Genetics and Molecular Research

Locally linear embedding and neighborhood rough set-based gene selection for gene expression data classification

L. Sun[1,2,3], J.-C. Xu[2,3], W. Wang[2] and Y. Yin[2]

[1]Post-doctoral Mobile Station of Biology, College of Life Science, Henan Normal University, Xinxiang, China
[2]College of Computer and Information Engineering, Henan Normal University, Xinxiang, China
[3]Engineering Technology Research Center for Computing Intelligence and Data Mining, Henan Province, China

Corresponding authors: L. Sun / J.C. Xu
E-mail: linsunok@gmail.com / xjc@htu.cn

Genet. Mol. Res. 15 (3): gmr.15038990
Received July 21, 2016
Accepted August 1, 2016
Published August 30, 2016
DOI http://dx.doi.org/10.4238/gmr.15038990

ABSTRACT. Cancer subtype recognition and feature selection are important problems in the diagnosis and treatment of tumors. Here, we propose a novel gene selection approach applied to gene expression data classification. First, two classical feature reduction methods including locally linear embedding (LLE) and rough set (RS) are summarized. The advantages and disadvantages of these algorithms were analyzed and an optimized model for tumor gene selection was developed based on LLE and neighborhood RS (NRS). Bhattacharyya distance was introduced to delete irrelevant genes, pair-wise redundant analysis was performed to remove strongly correlated genes, and the wavelet soft threshold was determined to eliminate noise in the gene datasets. Next,

prior optimized search processing was carried out. A new approach combining dimension reduction of LLE and feature reduction of NRS (LLE-NRS) was developed for selecting gene subsets, and then an open source software Weka was applied to distinguish different tumor types and verify the cross-validation classification accuracy of our proposed method. The experimental results demonstrated that the classification performance of the proposed LLE-NRS for selecting gene subset outperforms those of other related models in terms of accuracy, and our proposed approach is feasible and effective in the field of high-dimensional tumor classification.

Key words: Gene selection; Locally linear embedding; Classification; Neighborhood rough set

INTRODUCTION

Microarray techniques for tumor prognosis and various classification methods have been applied to analyze or interpret gene expression data (Liu et al., 2015). However, because of the limited availability of effective samples compared to the large number of genes in microarray data, many computational methods have failed to identify a small subset of important genes (Kar et al., 2015). In recent years, gene expression profiles for the molecular diagnosis of tumor have attracted attention for their potential in precise and early tumor diagnosis; however, dimensionality curse caused by high dimensionality and small sample size of tumor datasets challenges tumor classification (Wang et al., 2010; Sun et al., 2015). In general, the classification of cancer by microarray data involves data acquisition and pre-processing, gene selection, and classification (Elyasigomari et al., 2015). The aim of gene selection is to reduce the dimensionality of microarray data to enhance the accuracy of classification (Tabakhi et al., 2015).

The methods applied for gene selection are broadly divided into four categories including filter, wrapper, embedded, and hybrid approaches (Li et al., 2015). Filter methods are easily trapped into a local optimum. Wrapper approaches suffer from high computational cost, particularly given the high-dimensionality of microarray datasets. The main advantage of embedded approaches is the interaction with the learning model, but training a given classifier with a full gene set is time-consuming. The major disadvantage of hybrid approaches is that the filter and wrapper approaches are not truly integrated with each other, which may lead to lower classification performance (Tabakhi et al., 2015). Liu et al. (2010) introduced a conditional mutual information-based ensemble gene selection method for cancer microarray. Li et al. (2011) developed an embedded feature selection algorithm. However, this method requires the adjustment of a large number of parameters and its performance largely depends on those parameters. Cai et al. (2009) constructed a gene selection algorithm based on mutual information, but its computational cost increases as the number of selected genes is increased. Gan et al. (2008) proposed a gene selection method involving a Bayesian discriminant cost function. However, the proposed strategy is only applicable to one representative gene selection model-Bayesian discriminant-based genetic algorithms search. Sun and Xu (2014) improved the computational efficiency of a heuristic algorithm for gene selection.

Currently, the numerous available methods of gene selection require improvements for the pre-processing of data. The limitations are as follows: 1) Since some gene expression levels are very similar in all samples of tumor gene expression profiles, the corresponding genes are unrelated to classification. In order to select useful information regarding discrimination and decrease the computational complexity of searching gene subsets, these unrelated genes should be deleted. 2) Because strong correlation exists among genes, many genes are associated with subtype recognition for particular tumors. However, few genes are directly related to the tumor. Thus, strongly correlated genes should be removed. 3) Since noise is introduced at each processing stage, it is inevitably present in gene expression profiles. When the intensity of the noise is high, data points may be completely obscured. Extracting genes with noise from gene expression profiles can produce deviation. Thus, noise should be eliminated. Based on these limitations, the pre-processing of tumor genes should be used for forward optimization. It is known that familiar pre-processing methods of dimension reduction include principal component analysis (PCA), linear discriminant analysis, and locally linear embedding (LLE), among others. Sahu et al. (2014) proposed a feature selection procedure that augmented kernel PCA to obtain importance estimates of the features using noisy training data. However, PCA can only identify linear relationships among features in the data. Sharma et al. (2014) proposed a feature selection method using an improved regularized linear discriminant analysis technique. Lang et al. (2011) developed a LLE-based gene selection method for cancer classification. LLE and its extensions are a promising technique that can be used to solve the dimension reduction problem of high-dimensional data (Roweis and Saul, 2000). To evaluate gene selection methods, in addition to the predictive ability of gene subsets, two other important aspects that must be considered include stability of the selected genes and computational costs (Nguyen et al., 2015). Gene subsets with low dimensionality and high classification ability can be selected from gene expression profiles. Although it is clear that classical rough set (RS) algorithms are acceptable tools for selecting tumor genes, RS only can handle the character data subtype. Numeric and continuous data can only be handled after discretization. To avoid information loss and improve the classification accuracy of gene subsets, neighborhood RS (NRS) models are introduced into feature reduction and do not require discretization processing for continuous features. For example, Hu et al. (2008) proposed a heterogeneous gene selection method based on NRS. Liu et al. (2014) calculated the positive region of NRS and presented a quick NRS reduct algorithm.

In this study, classical reduction methods and gene expression profile analysis were combined to overcome the limitations of dimension reduction. We first removed noise and irrelevant and redundant genes from the original gene space by effective gene selection methods. Bhattacharyya distance (Sun et al., 1996) was introduced to delete irrelevant genes, pair-wise redundant analysis (Li and Ruan, 2005) was performed to remove strongly correlated redundant genes, and a wavelet method (Liu et al., 2007) was applied to eliminate the noise of genes by a given soft threshold. Next, prior optimized pre-processing was carried out, generating a reduction gene subset with more classification information. In the process of gene selection, relevance and redundancy analysis was performed to identify irrelevant and redundant genes. Next, LLE and NRS were combined to build an effective tumor gene selection and classification model, which improved the classification ability of selected gene subsets. For acute leukemia and colon cancer datasets, useful genes were selected by transforming neighborhood and several parameters. We validated the proposed LLE-NRS approach, which

improved the performance of gene selection and showed precise classification ability in the analysis of public microarray datasets.

MATERIAL AND METHODS

LLE-based dimension reduction

LLE is a dimension reduction technique for nonlinear data (Lang et al., 2011). The basic idea involves converting global nonlinear data into local linear, and obtaining global structure information by overlapping local areas. After linear dimension reduction for each local area, the low-dimensional global coordinates were obtained by combining the results according to certain rules. The specific steps of LLE-based dimension reduction algorithm are as follows:

Step 1: Input N original data x_{ij} with D-dimension, and identify k neighbor points for every sample point. When k neighbor points of the ith point are obtained, the Euclidean distance between the ith point and any other point is calculated. All of the calculated distances are sorted, and the first k^2 points close to the ith point are selected. Two conditions of the distance space are as follows: zero conditions $d_{ij} = 0$ if $i = j$, and triangle side ranging principle $d_{ij} + d_{jk} \geq d_{ik}$, where $d_{ij} = \sqrt[p]{\sum_{k=1}^{\gamma} |x_{ik} - x_{jk}|^p}$, $i, j \in [1, N]$, $k \in [1, D]$, d_{ij} is the Euclidean distance if $p = 2$, d_{ij} is the City-Block distance if $p = 1$, and d_{ij} is the dominance distance if $p = t$. Each point x_i in the samples has corresponding similarity with the remaining $N - 1$ points, which can be directly measured by determining the distance between two points. k is artificially set according to experience parameters, which should be greater than the output dimension of the samples. Here, if k is too large, the output results cause the different categories of data to become superimposed. If k is too small, the topological structure of the sample points cannot be maintained in low-dimensional space. From the processing results of Swiss-Roll data, when k is 3 or 4, the 3-dimensional data cannot be mapped into 2-dimensional space using LLE. When k gradually increases, the distribution of 2-dimensional data improves. When k is between 20 and 30, the distribution is optimal. When k continues to increase, the distribution effect gradually deteriorates. When k is up to 50, the data appears to be superimposed.

Step 2: Design a local reconstructed weight matrix of the sample points, after which the optimal linear reconstructed weight can be calculated from:

$$\varepsilon_i(w) = \min \| x_i - \sum_{j=1, j\neq i}^{k} w_{ij} x_j \|^2 = \min \| \sum_{j, j\neq i} w_{ij}(x_i - x_j) \|^2 = \sum_{j,k} w_j w_k G_{jk} \quad \text{(Equation 1)}$$

where e_i is an error function of linear reconstruction between x_i and k neighbor points $x_1, x_2, \ldots, x_k$. A smaller $e_i(w)$ value results in a better local reconstructed weight matrix. This means that x_i is closer to the linear combination point of its neighboring points. $G_{jk} = (x_i - x_j)^T(x_i - x_k)$ is a local Gramian matrix. w_{ij} is a linear reconstructed weight and satisfies two following constraint conditions: $w_{ij} = 1$ if x_j is a neighbor point of x_i, otherwise $w_{ij} = 0$, where the sum of all w_{ij} is 1. These values can be also referred to as sparse constraint conditions. The sample points must be refactored from their neighbor points. Additionally, an optimal weight $w_j = \frac{\sum_k G_{jk}^{-1}}{\sum_{bm} G_{bm}^{-1}}$ is calculated by the Lagrange multiplier approach, where an inverse matrix G_{jk}^{-1} exists when G_{jk}

is positive definite. Next, the weight of the minimum reconstructed error functions can satisfy the above constraint conditions.

Step 3: Map all of the sample points into a low-dimension space and calculate a low-dimension embedding matrix Y of the input samples using the weight matrix determined in Step 2 as follows:

$$\varepsilon(Y) = \min \| Y_i - \sum_{j=1}^{k} w_{ij} Y_{ij} \|^2 = \min \sum_{j,k} M_{ij} (Y_i \bullet Y_{ij}) \quad \text{(Equation 2)}$$

where $i, j = 1, 2, \ldots, k$, $(Y_i \bullet Y_{ij})$ represents the relationship of inner product between Y_i and Y_{ij}, M_{ij} is a sparse symmetric positive semi-definite matrix, and $Y_{i1}, Y_{i2}, \ldots, Y_{ik}$ are k neighbor points of Y_i which satisfies the following conditions: $\sum_{i=1}^{N} Y_i = 0$ and $\frac{\sum_{i=1}^{N} Y_i Y_i^T}{N} = I$. Here, I is a $d \times d$ unit matrix. When Y_i is moved to any location, the reconstructed error function ε is not affected, the translational freedom degree is eliminated, and the core of output results after mapping should be on the origin of the coordinates. When ε is not affected by Y_i, the rotation and proportion freedom degrees should be eliminated.

Subsequently, the Lagrange multiplier approach is employed to calculate eigenvectors, which correspond to the relatively smaller $d + 1$ eigenvectors of cost matrix M_{ij}. Here, eigenvectors with the smallest eigenvalue are all 1 vector. This represents a free translation mode corresponding to the eigenvalue 0. These vectors should be deleted, and then the output results of LLE are composed of the reserved d eigenvectors. To more intuitively demonstrate the characteristics of the LLE method, the acute leukemia dataset was downloaded from http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi, and the number of selected genes in the training samples was 38. There are 8 types of acute lymphoblastic leukemia (T-cell), 19 acute lymphoblastic leukemia (B-cell), and 11 acute myeloid leukemia (Sun et al., 2015). Each sample point has 7129 features. When k was 25, the output was a 38 x 2 eigenvector matrix; the mapping results are shown in Figure 1A. Next, 2-dimension mapping data was easily observed after dimension reduction using the LLE algorithm. Furthermore, because the data volume was too small and the results mapped on the same straight line, the visualization results of the sample points are shown more clearly in Figure 1B. Translation and zooming clearly improved the classification ability of the sample points through the processing of classifiers using the LLE algorithm. Through computing, the cross-validation recognition classification accuracy was up to 86.84%, which is a 17% increase compared to the original dataset.

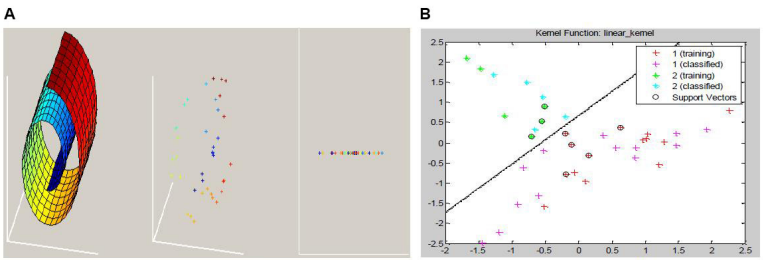

Figure 1. A. Mapping results of acute leukemia dataset, using the LLE algorithm. **B.** Visualization results of acute leukemia dataset, using the LLE algorithm.

To further illustrate the performance of the LLE method, the public colon tumor dataset was downloaded from http://www.molbio.princeton.edu/colondata. Sixty-two types of information genes were selected from the training samples, for which there are 40 types of tumor samples and 22 types of normal samples. Each sample point has 2000 features. For k = 25, the visualization results of the sample points are shown in Figure 2, where the sample points are divided into class1 and class2. All output results of class2 were –1, and the classification results of a small number of class1 were not ideal. However, the cross-validation accuracy was 80.65%, which is 11.3% higher than the 69.35% accuracy of the original colon tumor dataset.

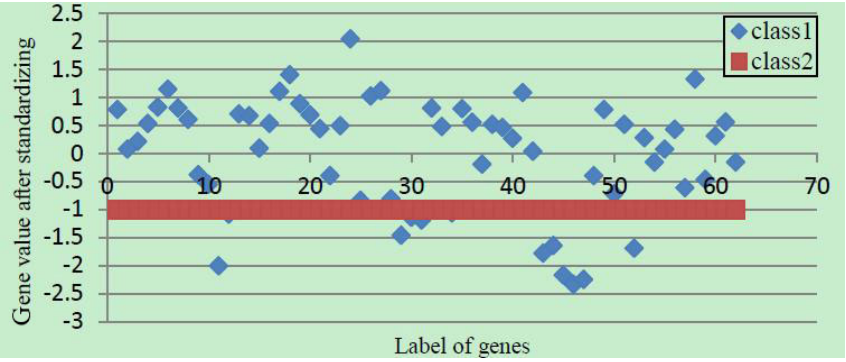

Figure 2. Visualization results of colon tumor dataset by LLE algorithm.

Improved RS-based feature selection

RS has been successfully applied for feature selection, as it can eliminate redundant features using individual feature information and mutual information (Meng et al., 2014). It is a useful tool for dealing with vague, uncertain, and incomplete information. Based on classical RS models, the selection criteria are constructed using feature dependence and significance measure for feature selection (Maji and Paul, 2011). However, some RS models can only deal with data with nominal features, and thus the datasets must be discretized before feature selection. In combination with PCA, the specific steps of the improved RS-based feature selection algorithm are as follows:

Step 1: Input an original gene dataset, calculate an average value $\overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$ and a standard deviation of feature values $s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1}$, and obtain a standardized formula $x_{ij}* = \frac{x_{ij} - \overline{x}_j}{s_j}$, where x is a feature value of original variable, i = 1, 2, ..., n, and j = 1, 2, ..., p.

Step 2: Apply the denoising method of soft thresholding-based wavelet transform to process noise in the gene dataset, where $\tilde{d}_{jk} = d_{jk} - \lambda$ if $d_{jk} \geq \lambda$, $\tilde{d}_{jk} = 0$ if $|d_{jk}| < \lambda$, $\tilde{d}_{jk} = d_{jk} + \lambda$ if $d_{jk} < -\lambda$, λ is a threshold, d_{jk} is a wavelet coefficient, and $\tilde{d}_{jk}$ is a processed wavelet coefficient.

Step 3: Employ a PCA method to reduce dimensionality of the data after denoising, and discretize the features of the obtained gene dataset.

Step 4: Use RS-based feature reduction for the discretized gene dataset to reduce feature columns.

Step 5: Eliminate the reduced feature columns and delete the duplicated rows if the

same tuples appear after the eliminating operation.

Step 6: Repeat Step 4 until no redundant features are present.

Step 7: Use Weka software to classify the features of the reduced dataset, output the corresponding classification results, and achieve cross-validation classification accuracy.

An acute leukemia dataset was used. Three features {feature2288, feature2795, feature4117} were selected using the improved RS-based feature selection method. The dimensionality of the dataset was greatly reduced. Through computation, the classification accuracy was 89.83% with Weka, which was approximately 20% higher than the value for the original dataset and 3% higher than that for the LLE algorithm. Thus, the improved RS-based feature selection method is effective. Figure 3A shows a 2-dimensional scatter plot of the above three selected features. The boundary between the two tumor subtypes for ALL and AML sample points was clearer. Similarly, two features {feature1668, feature1309} of the colon tumor dataset were selected and the dimension of data was greatly reduced. Figure 3B shows a 2-dimensional scatter plot of the two features, in which the feature values of the original data can be divided into positive and negative classes. However, the features of the two classes of sample points were mixed together, and the classification condition was not clear. The cross-validation classification accuracy was 75.81% under Weka, which was increased by only approximately 5% compared to the original dataset. Thus, the classical RS method was not effective for gene selection of the colon cancer dataset, and its classification accuracy of selected features was not ideal.

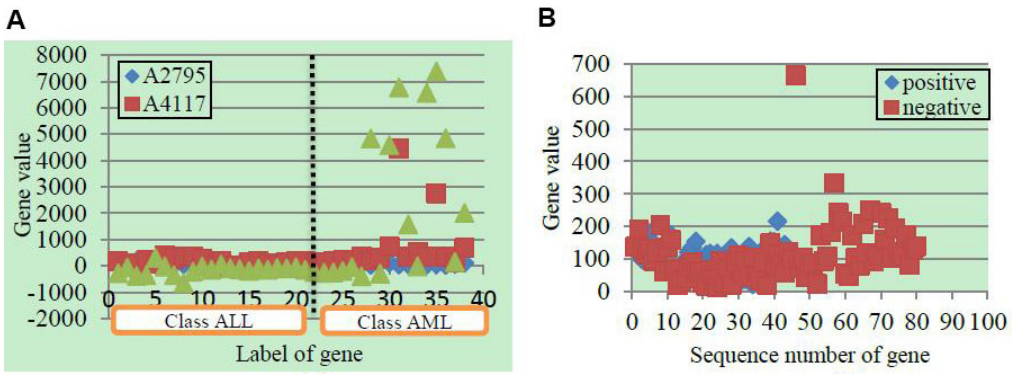

Figure 3. A. Two-Dimensional scatter plot for three features of the acute leukemia dataset; **B.** 2-dimensional scatter plot for two features of the colon tumor dataset.

LLE-NRS-based optimized gene selection

Discretization of classical RS model can lead to information dropout. To address this issue, many extended RS models have been employed, and NRS models were proposed for gene selection and classification (Wang et al., 2010; Hou et al., 2010). These methods employ δ neighborhood to deal with numerical data directly and use a forward feature reduction algorithm to select genes. Three key components of the gene selection methods using the RS model based on neighborhood are the construction of neighborhoods, approximation operators, and feature reduction. How to construct neighborhoods to suit for various data structures and design more effective feature reduction algorithm require further investigation (Meng et al., 2014).

A neighborhood of an object x is a set of objects with similar characteristics to x. A generalized definition for neighborhood has been given according to the binary relation (Yao and Lin, 1996). For an object $x \in U$, U is a nonempty finite set of objects, and a binary relation R on U, the neighborhood of x is $N_R(x) = \{y \mid xRy, y \in U\}$. Based on the generalized definition, many specific formulas for neighborhood are proposed to deal with complex real datasets. Because all datasets in gene selection are numerical features, the δ neighborhood is an effective method for dealing with numerical features. For an object $x \in U$ and a subset of features $B \subseteq C$, the δ neighborhood of x induced by B is defined as:

$$\delta_B(x) = \{y | \Delta B(x, y) \leq \delta, y \in U\} \quad \text{(Equation 3)}$$

where $\Delta B(x, y)$ is a distance metric function to determine the shape of the neighborhood and δ is a threshold to control the size of the neighborhood.

Note that the original dataset requires dimension reduction before feature reduction of NRS, and the requirements of the reduction model can be satisfied. Next, forward optimization during the tumor gene selection process is necessary. The specific steps of LLE-NRS-based optimized gene selection algorithm are as follows:

Step 1: Input original data x_{ij} and delete unrelated genes with Bhattacharyya distance to obtain gene subset G_b. The Bhattacharyya distance is adopted to process tumor gene samples and the unrelated genes are eliminated. Next, the concrete model can be expressed as

$$B(x) = \frac{(\mu_+(x) - \mu_-(x))^2}{4(\sigma_+^2(x) + \sigma_-^2(x))} + \frac{1}{2}\ln\left(\frac{\sigma_+^2(x) + \sigma_-^2(x)}{2\sigma_+(x)\sigma_-(x)}\right) \quad \text{(Equation 4)}$$

where μ_+ and μ_- are the averages of the expression levels of feature x in two different gene samples respectively, and σ_+ and σ_- are the corresponding standard deviations. Greater Bhattacharyya distances of genes result in larger differences in the expression level distribution of gene samples in the two classes; sample classification is also stronger.

Step 2: Perform pair-wise redundant analysis to remove strongly correlated redundancy genes of G_b and get gene subset G_r. If the correlation coefficient is greater than the specified threshold, the two genes are strongly related. The genes with a smaller classification information index are eliminated, and the classified gene subsets will have a larger classification information index.

Step 3: Perform wavelet analysis to eliminate noise of G_r under a given soft threshold and obtain gene subset G_x. Deviation is produced when the genes are extracted from gene datasets containing noise. In order to ensure the effectiveness of gene extraction and accuracy of classification recognition, noise should be eliminated, and the genes are extracted to obtain the corresponding feature reduction subset.

For practical application, the essence of the wavelet denoising problem is a function approximation. The denoising wavelet threshold is mainly based on the threshold function of the wavelet high-frequency subspace. Coefficients that are less than the threshold are set to zero by appropriate thresholds. Wavelet coefficients that are greater than the threshold will be retained. Next, the estimated coefficients can be obtained by mapping of the threshold function. This can be achieved by inverse wavelet transform processing, and the signal after denoising will be rebuilt. The selected threshold function model of the wavelet soft threshold denoising

algorithm is expressed as $\tilde{w}_{jk} = \text{sign}(w_{jk})(|w_{jk}| - \lambda)$ if $|w_{jk}| \geq \lambda$, otherwise $\tilde{w}_{jk} = 0$, where w_{jk} is a wavelet coefficient of the original signal containing noise by wavelet transform, and $\tilde{w}_{jk}$ is an estimated coefficient after threshold processing, namely soft threshold function. The signal can be estimated as the maximum mean variance and minimum error by using the soft threshold denoising method. This signal after denoising is an approximate optimal estimation of the original signal. These signals show the same smoothness as the original signals and do not produce additional shocks, indicating its wide adaptability and practicability.

Step 4: Reduce dimensionality of G_x using the LLE algorithm and get gene subset G_p. Apply the maximum and minimum method expressed as $f(x_i) = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$ to perform normalization processing of G_p, where $i = 1, 2, \ldots, n$, x_i is a sample feature and x_{max} and x_{min} are the maximum and minimum values of the samples, respectively. To obtain accurate processing results from NRS, all genes G_o are distributed on [0, 1].

Step 5: Reduce genes of G_o, using the NRS algorithm to obtain the optimal approximate gene subset G_n according to transformation of the neighborhood setting. Classify genes of G_n after reduction, output corresponding classification results under Weka, and achieve cross-validation classification accuracy.

Note that feedbacks exist in the LLE-NRS algorithm, and the aim is to select the best parameter combination. These parameters include the given wavelet soft threshold, k of LLE algorithm, neighborhood of the NRS algorithm, and related parameters under Weka.

RESULTS

To demonstrate the classification performance of tumor genes by our proposed LLE-NRS algorithm, three subtypes of public microarray datasets with different sample sizes and number of genes were downloaded from http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi, including ALL-T (acute lymphoblastic leukemia, T_cell), ALL-B (acute lymphoblastic leukemia, B_cell), and AML (acute myeloid leukemia). The datasets include 72 case samples. Each sample includes 7129 features. Next, the results for the acute leukemia (ALL) dataset after adjustment are shown in Table 1, where the numbers of training data and test data are 38 and 34, respectively.

Table 1. Results of acute leukemia dataset after adjustment.

Dataset	Training data			Test data		
	ALL-T	ALL-B	AML	ALL-T	ALL-B	AML
Original sample classification	8	19	11	2	18	14
Adjusted sample classification	6	21	11	4	16	14
Sample proportion	27		11	20		14

The training samples are processed using Bhattacharyya distance. Three hundred genes were extracted as a foundation for further reduction. The redundant genes with strong correlations were extracted using a pair-wise redundant analysis algorithm. A total of 276 genes were obtained. When a wavelet transform denoising method was adopted, the db1 wavelet was used for decomposition of three levels and extraction of coefficients. Soft threshold values {1.465, 1.823, 2.768} were used to deal with noise. Next, a new signal was refactored and

extracted. The signal analysis diagram is shown in Figure 4, in which the signal obtained from wavelet denoising was better than that from the forced denoising. This did not omit the small percentage of detailed data points, and the topology of the original signal was retained. Compared with the signal from acquiescent threshold denoising, the signal from wavelet denoising was clear and integrated, and the denoising process of genes was effective.

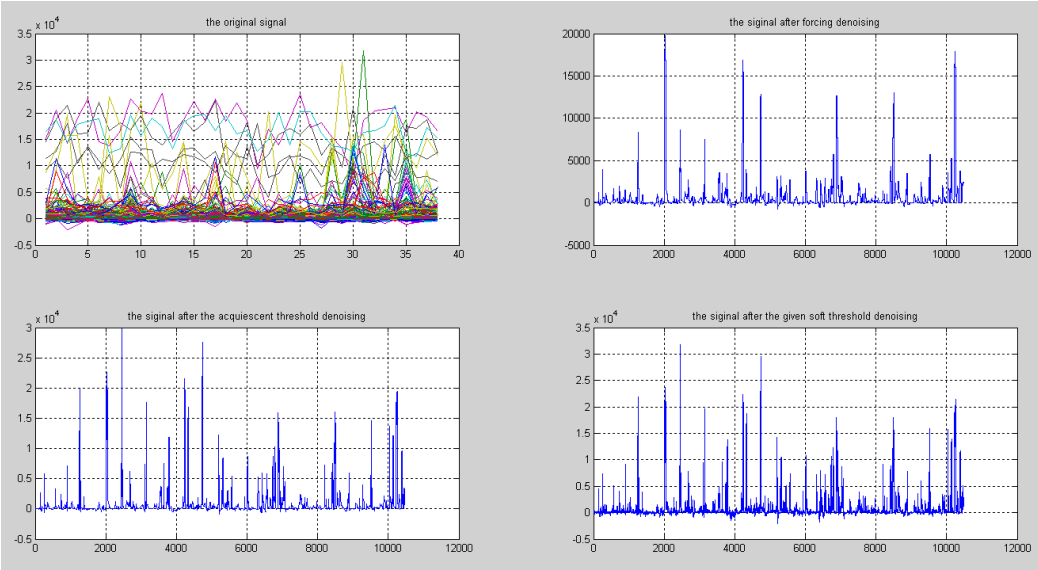

Figure 4. Signal analysis diagrams of acute leukemia dataset.

The gene samples were subjected to reduction dimension using the LLE algorithm. Gene subsets with most classification information were obtained. Next, feature reduction was carried out using the NRS algorithm. Weka was used to obtain the classification accuracy of each gene subset. In the following experiments, to obtain better experimental results and reflect the comparability and repeatability of the data, the proposed LLE-NRS algorithm was compared using several different gene selection and classification algorithms with an acute leukemia dataset. Here, PCA (Whipple et al., 2004) selected principal components whose contribution rates were greater than 85%. Supervised LLE (SLLE) was applied to the classification problem of tumor genes (Pillati and Viroli, 2005). When the sample supervised information was added to the LLE algorithm to guide the classification of data, locally linear discriminant embedding (LLDE) (Huang, 2009) was used to identify linear transformation, which met the requirements of minimum reconstruction error in LLE and best translation and scaling transformation. In order to verify the classification accuracy of our proposed algorithm, the LLE-NRS algorithm was compared with PCA, PCA+NRS, SLLE+KNN, and LLDE+KNN algorithms under 10-fold cross-validation, where *k*-nearest neighbor (KNN) was used for gene expression classification. The experimental results are shown in Table 2. The LLE-NRS algorithm outperformed some of the existing standard techniques and showed the highest classification accuracy.

Table 2. Classification results of five methods for acute leukemia dataset.

Dimension reduction methods	Genes	Cross-validation classification accuracy (%)
PCA	21	86.84
PCA + NRS	6	84.21
SLLE + KNN	6	87.21
LLDE + KNN	3	89.44
LLE-NRS	2	94.74

To further evaluate the classification performance of the LLE-NRS algorithm, using acute leukemia and colon cancer datasets, LLE-NRS was compared with the other seven gene selection methods including LLE (Lang et al., 2011), RS (Sun and Xu, 2014), SNRS (Xu et al., 2015), FRADM (Zhang, 2008), BAHSIC (Song et al., 2012), APBEFS (Meng and Wei, 2015), and LNB-MS (Wu et al., 2012). Table 3 shows the classification results of tumor datasets with several gene selection algorithms. The classification accuracy of the LLE-NRS algorithm was higher than those of the other algorithms, except for APBEFS on the acute leukemia dataset. Furthermore, the classification accuracy of the acute leukemia dataset using our optimized gene selection model was 23.69% higher than that of the original data. Similarly, the accuracy of the colon tumor dataset was 30.65% higher than that of original dataset, and in some cases reached 100%. Thus, our proposed LLE-NRS-based gene selection method is efficient for classification.

Table 3. Classification results of eight gene selection algorithms for two tumor datasets.

Tumor dataset	Cross-validation classification accuracy (%)								
	Original data	LLE	RS	SNRS	FRADM	BAHSIC	APBEFS	LNB-MS	LLE-NRS
Acute leukemia	71.05	86.84	89.93	68.06	87.54	95.7	98.32	93.42	94.74
Colon tumor	69.35	80.65	75.81	64.52	85.48	81	85.22	89.7	100

DISCUSSION

Tumor classification and gene selection from high-dimensional data have been widely examined in genetics and molecular biology studies (Algamal and Lee, 2015). In this study, irrelevant and redundant genes were identified and removed by Bhattacharyya distance and pair-wise redundant analysis. The wavelet soft threshold method was applied to eliminate the noise from genes. Prior optimized pre-processing was carried out. LLE and NRS were combined to develop gene selection and classification models, which improved the classification ability of selected gene subsets. Weka was applied to distinguish different tumor types and verify cross-validation classification accuracy of our proposed method. Our experimental results showed that LLE-NRS outperformed the other related feature selection algorithms with a positive tradeoff between classification precision and performance on public microarray datasets, ant it can be applied to further improve the performance of dimension reduction, feature selection, and classification of other datasets.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#61402153 and #61370169), the Key Project of Science and Technology Department of Henan Province

(#142102210056 and #162102210261), the Key Project of Educational Department of Henan Province (#16A520016), and the Ph.D. Research Startup Foundation of Henan Normal University (#qd15132, #qd15130, and #qd15129).

REFERENCES

Algamal ZY and Lee MH (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Comput. Biol. Med.* 67: 136-145. http://dx.doi.org/10.1016/j.compbiomed.2015.10.008

Cai R, Hao Z, Yang XW and Wen W (2009). An efficient gene selection algorithm based on mutual information. *Neurocomputing* 72: 991-999. http://dx.doi.org/10.1016/j.neucom.2008.04.005

Elyasigomari V, Mirjafari MS, Screen HRC and Shaheed MH (2015). Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization. *Appl. Soft Comput.* 35: 43-51. http://dx.doi.org/10.1016/j.asoc.2015.06.015

Gan ZH, Tommy WS and Huang D (2007). Effective gene selection method using Bayesian discriminant based criterion and genetic algorithms. *J. Signal Proc.* 50: 293-304.

Hou ML, Wang SL, Li XL and Lei YK (2010). Neighborhood rough set reduction-based gene selection and prioritization for gene expression profile analysis and molecular cancer classification. *J. Biomed. Biotechnol.* 2010: 726413. http://dx.doi.org/10.1155/2010/726413

Hu QH, Yu D, Liu JF and Wu CX (2008). Neighborhood rough set based heterogeneous feature subset selection. *Inform. Sci.* 178: 3577-3594. http://dx.doi.org/10.1016/j.ins.2008.05.024

Huang DS (2009). Research on Data Mining Methods to Gene Expression Profile. Science Press, 102-106.

Kar S, Sharma KD and Maitra M (2015). Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. *Expert Syst. Appl.* 42: 612-627. http://dx.doi.org/10.1016/j.eswa.2014.08.014

Lang YX, Qin Z and Li X (2011). An effective gene selection method for cancer classification based on locally linear embedding. *J. Comput. Theor. Nanosci.* 8: 2108-2111. http://dx.doi.org/10.1166/jctn.2011.1932

Li J, Jia Y and Li W (2011). Adaptive huberized support vector machine and its application to microarray classification. *Neural Comput. Appl.* 20: 123-132. http://dx.doi.org/10.1007/s00521-010-0371-y

Li J, Su L and Pang Z (2015). A filter feature selection method based on MFA score and redundancy excluding and it's application to tumor gene expression data analysis. *Interdiscip. Sci.* 7: 391-396. http://dx.doi.org/10.1007/s12539-015-0272-y

Li YX and Ruan XG (2005). Feature selection for cancer classification based on support vector machine. *J. Comp. Res. Dev.* 42: 1796-1801. http://dx.doi.org/10.1360/crad20051024

Liu H, Liu L and Zhang H (2010). Ensemble gene selection by grouping for microarray data classification. *J. Biomed. Inform.* 43: 81-87. http://dx.doi.org/10.1016/j.jbi.2009.08.010

Liu JX, Xu Y, Zheng CH, Kong H, et al. (2015). RPCA-based tumor classification using gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 12: 964-970. http://dx.doi.org/10.1109/TCBB.2014.2383375

Liu WD, Liu SH, Hu XF and Wang L (2007). Analysis of modified methods of wavelet threshold de-noising functions. *High Voltage Eng.* 33: 59-63.

Liu Y, Huang WL, Jiang YL and Zeng ZY (2014). Quick attribute reduct algorithm for neighborhood rough set model. *Inf. Sci.* 271: 65-81. http://dx.doi.org/10.1016/j.ins.2013.08.022

Maji P and Paul S (2011). Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. *Int. J. Approx. Reason.* 52: 408-426. http://dx.doi.org/10.1016/j.ijar.2010.09.006

Meng J and Wei SY (2015). Affinity propagation clustering based ensemble feature selection method. *Comput. Sci.* 42: 241-244.

Meng J, Zhang J, Lia R and Luan YS (2014). Gene selection using rough set based on neighborhood for the analysis of plant stress response. *Appl. Soft Comput.* 25: 51-63. http://dx.doi.org/10.1016/j.asoc.2014.09.013

Nguyen T, Khosravi A, Creighton D and Nahavandi S (2015). A novel aggregate gene selection method for microarray data classification. *Pattern Recognit. Lett.* 60-61: 16-23. http://dx.doi.org/10.1016/j.patrec.2015.03.018

Pillati M and Viroli C (2005). Supervised locally linear embedding for classification: An application to gene expression data analysis. In: Book of Short Papers (Zani S and Cerioli A, eds.). CLADAG2005, Parma, 6-8 Giugno, MUP, 147-150.

Roweis ST and Saul LK (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290: 2323-2326.

http://dx.doi.org/10.1126/science.290.5500.2323
Sahu A, Apley DW and Runger GC (2014). Feature selection for noisy variation patterns using kernel principal component analysis. *Knowl. Base Syst.* 72: 37-47. http://dx.doi.org/10.1016/j.knosys.2014.08.027
Sharma A, Paliwal KK, Imoto S and Miyano S (2014). A feature selection method using improved regularized linear discriminant analysis. *Mach. Vis. Appl.* 25: 775-786. http://dx.doi.org/10.1007/s00138-013-0577-y
Song L, Smola A, Gretton A, Bedo J, et al. (2012). Feature selection via dependence maximization. *J. Mach. Learn. Res.* 13: 1393-1434.
Sun L and Xu J (2014). A granular computing approach to gene selection. *Biomed. Mater. Eng.* 24: 1307-1314.
Sun L, Xu J and Yin Y (2015). Principal component-based feature selection for tumor classification. *Biomed. Mater. Eng.* 26 (Suppl 1): S2011-S2017. http://dx.doi.org/10.3233/BME-151505
Sun L, Han CZ, Dai N and Shen JJ (2006). Feature selection based on Bhattacharyya Distance: a generalized rough set method. In: 2006 6th World Congress on Intelligent Control and Automation. IEEE, 2006 2: 10101-10105.
Tabakhi S, Najafi A, Ranjbar R and Moradi P (2015). Gene selection for microarray data classification using a novel ant colony optimization. *Neurocomputing* 168: 1024-1036. http://dx.doi.org/10.1016/j.neucom.2015.05.022
Wang SL, Li X, Zhang S, Gui J, et al. (2010). Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. *Comput. Biol. Med.* 40: 179-189. http://dx.doi.org/10.1016/j.compbiomed.2009.11.014
Whipple ME, Mendez E, Farwell DG, Agoff SN, et al. (2004). A log likelihood predictor for genomic classification of oral cancer using principle component analysis for feature selection. *Stud. Health Technol. Inform.* 107: 823-826.
Wu MY, Dai DQ, Shi Y, Yan H, et al. (2012). Biomarker identification and cancer classification based on microarray data using Laplace naive Bayes model with mean shrinkage. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 9: 1649-1662. http://dx.doi.org/10.1109/TCBB.2012.105
Xu JC, Li T, Sun L and Li YH (2015). Feature gene selection based on SNR and Neighborhood rough set. *Data Acquis. Proc.* 30: 973-981.
Yao YY and Lin TY (1996). Generalization of rough sets using modal logic. *Intell. Autom. Soft Comput.* 2: 103-119. http://dx.doi.org/10.1080/10798587.1996.10750660
Zhang LJ (2008). Research on Feature Selection for Classification in Microarray Gene Expression Data. Ph.D. thesis, National University of Defense Technology, 45-69.