain applications due to their complexity and storage requirements. More recently, sparse representation which is derived by solving an optimization problem has been successfully used in computer vision and pattern recognition problems[14], [15], [16].

In this paper, we investigate the problem of microarray classification and propose a dscriminatively regularized sparse subspace learning algorithm, which will be abbreviated as DR-SSL. DR-SSL aims to find a novel framework for dimensionality reduction which can not only adopt sparse representation to model the data, but also can employ the label information to improve the classification quality. Specifically, we first construction the objective function of sparse subspace learning based on sparse coding using the training samples. Secondly , we construct a new discriminatively regularized term via the label information which is vital for the following dimensionality reduction and classification. Then the obtained discriminatively regularization is incorporated into the objective function of sparse subspace learning to form a novel framework for dimensionality reduction. Lastly, an iterative algorithm is developed to find the solution of the framework. We demonstrate the usefulness of our approach on gene expression data sets and the experiment results show that the proposed algorithm achieves better performance compared to the conventional dimensionality reduction methods such as

Chunming Xu is with the School of Mathematical Sciences, Yancheng Teachers University,Yancheng, 224002,China(e-mail: yctcxcm@gmail.com).

PCA and LPP.

The rest of the paper is organized as follows. In section II, we give a brief review of the sparse representation algorithm.Section III presents the proposed DR-SSL algorithm. Section IV presents the experimental results on gene expression data sets. Finally, we conclude this paper in Section V.

II. Sparse Representation

The problem of finding the sparse representation of a signal in a given overcomplete dictionary can be formulated as follows. Given a training set $X = [x_1, x_2, \ldots, x_n(x_i \in R_n)]$, each column of X is a sample vector.For each sample point x_i in the data set , we expect to reconstruct it using a few data points in X .The objective can be achieved by solving the follows minimization problem:

$$min|s_i|_1, s.t. \quad x_i = Xs_i \tag{1}$$

where $S_i \in R_n$ is the coefficient vector and $|s_i|_1$ is the l_1 norm of vector . However, one issue with the minimization problem (1) is that when the size of matrix X satisfied $m >> n$, it does not have exact solutions. A generalized version of Eq. (1) which allows for certain degree of noise can be expressed as follows:

$$J(S) = \arg\min \sum_{i=1}^{n} \|x_i - \sum_{i=1} XS_i\|^2 + \lambda_1 \|S_i\|_1 \tag{2}$$

In fact, Eq. (2) is an l_1- regularized least square problem where the positive parameter λ_1 is a scalar regularization that balances the contribution of the reconstruction error against the sparseness of the coefficients. The sparsity regularization term can not only ensure the under-determined system has a unique solution but also allow the learned representation to capture salient patterns of local descriptors.

III. Discriminatively Regularized Sparse Subspace Learning(DR-SSL)

In this section, we present a dscriminatively regularized sparse subspace learning algorithm(DR-SSL) for dimensionality reduction. The proposed DR-SSL algorithm can not only make use of the sparse representation to model the data, but also it can effective employ the label information to to improve the performance in the learned subspace.

A. Sparse Subspace Learning

Sparse subspace learning algorithm attempts to find a projection matrix which maps high dimensional data to lower dimensional data space for classification problems. Let $P \in R^{m \times d}$ denote transformation matrix. Project P onto x_i yields the low dimensional vector y_i

$$y_i = P^{\mathrm{T}} x_i \tag{3}$$

The sparse subspace learning can be formulated as an optimization problem

$$J(S, P) = \arg\min \sum_{i=1}^{n} \|P^{\mathrm{T}} x_i - P^{\mathrm{T}} \sum_{i=1} XS_i\|^2 + \lambda_1 \|S_i\|_1 \tag{4}$$

The sparse subspace learning algorithm can effectively deal with the out-of-sample problem [10],i.e., it can map a new testing point directly. One issue with the sparse subspace learning algorithm is that it doesn't consider the label information of the samples which is critical for the success of the dimensionality reduction and classification problems.To address the issue , we will propose a novel discriminatively regularized sparse subspace learning (DR-SSL) algorithm via the label information in the following subsection.

B. Discriminatively Regularized Term

Regularization theory has been used in a wide variety of applications to derive a large family of novel algorithms. There exists a lot of regularization methods. In [17], the authors constructed the discriminatively regularized term utilizing the underlying label knowledge which is vital for classification.Given l data points $x_1, x_2, \ldots, x_l \in R_d$ that are distribduted on a underlying submanifold.Let $l(x_i)$ be the class label of x_i and its k nearest neighbors be $N(x_i) = x_i^1, x_i^2, \cdots, x_i^k$ .By the label information, the set $N(x_i)$ can be further split into two subsets, $N_b(x_i)$ and $N_w(x_i)$. $N_w(x_i)$ contains the neighbors having the same label with x_i , while $N_b(x_i)$ contains the neighbors that sharing different labels. Specifically,

$$N_w(x_i) = \{x_i^j | l(x_i^j) = l(x_i), 1 \le j \le k\} \tag{5}$$

$$N_b(x_i) = \{x_i^j | l(x_i^j) \ne l(x_i), 1 \le j \le k\} \tag{6}$$

Define the weight matrices W_b and W_w respectively as follows:

$$W_{b,ij} = \begin{cases} 1 & x_i \in N_b(x_j) \quad or \quad x_j \in N_b(x_i) \\ 0 & else \end{cases}$$

$$W_{w,ij} = \begin{cases} 1 & x_i \in N_w(x_j) \quad or \quad x_j \in N_w(x_i) \\ 0 & else \end{cases}$$

The discriminatively regularized term aims to maximize $\sum (y_i - y_j)^2 W_{b,ij}$ while at the same time minimize $\sum (y_i - y_j)^2 W_{w,ij}$,where $y_i = P^{\mathrm{T}} x_i$ .Based on above analysis, we get the discriminatively regularized term

$$\Phi(P) = \alpha (y_i - y_j)^2 W_{w,ij} - (1 - \alpha) \sum (y_i - y_j)^2 W_{b,ij} \qquad y_i = P^{\mathrm{T}} x_i \tag{7}$$

where α is a positive parameter and $0 < \alpha < 1$.

Note that the discriminatively regularized term exploits not only the discriminant structure information but also the local manifold structure of given labeled samples.By the discriminatively regularized term, DR-SSL will apart the data samples from different classes at each local area well.

C. DR-SSL

Based on the discriminatively regularized term, the objective function of DR-SSL can be defined as follows:

$$J(S, P) = \arg\min \sum_{i=1}^{n} \|P^{\mathrm{T}} x_i - P^{\mathrm{T}} \sum_{i=1} XS_i\|^2 + \lambda_1 \|S_i\|_1 + \lambda_2 \Phi(P) \tag{8}$$

where λ_2 is the regularization parameter that controls the complex of the discriminatively regularized term.

Because $\sum(y_i - y_j)^2 W_{b,ij} = \sum(P^{\mathrm{T}}x_i - P^{\mathrm{T}}x_j)^2 W_{b,ij} = P^{\mathrm{T}}P(D_b - W_b)P^{\mathrm{T}}P$,where D_b is a diagonal matrix with entries $D_{b,ii} = \sum_j W_{b,ij}$ .On the other hand ,$\sum(y_i - y_j)^2 W_{w,ij} = \sum(P^{\mathrm{T}}x_i - P^{\mathrm{T}}x_j)^2 W_{w,ij} = P^{\mathrm{T}}X(D_w - W_w)X^{\mathrm{T}}P$. Also,$D_w$ is a diagonal matrix with entries $D_{w,ii} = \sum_j W_{w,ij}$. Then the discriminatively regularized term can be written as follows:

$$\Phi(P) = \alpha P^{\mathrm{T}}X(D_b - W_b)X^{\mathrm{T}}P - (1-\alpha)P^{\mathrm{T}}X(D_w - W_w)X^{\mathrm{T}}P \quad (9)$$

Define $L_b = D_b - W_b \;\; L_w = D_w - W_w$,we have:

$$\Phi(P) = \alpha P^{\mathrm{T}}XL_bX^{\mathrm{T}}P - (1-\alpha)P^{\mathrm{T}}XL_wX^{\mathrm{T}}P \quad (10)$$

Therefore,Eq.(8) is equivalent to:

$$J(S,P) = \arg\min \sum_{i=1}^{n} \|P^{\mathrm{T}}x_i - P^{\mathrm{T}}\sum_{i=1} XS_i\|^2 + \lambda_1\|S_i\|_1 + \lambda_2(\alpha P^{\mathrm{T}}XL_bX^{\mathrm{T}}P - (1-\alpha)P^{\mathrm{T}}XL_wX^{\mathrm{T}}P) \quad (11)$$

There are two parameters,i.e., S and P in Eq. (11) ,and there is not a closed-form solution for the optimization problem .In this paper, we propose an iterative algorithm to solve it.

Firstly,fix P , $J(S,P)$ is reduced to

$$J(S) = \arg\min \sum_{i=1}^{n} \|P^{\mathrm{T}}x_i - P^{\mathrm{T}}\sum_{i=1} XS_i\|^2 + \lambda_1\|S_i\|_1 \quad (12)$$

Eq.(12) is a l_1- regularized least square problem.We can use some standard convex optimization techniques to solve it.

On the other hand, when S is given, $J(S,P)$ becomes

$$J(P) = \arg\min \sum_{i=1}^{n} \|P^{\mathrm{T}}x_i - P^{\mathrm{T}}\sum_{i=1} XS_i\|^2 + \lambda_2(\alpha P^{\mathrm{T}}XL_bX^{\mathrm{T}}P - (1-\alpha)P^{\mathrm{T}}XL_wX^{\mathrm{T}}P) \quad (13)$$

Let $L = [l_1, l_2, \ldots, l_n], l_i = l_i - X_iS_i$. Following some simple algebraic steps, it is not easily to see that

$$\sum_{i=1}^{n} \|P^{\mathrm{T}}x_i - P^{\mathrm{T}}\sum_{i=1} XS_i\|^2 = \|P^{\mathrm{T}}L\|^2 = P^{\mathrm{T}}LL^{\mathrm{T}}P \quad (14)$$

So $J(P)$ can be further written as follows:

$$J(P) = \arg\min P^{\mathrm{T}}LL^{\mathrm{T}}P + \lambda_2(\alpha P^{\mathrm{T}}XL_bX^{\mathrm{T}}P - (1-\alpha)P^{\mathrm{T}}XL_wX^{\mathrm{T}}P) \quad (15)$$

By means of Lagrangian multiplier method, the projection matrix P can be constructed by the eigenvectors of $LL^{\mathrm{T}} + X(\alpha L_b - (1-\alpha)L_w)X^{\mathrm{T}}$ associated with the first d largest eigenvalues $p_1, p_2, \ldots, p_d$, i.e., P can be constructed as $A = (p_1, p_2, \ldots, p_d)$.

D. The Algorithm

The detail algorithm for DR-SSL is listed as follows:
Step1.Initialize P,$P = P_0$ and $t = 0$
Step2.While not convergent
Step3.Update the coefficient matrix S using Eq.(12)
Step4.Compute the eigenvectors $(p_1, p_2, \ldots, p_d)$ of Eq.(15) associated with the first d largest eigenvalues ,then $A = (p_1, p_2, \ldots, p_d)$
Step5.$t = t + 1$
Step6.End While
Step7.Feature extraction:$y_i = P^{\mathrm{T}}x_i$

One issue that deserves attention is the convergency. We use the reduction of S and P to check the convergence of DR-SSL. More specifically, let $S(t-1)$ and $S(t)$ be the S at the $(t-1)-th$ and $t-th$ iteration,respectively. let $P(t-1)$ and $P(t)$ be the P at the $(t-1)-th$ and $t-th$ iteration,respectively. The convergence of this algorithm can be judged by whether it can satisfy the following inequity.

$$\|S(t-1) - S(t)\|^2 + \|P(t-1) - P(t)\|^2 < \xi \quad (16)$$

where ξ is a small positive number.

IV. Experimental Results

In this section, we conduct extensive experiments on three public microarray data sets to evaluate the performance of the proposed DR-SSL algorithm. For comparison, we also present the results of two competing dimensionality reduction algorithms,i.e.,principal component analysis (PCA) and locality preserving projections (LPP).

A. Microarray Data Sets

In this paper, we use three public available microarray datasets to test the proposed method. The details of the data sets used in our experiments are summarized as follows:

The Central Nervous System(CNS) dataset [18]: Only dataset C is used in this paper which contains 60 patient samples. Among them, 21 are survivors and 39 are failures. There are 7129 genes in the dataset.

The Colon dataset [19]: contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors and 22 normal. There are 7129 genes in the dataset.

The Leukemia dataset [20]: contains two types of acute leukemia: 47 acute lymphoblastic leukemia and 25 acute myeloid leukemia. There are 2000 genes in the dataset.

In this paper, we perform a preliminary selection of genes on the basis of the ratio of their between-groups to within-groups sum of squares. For a gene j, this ratio is

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_c I(y_i = c)(\overline{x}_{cj} - \overline{x}_{.j})^2}{\sum_i \sum_c I(y_i = c)(x_{ij} - \overline{x}_{cj})^2} \quad (17)$$

where $\overline{x}_{.j}$ denotes the average expression level of gene j across all samples and $\overline{x}_{cj}$ denotes the average expression level of gene j across samples belonging to class c . In our experiments, the 100 genes with the largest BSS/WSS ratios are selected for all the datasets.

Then for each gene j , we subtract it the mean measurement of the gene u_j and divide it by the standard deviation σ_j.

$$x_{ij} = \frac{x_{ij} - u_j}{\sigma_j} \tag{18}$$

After this transformation, the mean of each gene j will be zero, and the standard deviation will be one.

B. Performance Evaluations and Comparisons

In this paper, the values of the regularization parameters λ_1 and λ_2 are both set to 1 for the DR-SSL algorithm. For the discriminatively regularized term, the number of nearest neighbors is empirically set to 8, the value of the regularization parameter α is set to 0.1. For all the dimensionality reduction algorithms, we reduce the dimension to 30. After the dimensionality reduction process, we can apply a suitable classifier to classify the data. Different classifiers have been applied for gene expression data classification, including K-neighbor [21], Bayesian [22], and Support Vector Machines [23], etc. In this paper, we apply K-neighbor classifier with K=1 for its simplicity. To obtain reliable experimental results, we employ 5-fold cross validation , a statistical method of evaluating and comparing learning algorithms,to obtain measures of accuracy. Briefly, the data is first partitioned into five data sets of approximately equal size respectively. The training data set, which contains four parts of the subsets, is used to learn a classification model while the remaining subsets is used to validate the model. The procedure should be repeated five times and the performance is evaluated by the averaged recognition results over the five subsets.

In general, the recognition rates varies with the dimension of the feature subspace. Figure 1-3 shows the plots of recognition rates versus dimensionality reduction for the PCA, LPP, and DR-SSL.

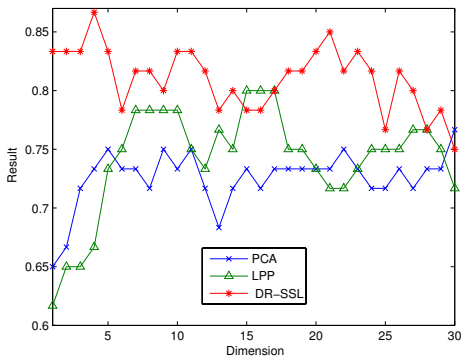

Fig. 1. Recognition results of dirrerent algorithms on the CNS dataset

The best result obtained in the optimal subspace and the corresponding dimensionality for each method are shown in Table I.

As can be seen, LPP has better performance than PCA on the CNS dataset and the Colon dataset. However, it gets the poorest results on the Leukemia dataset. Comparatively, the proposed DR-SSL method outperforms the other two

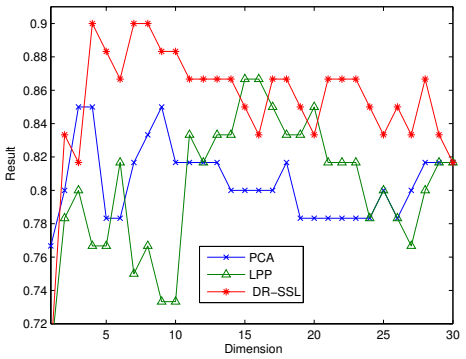

Fig. 2. Recognition results of dirrerent algorithms on the Colon dataset

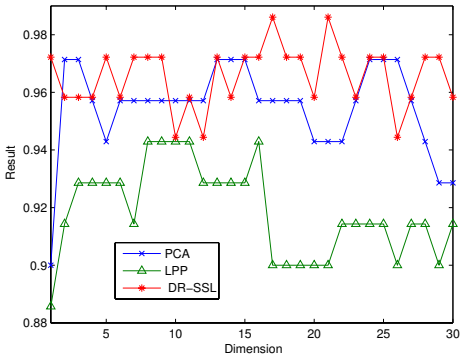

Fig. 3. Recognition results of dirrerent algorithms on the leukemia dataset

dimensionality reduction algorithms on all the gene-expression data sets. This is mainly because it is an effective sparse representation based learning algorithm. In addition, it can make full use of both the label information and the local manifold structure of given labeled samples to guide the dimensionality reduction process.

V. Conclusion

In this paper, an efficient dimensionality reduction algorithm called dscriminatively regularized sparse subspace learning(DR-SSL) is presented. The proposed method can make efficient use of the sparse representation to model the data. Moreover, it can make full use of both the label information and the local manifold manifold structure information of given labeled samples which are very helpful for dimensionality reduction and classification problems.The experiments on three gene-expression data sets show the effectiveness of the proposed algorithm.

TABLE I
The top recognition rates of different algorithms

Dataset	PCA	LLP	DR-SSL
CNS	76.67(30)	80(15)	86.67(4)
Colon	85(3)	86.67(15)	90(4)
Leukemia	97.14(2)	94.29(8)	98.61(17)

ACKNOWLEDGMENT

This work is partially supported by the Natural Science Foundation of Jiangsu Province of China (No. BK2010292).

REFERENCES

[1] T.R. Golub,D.K. Slonim ,P. Tamayo ,et al, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, vol.286,1999,pp. 531-537.
[2] I.T. Jolliffe,Principal Component Analysis.2nd edition. New York:Springer,2002.
[3] P. Comon, Independent Component Analysis-A New Concept?, Signal Process, vol.36,1994,pp.287-314.
[4] O.D. Richard,E.H. Peter and G.S. David, Pattern Classification,2nd edition.New York:Wiley-Interscience, 2000.
[5] S. Bicciato ,A. Luchini and C.D. Bello , PCA Disjoint Models for Multiclass Cancer Analysis using Gene Expression Data, Bioinformatics,vol.19,2003,pp.571-578.
[6] W. Liebermeister, Linear Modes of Gene Expression Determined by Independent Component Analysis, Bioinformatics, vol. 18,2002,pp. 51-60.
[7] X.W. Zhang ,Y.L. Yap ,D. Wei ,et al . Molecular Diagnosis of Human Cancer Type by Gene Expression Profiles and Independent Component Analysis . European Journal of Human Genetics , vol.5,2005,pp.46-56.
[8] S. Dudoit, J. Fridlyand, and T. P. Speed, Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data, Journal of the American Statistical Association, vol.97,2002,pp.77-87 .
[9] J.B.Tenenbaum,V.Silva and J.C.Langford,A global geometric framework for nonlinear dimensionality reduction,vol.290,2000,pp.2319-2323.
[10] S.T.Roweis,L.K.Saul,Nonlinear dimensionality reduction by locally linear embedding,vol.290,2000,pp.2323-2326.
[11] X. He and P. Niyogi, Locality Preserving Projections, Advances in Neural Information Processing Systems 16, , Cambridge,MIT Press, 2003.
[12] C. Shi and L.H. Chen, Feature Dimension Reduction for Microarray Data Analysis using Locally Linear Embedding. APBC,vol. 16, 2004,pp.1-7.
[13] G. Lee, C. Rodriguez and A. Madabhushi , Investigating the Efficacy of Nonlinear Dimensionality Reduction Schemes in Classifyi Gene- and Protein-Expression Studies, IEEE/ACM Transactions on Computational Biology and Bioinformatics,vol.5,2008, pp. -384.
[14] K. Huang and S. Aviyente, Sparse Representation for Signal Classification, Advances in Neural Information Processing Systems, vol.19,2006, pp. 609-616.
[15] S. Yan and H. Wang, Semi-Supervised Learning by Sparse Representation, SIAM International Conference on Data Mining, pp. 792-801 March,2009.
[16] John Wright, Yi Ma, Julien Mairal, et at, Sparse Representation For Computer Vision and Pattern Recognition .Proceedings of International Conference on Computer Vision and Pattern Recognition, vol.98,2010, pp. 1031-1044.
[17] H. Xue, S.C. Chen, Q. Yang, Discriminatively Regularized Least-Squares Classification. Pattern Recognition,vol.42,2009,pp. 93-104.
[18] S. Pomeroy , P. Tamayo and M. Gaasenbeek, et al, Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature,vol.415,2002,pp.436-442.
[19] T. R. Golub, D. K. Slonim and P. T., et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, 1999,Science,vol.286, pp.531-537 .
[20] U. Alon and N. Bkraai and D.A. Notterman, et al,Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences,vol.96,1999 ,pp.6745-6750.
[21] S. Deegalla and H. Bostrom,Classification of microarrays with kNN: comparison of dimensionality reduction methods, Lecture Notes in Computer Science,vol.4881,2007,pp.800-809.
[22] P. Helman ,R. Veroff and S.R. Atlas , et al,A Bayesian network classification methodology for gene expression data, Journal of Computational Biology,vol.11,2004,pp.581-615.
[23] T.S.Furey,N.Cristianini and N.Duffy,et al, Support vector machines classification and validation of cancer tissue samples using microarray expression data,Bioinformatics,vol.16,2000, pp.906-914.