Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 5490513, 11 pages
https://doi.org/10.1155/2018/5490513

Research Article

Feature Genes Selection Using Supervised Locally Linear Embedding and Correlation Coefficient for Microarray Classification

Jiucheng Xu,[1,2] **Huiyu Mu**,[1] **Yun Wang**,[1] **and Fangzhou Huang**[1]

[1]*College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China*
[2]*Engineering Technology Research Center for Computing Intelligence and Data Mining, Henan Province 453007, China*

Correspondence should be addressed to Jiucheng Xu; xjc@htu.cn

Received 27 September 2017; Revised 17 December 2017; Accepted 21 December 2017; Published 31 January 2018

Academic Editor: Xiaoqi Zheng

The selection of feature genes with high recognition ability from the gene expression profiles has gained great significance in biology. However, most of the existing methods have a high time complexity and poor classification performance. Motivated by this, an effective feature selection method, called supervised locally linear embedding and Spearman's rank correlation coefficient (SLLE-SC^2), is proposed which is based on the concept of locally linear embedding and correlation coefficient algorithms. Supervised locally linear embedding takes into account class label information and improves the classification performance. Furthermore, Spearman's rank correlation coefficient is used to remove the coexpression genes. The experiment results obtained on four public tumor microarray datasets illustrate that our method is valid and feasible.

1. Introduction

Cancer develops through either a series of genetic events or external influential factors that cause differential gene expression profile in the cancerous cells. The DNA microarray technology is pervasively used in the area of genomic research for diagnosing cancers [1]. Since the number of genes is typically larger than the number of samples, classification of microarray data is subjected to "the curse of dimensionality." However, only a small number of genes are required in cancer diagnosis whereas the search space can be huge. Feature selection is an important step to reduce both dimension and redundancy (there is some obvious inaccuracy of gene expression in the experiment to obtain the gene expression data) of gene expression data during the classification process. According to the literature [2], the selection of feature genes methods is usually more important than developing classifier in the genomic data analysis. Therefore, how to choose the feature genes in gene expression profile effectively is the key point of bioinformatics study at present.

When mining in high-dimensional data, "the curse of dimensionality" is one of the major difficulties to overcome. The aim of feature selection is to reduce computational complexity while some desired inherent information of the data is conserved [3, 4]. Manifold learning is an ideal tool for machine learning that discovers the structure of high-dimensional data and gives better understanding of the data [5]. The representative of such methods comprises locally linear embedding (LLE), isometric mapping (Isomap), Laplacian eigenmaps (LE), and local tangent space alignment (LTSA) [6], and so on. In between, LLE is one of the most noted manifold learning methods and widely used in spectral analysis [7], edit propagation [8], fault detection [9, 10], image recognition [11, 12], and so on.

Subsequently, various improved LLE methods are designed to enhance the performance. Lai et al. [31] proposed a unified sparse learning framework by introducing the sparsity or L1-norm learning, which further extended the LLE-based methods to sparse cases. Theoretical connections between the orthogonal neighborhood preserving projection and the proposed sparse linear embedding are discovered. The ideal sparse embedding derived from the proposed framework is computed by iterating the modified elastic net and singular value decomposition. Cheng et al. [32]

depended on the incremental locally linear embedding (ILLE) to improve the performance of fault-diagnosis for a satellite with high-dimensional telemetry data. Similarity, Liu et al. [33] put forward an incremental supervised LLE (I-SLLE) method for submersible plunger pump fault detection. In the I-SLLE algorithm, block matrix decomposition strategy is used to deal with out-of-sample data, while a part of original low-dimensional coordinates is also renovated, above which an iterative method is proposed to update all the dataset for improving the accuracy.

LLE has the advantage of global optimal solution of parsing without iteration. The low-dimensional embedding of calculation is summarized as sparse matrix eigenvalue calculation. So the complexity of calculation is relatively small. However, LLE mainly has the disadvantage of low self-learning ability and ignores the discriminant information. It is difficult to accurately capture the patterns on data and this could not gain higher effectiveness. Furthermore, the purpose of feature selection is to project the original data into a subspace with the following characteristics: the samples in the intraclass as close as possible and the samples in interclass far away from each other in the subspace. As mentioned before, feature genes selection distinguishes the pathogenic genes from normal genes. To solve this problem, de Ridder et al. extended the concept of LLE to multiple manifolds and proposed a supervised locally linear embedding (SLLE) algorithm which has been demonstrated to be a suitable feature for genes selection [34]. The dissimilarity between samples from different classes can be measured by metric function. It is commonly believed that the neighborhood of a sample in one class should consist of samples belonging to the same class. In the SLLE method, by taking into account class label information, the distance of interclass is larger than the Euclidean distance by adding a parameter to the pairs of points belonging to different classes. Otherwise, it remains as the Euclidean distance.

Feature selection reduces the dimension of feature and ensures the integrity of original dataset. It can improve the efficiency of data mining and dig out the results which are basically identical to the original dataset. More broadly, it is the problem of "the curse of the dimension." However, the major consideration of SLLE is the relationship between the attributes and categories. The way to judge if an attribute is redundant is based on whether the attribute affects information discrimination of the class label. That is to say, SLLE remains not fully considered by the relationship between the attributes. In practice, it is not independent between the attributes, and there is a certain correlation between them. For instance, the dressing index and temperature are usually related: a high temperature means a low clothing index; otherwise the opposite occurs. It is inevitable that data redundancy will be caused by placing a large number of associated attributes in the reduction result. Correlation coefficient reflects the coexpression relationship between genes. The two genes are considered as coexpression when their correlation coefficient value is greater than a certain threshold; thus it can be removed [35, 36].

In order to solve the problem of poor classification performance in tumor classification, a novel feature genes selection method, called supervised locally linear embedding and Spearman's rank correlation coefficient (SLLE-SC^2), is put forward in this paper. Supervised LLE algorithm, by taking into account class label information, is utilized to delete redundant genes. Meanwhile, Spearman's rank correlation coefficient is used to remove the coexpression genes. We also show biological investigation of the selected genes. Finally, we compared the performance of various classifiers based on the selected feature genes datasets. Results show that the SLLE-SC^2 method selects a small set of nonredundant disease related genes with high specificity and achieves better efficiently compared with other related methods.

2. Research Methodology

2.1. Locally Linear Embedding. LLE approximates the input data with a low-dimensional surface and reduces its dimensionality by learning a mapping to the surface [37]. It first finds a group of the nearest neighbors of each data point. Then it calculates a set of weights for each data point that wonderfully describe the point as a linear combination of its neighbors. Finally, it finds the low-dimensional embedding of points by using an eigenvector-based optimization technique; thus each point is also described with the same linear combination of its neighbors. LLE is designed to establish such a feature mapping: low-dimensional embedding maintains the same local neighborhood relationship in high-dimensional space. It gets the corresponding low-dimensional embedding from the nearest neighbor graph of geometric properties in high-dimensional space under certain conditions. In fact, LLE considers the point of nearest neighbors, rather than distant points.

(a) Assigning Neighbors to Each Data Point. To find a group of nearest neighbors, LLE adopts k nearest neighbors (i.e., Euclidean distance) standard. Let $X = \{x_1, \ldots, x_N\}$ be a given dataset of N points, $x_i \in R^D$; Euclidean distance is adopted to calculate the distance between samples D_{ij} $(i, j \in 1, 2, \ldots, n)$ and find refactoring neighborhood of the k nearest neighbors for each data point.

(b) Computing the Weights Best Linearly Reconstructed from Its Neighbors. LLE computes the barycentric coordinates of a point X_i based on its neighbors X_j. The original point is reconstructed by a linear combination and given by the weight matrix W_{ij} of its neighbors. Reconstruction errors are measured by the cost function

$$\begin{aligned}\varepsilon_i(W) &= \min \left\| X_i - \sum_{j=1}^{k} W_{ij} X_j \right\|^2 \\ &= \min \left\| \sum_j W_{ij} \left(X_i - X_j\right) \right\|^2 = \sum_{j,k} W_j W_k G_{jk},\end{aligned} \tag{1}$$

where ε_i is reconstruction error; G_{jk} is a local graham matrix.

$$G_{jk} = \left(X_i - X_j\right)^T \left(X_i - X_k\right), \tag{2}$$

where G_{jk} is a positive definite symmetric matrix. Equation (1) is a constrained least squares problem, and it is minimized under two constraints:

$$W_{ij} = \begin{cases} 1 & X_j \text{ is a neighbor of } X_i \\ 0 & \text{the others} \end{cases} \tag{3}$$

$$\sum_j W_{ij} = 1 \tag{4}$$

in which, (3) is a constraint of coefficient. That is to say, each data point is reconstructed only from its neighbors. Equation (4) means the sum of every row of weight matrix equals 1. Thus (1) is rewritten as constrained optimization form:

$$\begin{aligned} \min \quad & \sum_{j,k} W_j W_k G_{jk} \\ \text{s.t.} \quad & \sum_j W_{ij} = 1. \end{aligned} \tag{5}$$

Equation (5) is calculated by Lagrange multiplier approach. As G_{jk} is positive definite symmetric matrix, the inverse of the matrix G_{jk} exists. The optimal weight is calculated by

$$W_j = \frac{\sum_k {G_{jk}}^{-1}}{\sum_{lm} {G_{lm}}^{-1}}. \tag{6}$$

(c) Computing the Low-Dimensional Embedding Vector Best Reconstructed and Finding the Smallest Eigenmodes of the Sparse Symmetric Matric. Each point X_i in the high-dimensional space is mapped onto a point Y_i in the low-dimensional space. The low-dimensional space Y is calculated by the following function:

$$\varepsilon(Y) = \min \left\| Y_i - \sum_{j=1}^{k} W_{ij} Y_j \right\|^2 = \min \sum_{j,k} M_{ij} \left(Y_i \cdot Y_j \right). \tag{7}$$

Cost function (7) is based on locally linear reconstruction errors, in which $(Y_i \cdot Y_j)$ is inner product; M_{ij} is a sparse $N \times N$ matrix (N being the number of data points).

$$M_{ij} = \delta_{ij} - W_{ij} - W_{ji} + \sum_k W_{ki} W_{kj}, \tag{8}$$

where M_{ij} is a positive definite symmetric matrix. Equation (7) is a minimization problem. Significantly, we can translate Y_i to any position without affecting the reconstruction error. Thus a constraint is added to eliminate the translational degree of freedom in (7). It requires all the center of low-dimensional embedding Y_i at the origin. Namely,

$$\sum_i Y_i = 0. \tag{9}$$

In order to eliminate the rotational and proportion degree of freedom, we add a constraint of unit covariance:

$$\frac{1}{n} * \sum_i Y_i Y_i^T = I; \tag{10}$$

then (7) is regarded as a constrained optimization problem.

$$\begin{aligned} \min \quad & \sum_{j,k} M_{ij} \left(Y_i \cdot Y_j \right) \\ \text{s.t.} \quad & \sum_i Y_i = 0 \\ & \frac{1}{n} \cdot \sum_i Y_i Y_i^T = I. \end{aligned} \tag{11}$$

Equation (11) can be solved in multiple ways. One of the most effective methods is calculating cost matrix M relatively minimum $d + 1$ eigenvalue with its eigenvector which is optimized by using Lagrange multipliers. Notice that eigenvalue with its eigenvector is a fully 1 vector; it represents translation degrees of freedom corresponding to the 0 eigenvalue and requires removing. The retained d eigenvectors formed the output of LLE.

2.2. Supervised Locally Linear Embedding. LLE is an unsupervised manifold feature selection algorithm, which ignores the discriminant information of data. In order to improve the classification capability of LLE, discriminant information is assembled in the cost function of LLE (i.e., SLLE). SLLE is based on assumptions of the distance of data point from the same class less than the data point from the different classes and adds the discriminant information to the interclass distance. One of the solutions is to increase the Euclidean distance by adding a constant to the pairs of points from different classes, and the distance of data points from the same class is kept.

In a given set $X = \{x_1, x_2, \ldots x_n\}$, the distance metric is defined as

$$\Delta'(i, j) = \Delta(i, j) + \lambda \cdot \max\left(\{\Delta(i, j)\}\right) \cdot \delta_{ij}, \tag{12}$$

where $\Delta(i, j)$ is the Euclidean distance between x_i and x_j. $\lambda \in [0, 1]$ is a tunable parameter. $\max(\{\Delta(i, j)\})$ is the maximum of Euclidean distance set $\{\Delta(i, j)\}$. δ_{ij} is equal to 0 or 1 which is used to indicate whether the points belong to the same class; if x_i and x_j belong to the same class, $\delta_{ij} = 0$; otherwise, $\delta_{ij} = 1$.

It is worth noting that when $\lambda = 0$, the SLLE is turned into the original unsupervised LLE; when $\lambda = 1$, it is the supervised LLE; otherwise, it is a semisupervised LLE.

2.3. Spearman's Rank Correlation Coefficient. The relationship between attributes and categories relates to the feature reduction effectiveness and classification accuracy. Similarity, this connection is similar for attributes. In general, the connection between attributes is measured by correlation coefficient. The conventional measures of correlation coefficient are bivariate normal distribution, chi-square test for independence and rank correlation coefficient, and so on. Among them, Spearman's rank correlation coefficient is a nonparametric measure of rank correlation (statistical dependence between the ranking of two variables). It assesses how well is the relationship between two variables which is described with the monotonic function.

Table 1: Example sample X.

Sample	a_1	a_2
$x1$	0.7	0.9
$x2$	0.3	0.3
$x3$	0.5	0.4
$x4$	0.2	0.1
$x5$	0.8	0.7

Table 2: The rank sequences R_a and S_a.

Sample	a_1	R_a	a_2	S_a
$x1$	0.7	4	0.9	5
$x2$	0.3	2	0.3	2
$x3$	0.5	3	0.4	3
$x4$	0.2	1	0.1	1
$x5$	0.8	5	0.7	4

In a given dataset sample $X = \{x_1, x_2, \ldots, x_n\}$, attribute $C = \{a_1, a_2, \ldots, a_n\}$. The sequence A_i in sample X, relatively, attribute a_i with its attribute value is $A_i = \{x_1 = v_1, x_2 = v_2, \ldots, x_n = v_n\}$. Then the sequence A_i is sorted in descending order with rank for each sample (i.e., sample of the smallest attribute value with rank of 1, sample of the largest attribute value with rank of $|X|$; the rank takes an average with the attribute with the same value). Next, according to original sample order, we reorder the new rank sequence $A'_i = \{x_1 = v'_1, x_2 = v'_2, \ldots, x_n = v'_n\}$.

For the attributes a_i, a_j of sample k, its rank sequence is R_k and S_k, respectively. So we obtain $|U|$ pairs rank combination $(R_1, S_1), (R_2, S_2), \ldots, (R|U|, S|U|)$. Spearman's rank correlation coefficient of attributes a_i, a_j is defined as

$$r_{ij} = \left| r\left(a_i, a_j\right) \right| = \frac{\sum_{k=1}^{|U|} \left[\left(R_k - \overline{R}\right)\left(S_k - \overline{S}\right)\right]}{\sqrt{\sum_{k=1}^{|U|} \left[\left(R_k - \overline{R}\right)^2 \left(S_k - \overline{S}\right)^2\right]}}, \quad (13)$$

where $\overline{R} = 1/|U| * \sum_{k=1}^{|U|} R_i$, $\overline{S} = 1/|U| * \sum_{k=1}^{|U|} S_i$. Correlation coefficient r_{ij} meets the following properties:

(1) $0 \le r_{ij} \le 1$.

(2)r_{ij} always gives an answer between 0 and 1. The numbers in between are like a scale, where 1 indicates a very strong link and 0 indicates no link.

For more detailed instructions, we use an example to work out r_{ij} in Table 1. Sample $X = \{x_1, x_2, x_3, x_4, x_5\}$; attribute $C = \{a_1, a_2\}$.

(1) Obtain the sequence A_1 in sample X; relatively attribute a_1 with its attribute value is $A_1 = \{x_1 = 0.7, x_2 = 0.3, x_3 = 0.5, x_4 = 0.2, x_5 = 0.8\}$.

(2) The sequence A_1 is sorted in descending order with rank for each sample. Thus we obtain an ordered sequence of attribute $\{x_4, x_2, x_3, x_1, x_5\}$ and rank sequence $\{x_4 = 1, x_2 = 2, x_3 = 3, x_1 = 4, x_5 = 5\}$.

(3) According to original sample order, we reorder the new rank sequence $R_a = \{x_1 = 4, x_2 = 2, x_3 = 3, x_4 = 1, x_5 = 5\}$.

(4) In the same way, the rank sequence S_a in sample X relative attribute a_2 with its attribute value is $A_2 = \{x_1 = 5, x_2 = 2, x_3 = 3, x_4 = 1, x_5 = 4\}$.

(5) The rank sequences R_a and S_a in sample X relatively attributed to its attribute value are shown in Table 2.

(6) Finally, according to (13), Spearman's rank correlation coefficient is 0.9 for this set of data.

2.4. Feature Genes Selection Using Supervised Locally Linear Embedding and Correlation Coefficient. Microarray data often contain redundant and noise features. These features could lead to poor classification performance and overfitting problems. Meanwhile, the gene expression data are in high-dimension and the number of feature gene datasets is very small which leads to the calculation falling into local optima and being computationally expensive. The key technique is to find a new feature genes selection method which can provide understanding and insight into tumor related cellular processes.

SLLE (by taking into account class label information) finds an ideal low-dimensional manifold of mapping for separating the intraclass and interclass. However, the main consideration of supervised algorithm is the relationship between the attributes and categories. That is to say, supervised learning algorithm is not fully considering the relationship between the attributes. In practice, the relationship between the attributes affects the reduction results and classification accuracy. It is inevitable that data redundancy will be caused by placing a large number of associated attributes in the reduction result. In general, the connection between attributes can be measured by correlation coefficient. Correlation coefficient reflects the coexpression relationship between genes. The two genes are considered as coexpression when their value of correlation coefficient is greater than a certain threshold; thus they are removed in feature genes selection. Spearman's rank correlation coefficient is a nonparametric measure of rank correlation (statistical dependence between the ranking of two variables).

Therefore we propose an effective SLLE-SC^2 method for the selection of feature genes. Firstly, SLLE is used for reduction, mapping into the original data in a new feature space. Then considering the relationship between the attributes in the new feature space, Spearman's rank correlation coefficient is used for feature selection. Specifically, the PCA is used to compute the contribution of attributes, respectively, in the new feature space. Spearman's rank correlation coefficient is used to compute the maximum contribution of attribute and other attributes, respectively. If the value of correlation coefficient between attributes is greater than or equal to a preset threshold, the attribute is removed. Then loop is over the other attributes. SLLE method description is shown in Algorithm 1. Spearman's rank correlation coefficient method description is shown in Algorithm 2. Feature genes selection using SLLE-SC^2 method description is shown in Algorithm 3.

Input: $X = \{U, C, V\}$
Output: Reduction set
Step 1. For each data X_i in high-dimensional space, find the k nearest points X_j in terms of the Euclidean distance;
Step 2. Calculate the local reconstructed weight matrix for each sample point. The current sample point is expressed by the k nearest neighboring points and gets the weight matrix, the error function is defined as: $W_j = \sum_k G_{jk}^{-1} / \sum_{lm} G_{lm}^{-1}$;
Step 3. According to the weight W for the sample point X_i and neighboring point X_j in the high-dimensional space. Then the embedding space in low-dimension is calculated. The weight is fixed to a constrained optimization problem;
Step 4. By minimizing the loss functions to get the corresponding weight matrix and reconstructed coordinates. The retained d eigenvectors are formed the output of LLE algorithm;
Step 5. Return reduction set.

Algorithm 1: Supervised locally linear embedding method description.

Input: $X = \{U, C, V\}, a_i, a_j$
Output: Correlation coefficient r_{ij}
Step 1. Obtain the sequence A_i, A_j in sample X relatively attribute a_i, a_j with its attribute value;
Step 2. The sequenceA_i, A_j is sorted in descending with rank for each sample. The rank takes an average when the attributes with the same value;
Step 3. The new rank sequence R_k, S_k are obtained according to original sample order for A_i, A_j;
Step 4. for $k = 1$ to $|X|$ do;
calculate the average rank sequence $\overline{R}, \overline{S}$ for rank sequence R_k, S_k;
Step 5. for $k = 1$ to $|X|$ do;
calculate $r(a_i, a_j)$;
Step 6. Calculate r_{ij};
Step 7. Return correlation coefficient r_{ij}.

Algorithm 2: Spearman's rank correlation coefficient method description.

3. Experiments and Results

3.1. Data Preparation. In order to verify the effectiveness of the proposed algorithm, four public tumor microarray datasets are used for making simulation experiment. Particularly, all of them represent binary classification tasks. Detailed information of datasets is shown in Table 3.

All numerical experiments are performed on a personal computer with 3.1 GHz AMD Athlon (tm) II and 4-G-byte memory. This computer runs Windows 7, with Matlab-R2010 and Weka-3.9.0.

3.2. Results and Analysis. In order to illustrate the reliability and comparability of tumor microarray datasets, we do experiment many times for the average. Experiments use 10-fold cross-validation. Specifically, based on preliminary tuning experiment, we set the nearest neighbors k for each data point as 5 for SLLE-SC^2 method.

PCA algorithm is used for analyzing four tumor microarray datasets before SLLE-SC^2 method test and drawing Pareto diagram (i.e., the information in genomic datasets) of the principal components explained variance for each dataset (blue curve said before the information content of total n genes in Figure 1). The results are shown in Figures 1(a), 1(b), 1(c), and 1(d).

The accumulation contribution rate of most datasets (except lung dataset) reaches more than 90 percent when the principal components of datasets are 50 (see Figure 1). It illustrates that gene expression profile datasets contain a large amount of redundancy (i.e., irrelevant and confounding factors) and the number of feature genes are a small part, so it is necessary to remove the redundancy genes.

The classification accuracies vary with the threshold λ of correlation coefficient; threshold λ takes values from 0 to 1 with step 0.1. For each value of the threshold, SLLE-SC^2 obtains a subset of genes based on the average classification accuracies of SVM classifier. Experiments use 10-fold cross-validation. Classification accuracies with threshold λ are shown in Figure 2.

All the results show a common rule that the classification accuracies based on SVM increase with the value of threshold λ at first, arrive at a peak value, and then are stable relatively. It is easier for the classification of leukemia data than the others. When λ is among 0 to 0.3, classification accuracy increases faster, and when $\lambda > 0.3$, classification accuracy is relatively stable. It conforms to the actual performance. When λ is large, it has less strict requirements for removing redundant attributes, so the classification accuracy has no obvious change. Instead, when λ is small, it has many strict requirements for removing redundant attributes and causes

Input: Data set $X = \{U, C, V\}$
Output: Feature genes set F
Step 1. F = null; flag set $flag$ = null; // the initial state is empty;
Step 2. SLLE(X) // using Algorithm 1 for feature genes selection;
Step 3. for $i = 1, 2, \ldots, n$ do;
calculate the contribution $Con(a_i)$ of attributes a_i respectively by PCA, where attribute $a_i \in C\text{-}(F \cup flag)$;
if $Con(a_k) = \max Con(a_i)$, output attribute a_k;
end for
Step 4. for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, n$ do
calculate correlation coefficient r_{jk} for attribute a_j and a_k by Algorithm 2, where $\forall a_j \in a_i$;
if $r_{jk} \geq \lambda$ then
$flag = flag \cup a_k$;
go to Step 3;
end if
if $red = red \cup a_k$ then
go to Step 5;
end for
Step 5. Return red.

Algorithm 3: SLLE-SC^2 method description.

Table 3: Experiment dataset.

Dataset	Number of features	Classes	Number of instances
Leukemia	7129	ALL (47), AML (25)	72
Colon	2000	Tumor (40), normal (22)	62
Lung	12600	Tumor (186), normal (17)	203
Prostate	12600	Tumor (52), normal (50)	102

Table 4: The results of various performance metrics.

Dataset	Acc	TPR	TNR	F-measure	G-mean	AUC
Leukemia	0.997	0.86	0.882	0.909	0.895	0.914
Colon	0.948	0.89	0.877	0.85	0.911	0.864
Lung	0.942	0.793	0.827	0.842	0.837	0.858
Prostate	0.968	0.863	0.873	0.858	0.848	0.904

Table 5: Classification performance of leukemia data.

Classifiers	SLLE-SC^2	LE	LLE	SLLE	SC^2
SVM	99.7	85.9	92.3	97.4	85.2
C4.5	97.4	84.6	87.5	93.2	81.1
Naive Bayes	98.8	79.7	82.7	99.1	74.4
kNN	100	93.2	92.3	98.8	83.6

Table 6: Classification performance of colon data.

Classifiers	SLLE-SC^2	LE	LLE	SLLE	SC^2
SVM	94.8	81.2	89.1	91.9	80.5
C4.5	93.1	83.3	87.5	92.6	77.2
Naive Bayes	92.7	95.6	85.7	89.6	73.4
kNN	94.6	79.3	89.3	92.7	78.7

Table 7: Classification performance of lung data.

Classifiers	SLLE-SC^2	LE	LLE	SLLE	SC^2
SVM	94.2	80.5	87.1	91.6	80.6
C4.5	92.7	79.2	87.5	92.3	79.1
Naive Bayes	94.8	78.1	90.7	94.7	80.5
kNN	89.9	81.4	87.3	89.6	75.8

Table 8: Classification performance of prostate data.

Classifiers	SLLE-SC^2	LE	LLE	SLLE	SC^2
SVM	97.9	85.5	88.2	96.9	79.5
C4.5	95.4	81.3	90.7	95.3	81.1
Naive Bayes	94.8	79.1	86.7	89.9	73.7
kNN	96.8	82.9	87.3	97.8	74.8

decline of classification accuracy. For overall consideration, the threshold of correlation coefficient is 0.3.

For convenient description, the datasets in Table 3 are divided into positive and negative: positive ones are ALL

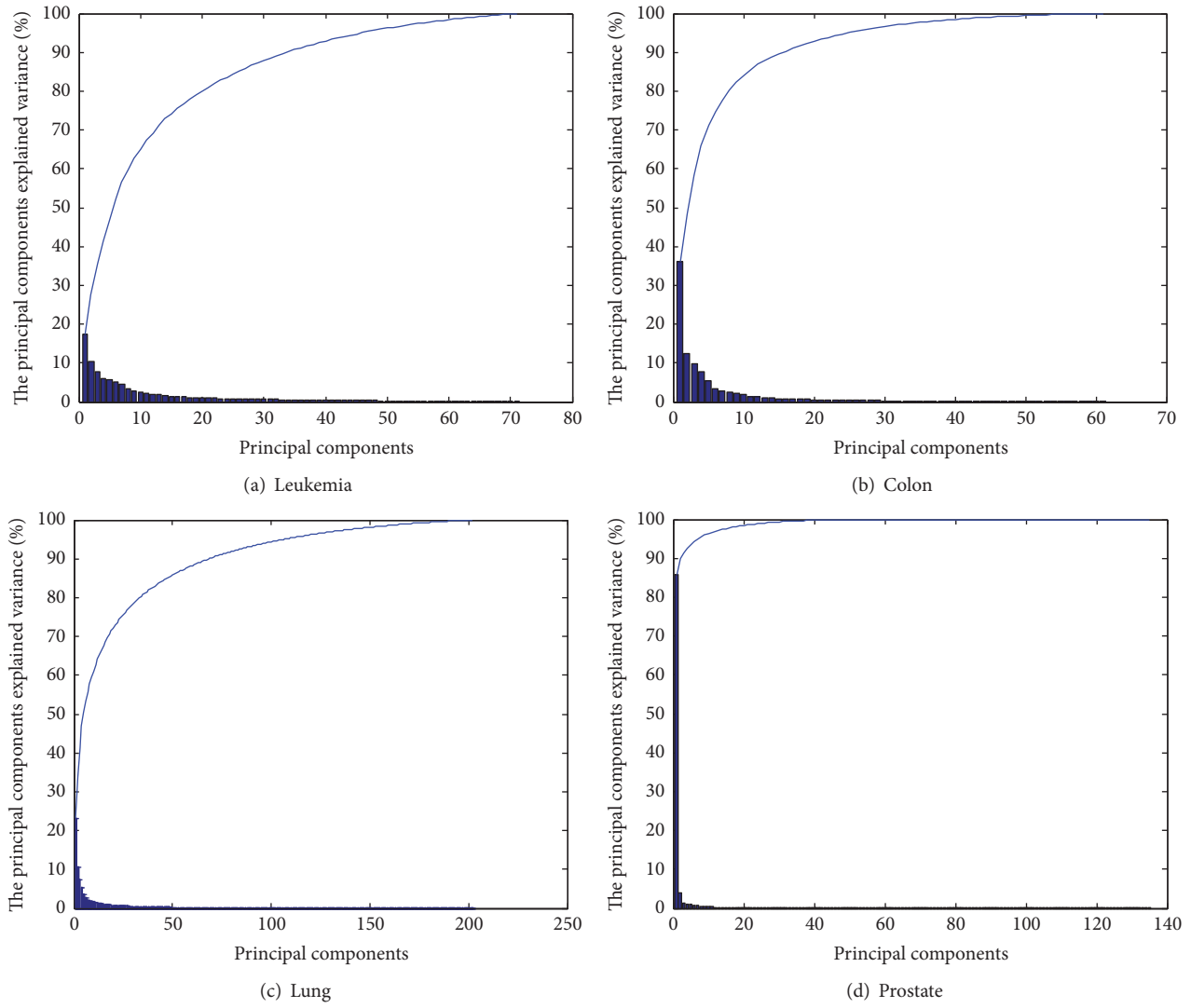

Figure 1: Pareto diagram of the principal components explained variance.

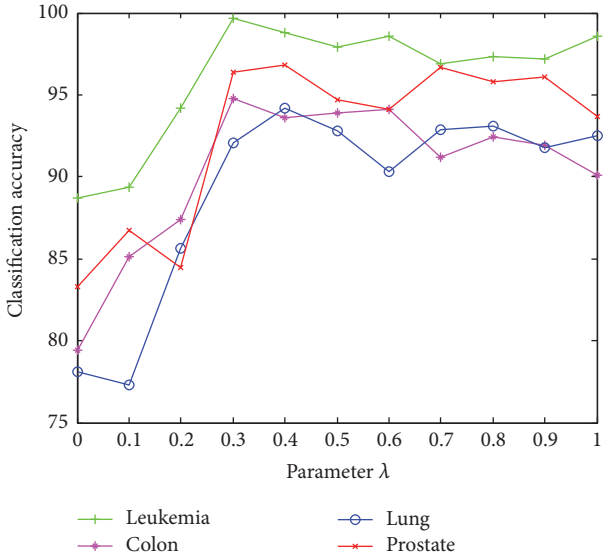

Figure 2: Classification accuracies with threshold λ.

Table 9: The number of feature genes and classification results.

Method	Leukemia	Colon	Lung	Prostate
IGA-FBFE [13]	94.20 (35)	90.09 (30)	91.23 (80)	88.12 (50)
BQPSO [14]	100 (7)	92.52 (11)	99.96 (9)	99.25 (10)
CAGC [15]	95.3 (866)	91.9 (135)	—	68.9 (3071)
ILASSO [16]	98.61 (14)	90.32 (4)	100 (7)	96.08 (9)
RT-PLSDA [17]	94.12 (9)	—	97.99 (4)	91.18 (18)
MAHP [18]	92.78 (5)	83.47 (5)	88.77 (5)	—
SU [19]	100 (6)	83.87 (4)	100 (3)	93.14 (4)
DRF0-CFS [20]	91.18 (13)	90.0 (10)	98.66 (17)	85.29 (113)
IG-SGA [21]	97.06 (3)	85.48 (60)	—	100 (26)
SLLE-SC^2	99.7 (5)	95.4 (4)	94.8 (3)	97.3 (5)

Table 10: Biological significance of leukemia data.

Index	Gene selection	Description
1834	M23197	CD33 antigen (differentiation antigen) [22]
1882	M27891	CST3 cystatin C [22]
3847	U82759	GB DEF = homeodomain protein HoxA9 mRNA [23]
4847	X95735	Zyxin [23]
6041	L09209	APLP2 [22]

Note. Index denotes the serial number of the selected genes in the original data.

Table 11: Biological significance of colon data.

Index	Gene selection	Description
792	R88740	ATP synthase coupling factor 6, mitochondrial precursor [24]
1346	T62947	60S ribosomal protein l24 (*Arabidopsis thaliana*) [24]
1400	M59040	Human cell adhesion molecule (CD44) mRNA [25]
1772	H08393	Collagen alpha 2(xi) chain (*H. sapiens*) [24]

Note. Index denotes the serial number of the selected genes in the original data.

Table 12: Biological significance of lung data.

Index	Gene selection	Description
4336	AL050224	*Homosapiens* mRNA; cDNA DKFZp586L2123 [26]
7765	X05323	Human MOX2 gene for OX-2 membrane glycoprotein, exon 1, and joined CDS [27]
8537	AJ011497	*Homosapiens* mRNA for claudin-7 [27]

Note. Index denotes the serial number of the selected genes in the original data.

and tumor, negative ones are AML and normal, respectively. TP and TN mean the number of right positive and negative examples; FN and FP denote the number of misclassified positive and negative examples, respectively.

$$\begin{aligned}
\text{Acc} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\
\text{TPR} = \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{TNR} &= \frac{\text{TN}}{\text{TN} + \text{FP}} \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
F\text{-measure} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\
G\text{-mean} &= (\text{TPR} \times \text{TNR})^{1/2}.
\end{aligned} \tag{14}$$

(Note: Acc: overall accuracy; TPR: true positive rate; TNR: true negative rate; FPR: false positive rate; AUC: area under the receiver operating characteristic curve—it is the area below the ROC curve that depicts the performance of a classifier using the FPR and TPR pairs [38])

To present the superiority of SLLE-SC^2 method, we evaluate it in comparison with that of SVM classification approaches and adopt the procedure of 10-fold cross-validation. Table 4 reports the results of various performance metrics on four biomedicine datasets.

From the results in Table 4, our method with that of SVM classification results in better performance. Lung data acquire the lowest Acc value on all datasets. In terms of six important performance metrics, leukemia data obtain the largest Acc value, as well as taking the first place on four datasets for TNR, F-measure, and AUC criteria, respectively. In general, SLLE-SC^2 algorithm gets a better effect in the aspects of high-dimensional and imbalanced classification tasks.

(i) Classification Performance of Feature Genes. Laplacian eigenmaps (LE), locally linear embedding (LLE), supervised locally linear embedding (SLLE), and Spearman's rank correlation coefficient (SC^2) are implemented as competing methods to compare with the proposed SLLE-SC^2 method.

Table 13: Biological significance of prostate data.

Index	Gene selection	Description
5890	AJ001625	*Homosapiens* mRNA for Pex3 protein [28]
6462	M11433	Human cellular retinol-binding protein mRNA, complete cds [29]
9172	AI207842	Ao89h09.x1 *Homosapiens* cDNA, 3 ends [30]
9850	M84526	Human adipsin/complement factor D mRNA, complete cds [29]
12495	M98539	Human prostaglandin D2 synthase gene, exon 7 [29]

Note. Index denotes the serial number of the selected genes in the original data.

The nearest neighbor k is 5 for LE, LLE, SLLE, and SLLE-SC^2. Four classifiers are implemented for classification including SVM, C4.5 (a classification algorithm of decision tree), Naive Bayes (naive Bayesian classification), and k-nearest neighbors (kNN). Experiments use 10-fold cross-validation; the results are shown in Tables 5–8.

Each result composes the classification accuracy of 20 independent outcomes in Tables 5–8. We see that SLLE-SC^2 gains the greatest average accuracy in four datasets. By averaging across four classifiers, SLLE-SC^2 obtains the top accuracy, with 100% (kNN classifier), 94.8% (Naive Bayes classifier), and 97.9% (SVM classifier) in the leukemia, lung, and prostate datasets, respectively. SC^2 achieves the worst performance, and its accuracy is much lower than that of SLLE-SC^2. SLLE by taking into account class label information gets much better classification performance.

(ii) Comparison of the Classification Effect with the Gene Selected by Different Methods. To verify classification effect with the gene selected by different methods, IGA-FBFE and other 9 feature selection methods are used for comparison in gene expression profiles. Lib-SVM classifier in Weka tool is used for simulation experiment. The number of feature genes and classification results are shown in Table 9.

As shown in Table 9, in terms of the number of selected genes, the difference between methods can be clearly found. For some methods, the number is as high as 60 (e.g., lung data with IGA-FBFE method) or even more, but for some methods the number is less than 10 (such as MAHP, SU, and SLLE-SC^2 methods). However, it is hard to do a further comparison of the selected genes for the listed methods, as the genes selected by the other methods are not offered.

As for the classification accuracies, our method produces the results of 99.7% and 5 selected genes for the leukemia data. The results are not inferior to most of the published works. Colon data get small number of selected genes and higher accuracy. For lung data, ILasso and SU methods obtain better classification than our method but failure in number of feature genes. For prostate data, though BQPSO and IG-SGA acquire higher accuracy 99.25% and 100%, respectively, the number of feature genes is more than ours. Clearly, SLLE-SC^2 cannot overcome all the existing methods. However, it can outperform some of the published methods and obtain a comparable result with most of the listed methods. Some of the methods produce high classification accuracy which use too large numbers of the selected genes in the classification (e.g., in prostate data, 26 genes are employed by IG-SGA method). However, such results may be difficult for a biological interpretation, all of which go to prove that our method selects the feature genes which have high classification ability and can reflect the structure of the data actuality. The small numbers of feature genes not only improve the running efficiency of the algorithm, but also can enhance the understanding of the microarray data.

(iii) Biological Significance. In order to validate the selected genes, Tables 10–13 summarize the index, gene, and description of the selected genes.

We search genes from the web of National Center for Biotechnology Information (NCBI) to further understand the selected genes (https://www.ncbi.nlm.nih.gov/). It can be seen that most of genes are closely associated with cancer as seen in Tables 10–13. Most of the selected genes are consistent with the results shown in the previous research [22–30]; for example, gene M23197 has been certified for targeted antibody therapy to make leukemia AML die [22], and the gene X95735 codes an LIM domain protein that is significant in cell adhesion of fibroblasts [23]. Gene AL050224 takes effect in the RNA polymerase and finds the overexpression in lung tissues [26]. Gene AJ011497 shows low-expression in MPM while showing high-expression in ADCA [27]. It is considered as a biomarker for the lung cancer. Gene M84526 codes another serine protease adipsin which is secreted by adipocytes into the bloodstream and functions as part of the alternative complement pathway of the innate immune system [29].

4. Conclusions

In this work, we explore the effects and benefits of SLLE-SC^2 in the context of feature selection from high-dimensional genomic data. Specifically, supervised LLE is used to remove redundant genes. Considering the relationship between the attributes, the coexpression relationship between genes is deleted by Spearman's rank correlation coefficient. Our results on four microarray datasets are very promising and supported by existing biological knowledge. The results of our experiments give insight into both predominance and inferior position of SLLE-SC^2 method and could represent a useful starting point to better understand the behavior of these techniques as well as the extent of their applicability to specific tumor problems. In more detail, we study genomic information to better understand pathogenesis of tumor and provide reference for the clinical treatment of tumor.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61370169, 61772176, and 61402153) and Key Project of Science and Technology Department of Henan Province (no. 162102210261).

References

[1] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," *Neurocomputing*, vol. 256, pp. 56–62, 2017.

[2] B. Schölkopf, K. Tsuda, and J. Vert, "Gene expression analysis: joint feature selection and classifier design," *MIT Press*, pp. 299–317, 2004.

[3] H. Wang, X. Jing, and B. Niu, "A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data," *Knowledge-Based Systems*, vol. 126, pp. 8–19, 2017.

[4] J. Thomas, T. Hepp, A. Mayr, and B. Bischl, "Probing for sparse and fast variable selection with model-based boosting," *Computational and Mathematical Methods in Medicine*, Art. ID 1421409, 8 pages, 2017.

[5] L. Zhao and Z. Zhang, "Supervised locally linear embedding with probability-based distance for classification," *Computers & Mathematics with Applications*, vol. 57, no. 6, pp. 919–926, 2009.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[7] J. L. Ward and S. L. Lumsden, "Locally linear embedding: dimension reduction of massive protostellar spectra," *Monthly Notices of the Royal Astronomical Society*, vol. 461, no. 2, pp. 2250–2256, 2016.

[8] Y. Chen, B. An, J. Dong, and M. Zhao, "A novel manifold preserving edit propagation based on K nearest neighbors and locally linear embedding," *Optical and Quantum Electronics*, vol. 48, no. 11, article no. 488, 2016.

[9] Y. Liu, Z. Yu, M. Zeng, and Y. Zhang, "LLE for submersible plunger pump fault diagnosis via joint wavelet and SVD approach," *Neurocomputing*, vol. 185, pp. 202–211, 2016.

[10] S. Kang, D. Ma, Y. Wang, C. Lan, Q. Chen, and V. I. Mikulovich, "Method of assessing the state of a rolling bearing based on the relative compensation distance of multiple-domain features and locally linear embedding," *Mechanical Systems and Signal Processing*, vol. 86, pp. 40–57, 2017.

[11] W. Ou, S. Yu, G. Li, J. Lu, K. Zhang, and G. Xie, "Multi-view non-negative matrix factorization by patch alignment framework with view consistency," *Neurocomputing*, vol. 204, pp. 116–124, 2016.

[12] C.-T. Huang, Z. Wang, and C.-C. J. Kuo, "Visible-light and near-infrared face recognition at a distance," *Journal of Visual Communication and Image Representation*, vol. 41, pp. 140–153, 2016.

[13] R. Rabia, C. K. Verma, and N. Namita, "A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data," *Genomics Data*, vol. 8, pp. 4–15, 2016.

[14] M. Xi, J. Sun, L. Liu, F. Fan, and X. Wu, "Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine," *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 3572705, 9 pages, 2016.

[15] G. Lu, D. Chen, Y. Du et al., "Using co-classification approach to detect the type of cancer," *Computer Science*, vol. 37, no. 2, pp. 232–236, 2010.

[16] J. Zhang, G. Hu, and P. Li, "Informative gene selection for tumor classification based on iterative lasso," *Pattern Recognition and Artificial Intelligence*, vol. 27, no. 1, pp. 49–59, 2014.

[17] Z. Mao, W. Cai, and X. Shao, "Selecting significant genes by randomization test for cancer classification using gene expression data," *Journal of Biomedical Informatics*, vol. 46, no. 4, pp. 594–601, 2013.

[18] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "A novel aggregate gene selection method for microarray data classification," *Pattern Recognition Letters*, vol. 60-61, pp. 16–23, 2015.

[19] Q. Ye, Y. Gao, R. Wu et al., "Informative gene selection method based on symmetric uncertainty and svm recursive feature elimination," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 30, no. 5, pp. 429–438, 2017.

[20] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Distributed feature selection: an application to microarray data classification," *Applied Soft Computing*, vol. 30, pp. 136–150, 2015.

[21] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Applied Soft Computing*, vol. 50, pp. 124–134, 2017.

[22] 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, no. 5439, pp. 531-527, 1999.

[23] W. Chu, Z. Ghahramani, F. Falciani, and D. L. Wild, "Biomarker discovery in microarray gene expression data with Gaussian processes," *Bioinformatics*, vol. 21, no. 16, pp. 3385–3393, 2005.

[24] S. K. Shevade and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression," *Bioinformatics*, vol. 19, no. 17, pp. 2246–2253, 2003.

[25] I. Guyon, "Erratum: gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2001.

[26] X. Wang and R. Simon, "Microarray-based cancer prediction using single genes," *BMC Bioinformatics*, vol. 12, article 391, 2011.

[27] G. J. Gordon, R. V. Jensen, L. Hsiao et al., "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, no. 17, pp. 4963–4967, 2002.

[28] C. Wei, Z. Ghahramani, F. Falciani, and D. L. Wild, "Biomarker discovery in microarray gene expression data with gaussian processes," *Bioinformatics*, vol. 21, no. 16, pp. 3385–3393, 2005.

[29] D. Singh, P. G. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[30] D. Karan, D. L. Kelly, A. Rizzino, M.-F. Lin, and S. K. Batra, "Expression profile of differentially-regulated genes during progression of androgen-independent growth in human prostate cancer cells," *Carcinogenesis*, vol. 23, no. 6, pp. 967–975, 2002.

[31] Z. Lai, W. K. Wong, Y. Xu, J. Yang, and D. Zhang, "Approximate orthogonal sparse embedding for dimensionality reduction,"

IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 4, pp. 723–735, 2016.

[32] Y. Cheng, B. Jiang, N. Lu, T. Wang, and Y. Xing, "Incremental locally linear embedding-based fault detection for satellite attitude control systems," *Journal of The Franklin Institute*, vol. 353, no. 1, pp. 17–36, 2016.

[33] Y. Liu, Y. Zhang, Z. Yu, and M. Zeng, "Incremental supervised locally linear embedding for machinery fault diagnosis," *Engineering Applications of Artificial Intelligence*, vol. 50, pp. 60–70, 2016.

[34] D. de Ridder, O. Kouropteva, O. Okun, M. Pietikäinen, and R. P. Duin, "Supervised locally linear embedding," *Joint International Conference on Artificial Neural Networks and Neural Information Processing*, vol. 2714, pp. 333–341, 2003.

[35] Y. S. Son and J. Baek, "A modified correlation coefficient based similarity measure for clustering time-course gene expression data," *Pattern Recognition Letters*, vol. 29, no. 3, pp. 232–242, 2008.

[36] T. Anna, N. Robert, and I. Nanako, "Correlation analysis based on Spearman's rank correlation coefficient between gene expression data and the histological grade (or metastasis status)," *Biological Sciences*, 2014.

[37] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[38] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.