Relevant Component Locally Linear Embedding Dimensionality Reduction for Gene Expression Data Analysis

Xiaoping Min, Hai Wang, Zhiwei Yang

Department of Computer Science,
Xiamen University, Xiamen 361005, China

Shengxiang Ge, Jun Zhang, Ningxia Shao

National Institute of Diagnostics and Vaccine Development in Infectious Diseases,
Xiamen University, Xiamen 361005, China

Abstract
Gene expression data typically contain many genes, which generates a feature vector with high dimensionality and considerable irrelevant information. Moreover, datasets typically contain few samples, which leads to the 'curse of dimensionality'. This dimensionality degrades classification performance. Therefore, it is essential to determine a good method for reducing the feature set size. In this paper, we proposed a relevant component locally linear embedding (RLLE) algorithm. This method changes the feature space used for data representation through a global linear transformation that assigns large weights to relevant dimensions and low weights to irrelevant dimensions. Next, the new distance between the sample points is calculated. Through processing the feature space, sample points of the same class could be efficiently neighbored. Based on the new distance, a locally linear embedding algorithm was applied to reduce the dimensionality of the samples. We applied the techniques to six published DNA microarray data sets and compare the RLLE algorithm with PCA, ISOMAP, LLE and RELIEFF using Naive Bayes classifiers. The experimental results and comparisons demonstrate that the proposed method is highly effective.

Key words: GENOMIC MICROARRAY, DIMENSIONALITY REDUCTION, RELEVANT COMPONENT, LOCALLY LINEAR EMBEDDING

Introduction

DNA microarray technology measures the level of expression for tens of thousands of genes on a fingernail-size chip. Gene and protein expression profile analyses have become an effective prediction technique for diseases. Based on the analyses, the genes most correlated with diseases are selected, and the diseases are categorized based on patients' expression profiles. Based on these data, we can establish a relationship between genes and disease categories, which could be used to determine gene categories and predict gene function [1, 2, 3].

Researchers have proposed a range of approaches for gene selection using microarray data analyses [4, 5, 6]. Simultaneously, substantial research has considered classifications based on patients' expression profiles [7, 8]. However, DNA microarray data typically contains thousands of genes, but the number of samples is relatively small (tens or hundreds). These data sets are regarded as high-dimensional data for small samples, which encounter the "curse of dimensionality". Moreover, DNA microarray data sets contain many genes that are irrelevant to disease classification, and at the same time many "technical" and "biological" noise data. Technical noise data originates from different stages, such as producing a gene chip and preparing gene samples. However, the biological noise data stem from the diverse genetic backgrounds of samples or impurities mixed into the samples. Therefore, it is impossible to extract useful information from the data without processing it[9,10,11,12].

To solve the curse of dimensionality in analyzing gene expression profiles and minimize the effects from noise data, two solutions have been proposed, namely feature selection and dimensionality reduction. Feature selection is mainly used to select functional genes and has been extensively adopted in gene expression profile research. The most used independent criteria are distance measurements, information measurements, dependency measurements, consistency measurements, max-relevance, information gain, sum minority, twoing rule, F-statistics and so on[13,14,15,16,17]. Dimensionality reduction approaches include linear approaches, such as PCA and MDS, and nonlinear approaches, such as LLE, ISOMAP and LEM[18, 19, 20, 21,22].

George Lee has applied 6 dimensionality reduction approaches, including PCA, LDA, MDS, LLE, ISOMAP and LEM in their research to test and analyze gene expression profile data [23]. Their research indicated that gene expression profile data present a nonlinear structure. Therefore, nonlinear dimensionality reduction would yield better results than linear dimensionality reduction.

Xuehua Li proposed the kernel method based on locally linear embedding to select the optimal number of nearest neighbors and construct a uniform distribution manifold. The techniques were applied to two published DNA microarray datasets.

LLE, isomap aim to depict the high-dimensional data nonlinearly such that two adjacent points on the manifold are adjacent in the low dimensional embedding space. They have adopted Euclidean distance as the distance between two points, and the distance relationship was maintained during dimensionality reduction. However, due to the limited sample points in gene expression profiles, the data could not be densely and uniformly distributed in manifold. Moreover, due to high dimensionality and considerable noise, the sample neighbor points are not necessarily points in the same class. Therefore, reducing dimensionality does not always yield better classification results.

Shental proposed a relevant components analysis(RCA) algorithm in 2002[24]. They used small subsets of data from the same class, which are referred to as chunklets, to seek and reduce irrelevant variability in the data while amplifying relevant variability. The RCA transformation is intended to reduce clutter such that, in the new feature space, the inherent data structure can be more easily discerned. The method can be used to preprocess supervised data clustering or nearest neighbor classification[25].

In this research, a relevant component LLE algorithm is proposed. Through processing the feature space, sample points of the same class can be efficiently neighbored. First, the ReliefF algorithm was used to calculate the relevance between sample attributes and correlate classes with known class labels for data. Next, large weights were assigned to "relevant dimensions", and low weights were assigned to "irrelevant dimensions". After this global linear transformation of feature space, the distance between the sample points was re-calculated, and K nearest neighbor samples of each point were obtained. Based on the new distance, the LLE algorithm was applied to reduce the dimensionality of the samples. Finally, the dimensionality reduction results were classified using the NAÏVE BAYES classification algorithm. The classification results demonstrated that the algorithm described in this report significantly enhanced the classification accuracy.

Method

Description of the Data Sets

Six datasets in [26] were used in our research, which were summarized in Table 1.

•ALL: The ALL dataset contains 12625 genes for six acute lymphoblastic leukemia subtypes: 10 BCR, 27 E2A, 64 Hyperdip, 20 MLL, 43 T, and 79 TEL samples. More details on this data set can be found in [18].

•GCM: The GCM dataset [27] contains 16063 genes for fifteen types of human tumors: 12 breast, 14 prostate,12 lung, 12 colorectal, 22 lymphoma, 11 bladder, 10 melanoma, 10 uterus, 10 leukemia, 11 renal, 11 pancreas, 120 ovary, 11 mesothelioma, 20 CNS, and 9 MET samples.

•HBC: The HBC dataset consists of 22 hereditary breast cancer samples, each of which includes 3226 genes; it was first studied in [28].

•LYM: The Lymphoma dataset consists of 62 prevalent adult lymphoid malignancies samples; each sample has 4026 genes. This dataset was first studied in [29].

•MLL: The MLL-leukemia dataset consists of three classes and was first studied in[30].

•NCI60: The NCI60 dataset consists of 60 samples; it was first studied in [31]. cDNA microarrays were used to examine the variation in gene expression among the 60 cell lines from the National Center Institute's anticancer drug screen.

Table 1. The dataset description

dataset	samples	genes	classes
ALL	248	12558	6
GCM	198	16063	14
HBC	22	3226	3
LYM	62	4026	3
MLL	72	12582	3
NCI60	60	1123	9

Locally Linear Embedding (LLE)

Locally linear embedding (LLE) is an unsupervised learning algorithm that computes low dimensional embedding wherein nearby points in the high dimensional space remain nearby and similarly co-localized with respect to each other in the low dimensional space [22]. These coefficients do not change upon rotation, rescaling and translation; hence, they characterize the intrinsic geometric properties of each neighborhood.

Let X = {$x_1, x_2, . . ., x_n$} be n points in a high dimensional space.

LLE maps X to a data set Y ={$y_1, y_2, . . ., y_n$}, $y_i \in$ Rm(m <d). Assuming that the data lie on a nonlinear manifold, which can be locally approximated as linear, the outline of an LLE can be summarized as follows.

Step 1. Identify the k nearest neighbors of each point x_i ($i = 1, 2, . . ., n$) in X using Euclidean distances to measure similarity.

Step 2. Assign a weight to each pair of neighboring points. Compute the optimal reconstruction weights that can minimize error when linearly reconstructing Xi using its k nearest neighbors:

$$\varepsilon_i(w) = argmin\left|X_i - \sum_{j=1}^{k} W_{ij}X_j\right|^2 \qquad (1)$$

with the following constraints

$$\begin{cases} \sum_{j=1}^{k} W_{ij} = 1 \;\; if\; X_j \in N_i(X_i) \\ \quad W_{ij} = 0 \;\; if\; X_j \notin N_i(X_i) \end{cases} \qquad (2)$$

where $N_i(X_i)$ denotesthe k nearest neighbors of the point X_i. Minimizing ε_i subject to the above constraints is a constrained least squares problem.

Step 3 Compute the optimal low dimensional embedding Y based on the weight matrix W obtained in Step2 to solve the following optimization function:

$$\varepsilon(Y) = argmin\sum_i\left|Y_i - \sum_{j=1}^{k} W_{ij}Y_j\right|^2 = tr(YTMY) \qquad (3)$$

using the constraints $\frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^T = 1$ and $\sum_{i=1}^{n} Y_i = 0$. With the weight matrix W, a sparse, symmetric and positive semi-definite matrix M can be defined as follows.

$$M = (I - W)^T(I - W) \qquad (4)$$

Eq.4 can be minimized by finding the eigenvectors with the smallest eigenvalues of the sparse matrix M.

LLE then computes the bottom d + 1 eigenvectors of M associated with the d + 1 smallest eigenvalue. The first eigenvector with an eigenvalue nearest to zero is excluded. The remaining d eigenvectors yield the final embedding Y.

Relieff

The Relieff algorithm was proposed as an extension of Relief by Kononenko in 1994 to handle multi-class data[32]. It is a simple but efficient procedure to estimate the qualities of attributes in problems with strong dependencies between attributes and is typically applied in data pre-processing as a feature subset selection method.

Given a randomly selected instance from a class L, ReliefF determines k near hits from the same class referred to as nearest hits H and k near misses from each different class referred to as nearest misses M. Finally, it updates the weight of each feature based on H and M. Repeating the above steps m times, the weights for each feature will be generated.

The ReliefF algorithm is shown below.

Input: a training set D, the number of iteration m, the number of the nearest neighbor k, the number of features n, each sample is a vector

```
of features $A_i$, predefined feature weight threshold δ.
    Output: feature subset S constituted by features whose weights are all greater than the weight threshold δ.
    Step1. Let S=∅, set all feature weights W[A]=0.
    Step2. For j=1 to m do
    select an instance $R_i$ from D randomly
    find out k nearest neighbors $H_i(i = 1,2,\ldots,k)$ from the same class and k nearest neighbors $M_i(C)$ $(i = 1,2,\ldots,k)$ from each different class C.
    For $t$=1 to $n$ do
        $$W[A] = W[A] - \sum_{j=1}^{k} \frac{diff\left(A, R_j, H_j\right)}{m * k} + \sum_{C \neq class(R_i)} [\frac{P(C)}{1 - P(class(R_i))} \sum_{j=1}^{k} diff\left(A, R_j, M_j(C)\right)]/(m * k);$$
     End
    End
```

Function $\mathrm{diff}(\mathrm{A}, \mathrm{I}_1, \mathrm{I}_2)$ calculates the difference between the values of the attribute A for two instances I1 and I2, here it was defined as $diff(A, I_1, I_2) = \frac{|value(A,I_1) - value(A,I_2)|}{max(A) - min\,(A)}$.

Relevant component based LLE(RLLE)

The LLE algorithm attempts to compute low dimensional embedding wherein the local configurations of nearest neighbors are preserved. Under appropriate conditions, such as where the manifold is well-sampled, each data point and its neighbors lie on or close to a locally linear patch of the manifold, the algorithm is effective at deriving such embedding solely from the geometric properties of the nearest neighbors in the high dimensional space

However, for gene expression profile data, the sample number varies from dozens to hundreds; thus, it is impossible to generate the dense and uniform distribution required by the LLE algorithm. Simultaneously, the sample dimensionality is high and it contains many irrelevant or redundant features, and other noise data. Such noise data interferes with sample distance calculations. Ultimately, sample points for the same class were acquired that were not necessarily neighbored in the Euclidean distance space and not normally distributed in the same manifold. In this instance, reducing the dimensionality of neighboring preserved data in the LLE algorithm will not necessarily improve data classification accuracy.

To avoid this result, our research proposes a relevant component based LLE dimensionality reduction algorithm. Knowing the class information of certain samples, we can calculate the correlation between the samples' attributes and classes, eliminate the attributes that are significantly irrelevant to the classes, and generate the weight matrix V in accordance with the relevance scale.

We assume that the data are represented as feature vectors. Our method modifies the feature representation of the data space using the linear transformation V such that the Euclidean distance in the transformed space is less affected by irrelevant variability, and the neighboring for the sample points in the same class is enhanced.

Let us assume xi (i belongs to 1~N) composes the sample points. N is the number of sample points. Each xi is a D-dimensional vector.

First, the ReliefF algorithm was used to determine the weight of each gene, and the irrelevant genes were eliminated, so the dimensionality became D'. Next, each gene expression data set was multiplied by its relevant weight, and the distance between new sample points was re-calculated as follows:

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T A (x_i - x_j)} \qquad (5)$$

A is the rescaling transformation. When A is a unit matrix, the distance remains the Euclidean distance.

Similarly, the Euclidean distance in the second step of the LLE algorithm was replaced, thus, Eq.1 was converted as follows.

$\varepsilon(W) = \sum d^2\left(X_i - \sum_{j=1}^{k} W_{ij} X_j\right) = \sum\left(X_i - \right.$ *j=1kWijXjTAXi–j=1kWijXj* (6)

The constraint condition $\begin{cases} \sum_{j=1}^{k} W_{ij} = 1 \;\; if\ X_j \in N_i(X_i) \\ \quad W_{ij} = 0 \;\; if\ X_j \notin N_i(X_i) \end{cases}$ remained unchanged. After optimization, W was obtained, and the low-dimensional embedding was calculated.

The algorithm procedure is as follows:

1: Calculate the relevant weight between each gene attribute and sample class using the ReliefF algorithm;

2: Delete the irrelevant or redundant attributes;

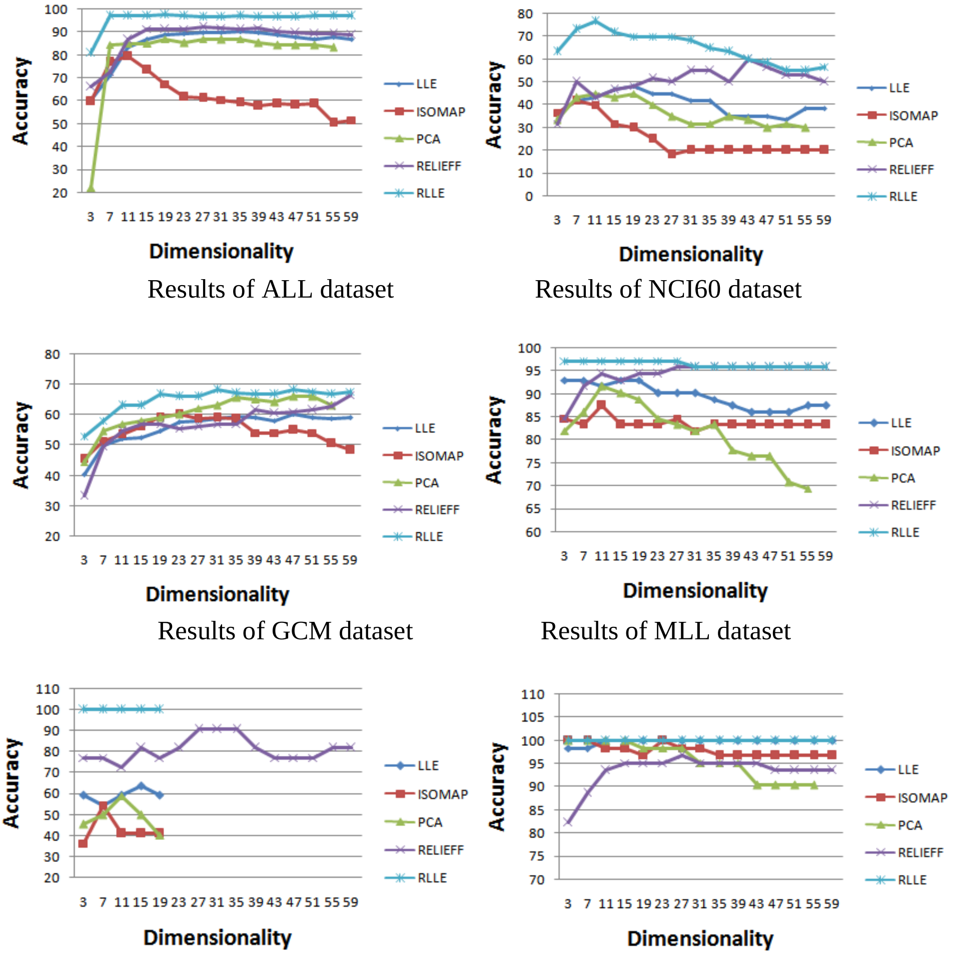
3: Calculate the distance between the new samples with the obtained weight vector V and matrix A;

4: Obtain the K nearest neighbor points of each sample and calculate the locally linear embedding based on the new distance.

Results and discussion

Each gene dataset was formatted as a MATLAB data structure file(.mat). The algorithm output was written as a CSV file, which contains information such as the data matrix after the dimensionality is reduced, the feature number and the category label. Next, the Native Bayesian classifier and 10-fold cross-validation in the WEKA kit were applied to classify the data.

Figure 1 is the classification results for 6 datasets after the dimensionality is reduced. Table 2 shows the optimal classification results for 6 datasets after the dimensionality was reduced to 3~60 dimensions under the conditions no treatment and treatments with the PCA, LLE, ISOMAP and ReliefF algorithms.



Results of ALL dataset

Results of NCI60 dataset

Results of GCM dataset

Results of MLL dataset

Results of Hbc dataset

Results of LYM dataset

Figure 1. Comparison between the RLLE and PCA, ISOMAP, LLE and RELIEFF algorithms. This figure describes the Native Bayes classification results when the dimensionality for the six datasets was reduced to 3~60.

Table 2 A comparison of the PCA, LLE, ISOMAP, ReliefF and RLLE algorithms (optimal native Bayes classification results after the dimensionality was reduced to 3~60)

	ALL	MLL	HBC	NCI60	GCM	SRBCT	LYM
NO	56.85	94.44	72.72	36.66	67.17	84.09	91.93
ReliefF	92.34	95.83	90.9	60	66.66	73.86	96.77
PCA	87.09	91.66	59	45	66.16	82.95	100
LLE	90.32	93.05	63.63	48.33	60.1	80.68	100

ISOMAP	79.43	87.5	54.54	41.66	59.09	84.09	100
RLLE	97.17	97.22	100	76.66	68.18	85.22	100

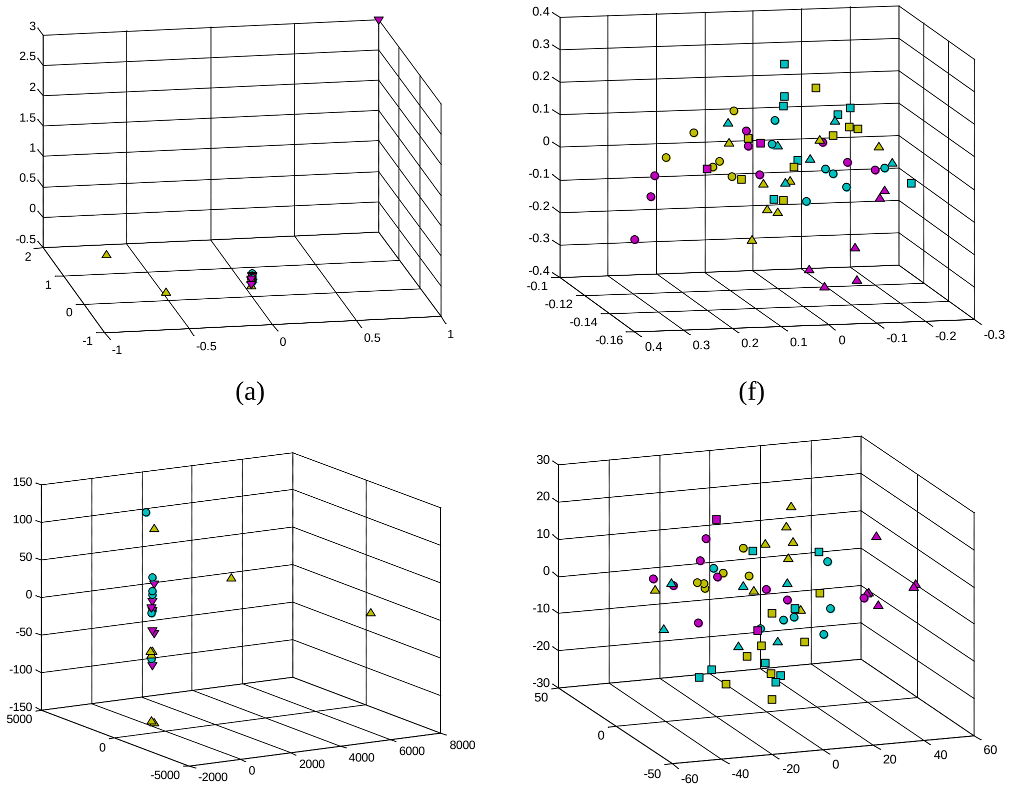
The above comparison results show that:

For the data sets herein, the PCA, LLE and ISOMAP algorithms for reducing dimensionality do not always improve the classification accuracy. Both the linear and nonlinear dimensionality reduction did not determine the classification results after the dimensionality was reduced. In other words, they may produce better or worse classification results. Compared with the linear dimensionality reduction algorithms PCA, the nonlinear dimensionality reduction algorithm LLE and ISOMAP did not significantly improve the classification accuracy.

The ReliefF algorithm is a supervised feature selection approach, the number of selected genes is not constrained by the number of samples, and most often, the classification outcome was better than from unsupervised algorithms such as PCA, LLE and ISOMAP.

The RLLE algorithm is shown to achieve better performance comparing with other dimensionality reduction algorithms on almost all datasets. The experimental results and comparisons demonstrate that the proposed method is highly effective.

Figure 2 presents a three-dimensional scatter diagram of HBC and GCM processed using the PCA, LLE, ISOMAP, ReliefF and RLLE algorithms respectively. The diagram shows that the RLLE algorithm more efficiently distinguished the data.

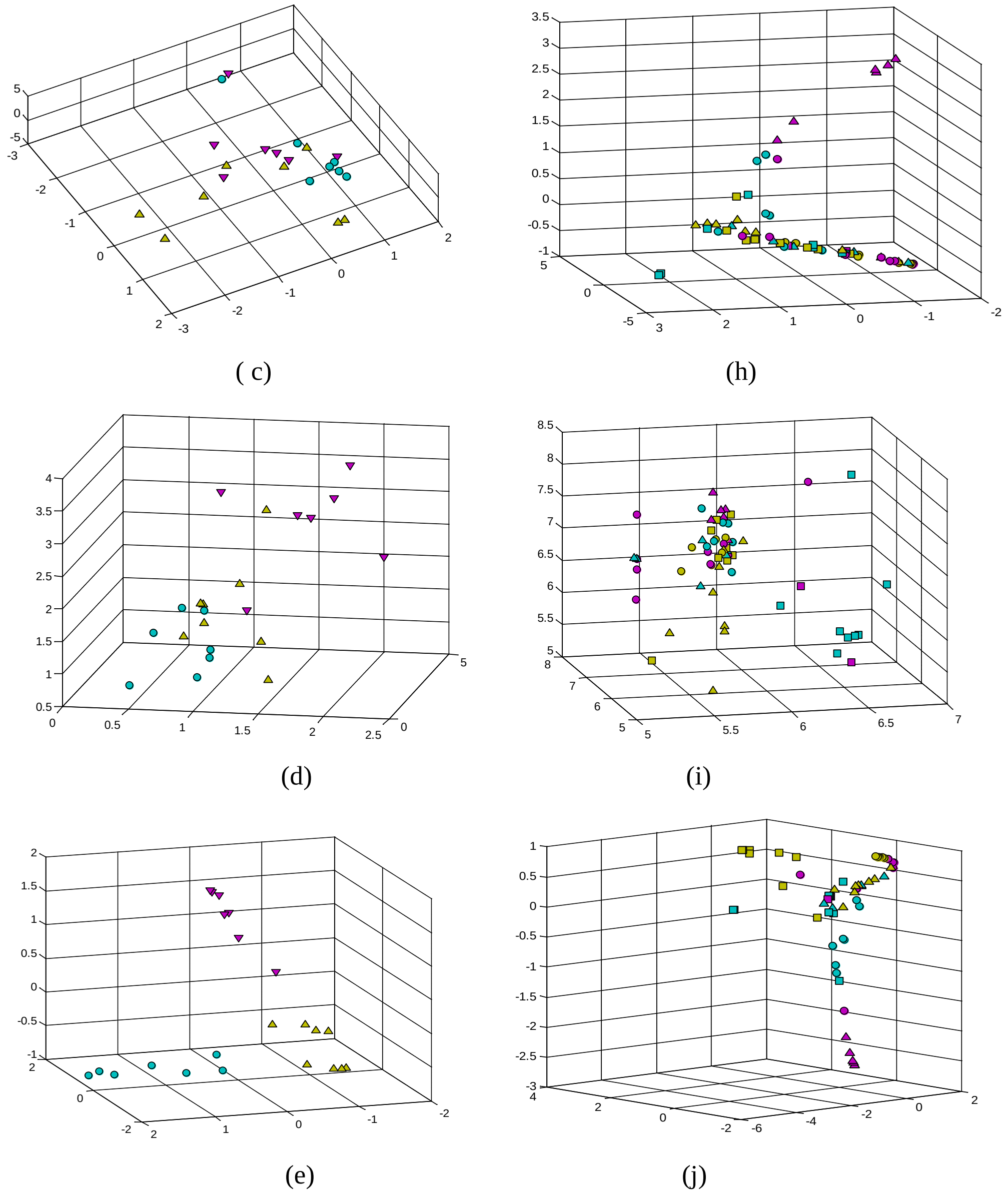


(a)

(f)

(b)

(g)



(c) (h)

(d) (i)

(e) (j)

Figure 3. Embedding graphs generated by plotting the three most dominant embedding vectorsfor (a) PCA,(b) ISOMAP, (c)LLE, (d) RELIEFF, and (e) RLLEusing the HBC data set in the left column. The right column shows the embedding results for the NCI60 dataset using(f) PCA, (g)ISOMAP, (h) LLE, (i) ReliefF and (j)RLLE.

Because the irrelevant and redundant features were eliminated, and the weights were assigned to the dimensions based on the relevance between the attributes and categories, more same-class samples are neighbors, which is important for classification. Table 3 shows the average number of same-class samples in five-neighbor samples after applying the Euclidean distance equation and the distance equation from our research. The table shows that the improved distance calculation equation increased the number of same-class samples among neighboring points.

Table 3. Average number of same-class samples in 5 neighbor samples

	Hbc	Lym	All	Mll	Nci60	Gcm
Euclidean distance equation	2.23	4.74	4.19	3.84	1.86	2.25
Our distance equation	3.86	4.95	4.77	4.69	2.83	2.77

Conclusions

Data mining and analysis for microarray analyses has rapidly garnered interest in recent years. The large number of gene expression profiles coupled with limited number of samples provides an immense space for genomic dimensionality reduction and selection. The goal of this study was to develop an algorithm to improve the accuracy of microarray classifications. In this paper, we present an effective approach that first obtains different weights through calculating the relevance between each attribute and class. Next, a new distance is calculated. Compared with Euclidean distance, the new distance could render more same-class sample points as neighbors. Ultimately, based on the new distance, locally linear embedding was calculated to reduce the dimensionality.

The classification experiment after the dimensionality was reduced demonstrated that for each data set, the RLLE algorithm produced more effective classification results than the PCA, ISOMAP, LLE and RELIEFF approaches. The algorithm can aid biological researchers in classifying different classes, such as cancer and noncancerous, which are difficult to classify due to high-dimensionality.

Acknowledgements

The authors thank the anonymous reviewers and editors for their valuable suggestions and comments for improving this paper. We also wish to thank the National Natural Science Foundation of China(Grants No. 61001144) and the the Key Research Item of Science and Technology of Fujian Province(Grant No. 2014Y0073) for supporting this work.

References

1. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286(5439), p.p.531–537.
2. U. Alon, N. Barki et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probes by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96(12), p.p. 6745–6750
   A. Ben-Dor, L. Bruhn et al. (2000) Tissue classification with gene expression profiles. *J. Comput. Biol*, 7(3-4), p.p. 559–583
3. Wang Y, Tetko I V, Hall M A, et al. (2005) Gene selection from microarray data for cancer classification-a machine learning approach. *Computational biology and chemistry*, 29(1), p.p.37-46
4. Ghosh A, Chandra Dhara B, De R K. (2014) Selection of genes mediating certain cancers, using a neuro-fuzzy approach. *Neurocomputing*, vol.133, p.p.122-140.
5. Peng H Y, Jiang C F, Fang X, et al. (2014)Variable selection for Fisher linear discriminant analysis using the modified sequential backward selection algorithm for the microarray data. *Applied Mathematics and Computation*, vol.238, p.p.132-140.
6. Somorjai R L, Dolenko B, Baumgartner R (2003)Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, 19(12), p.p.1484-1491.
7. L. Nanni, A. Lumini, S. Brahnam (2010)Advanced machine learning technique for microarray spot quality classification. *Neural Comput. Appl*, 19 (3), p.p. 471–475.
8. Swain P S, Elowitz M B, Siggia E D (2002)Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20), p.p.12795-12800.
9. Y. Tu, G. Stolovitzky, and U. Klein (2002) Quantitative noise analysis for gene expression microarray experiments, *Proc. Nat. Acad. Sci. USA*, 99(22), p.p. 14031–14036.
10. C. L. Wilson, S. D. Pepper, Y. Hey, and C. J. Miller (2004) Amplification protocols introduce systematic but reproducible errors into gene expression studies. *BioTechniques*, 36(3), p.p. 498–506.
11. Raser J M, O'Shea E K. (2005)Noise in gene expression: origins, consequences, and control. *Science*, 309(5743), p.p.2010-2013.
12. Blanco,R., et al. (2004) Gene selection for cancer classification using wrapper approaches. *Int. J. Pattern Recognit. Artif. Intell*, 18(8), p.p.1373–1390.
13. Inza I, Larrañaga P, Blanco R, et al. (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial intelligence in medicine*, 31(2), p.p.91-103.
14. Peng H, Long F, Ding C. (2005)Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal and Mach Intell*, 27(8), p.p.1226-1238.

15. Dudoit S, Fridlyand J, Speed TP. (2002)Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457), p.p.77-87.
16. Zhang Y, Ding C, Li T. (2008)Gene selection algorithm by combining reliefF and mRMR. *BMC genomics*, 9(Suppl 2), p.p. S27.
17. J. Venna and S. Kaski, (2006)Local Multidimensional Scaling, *Neural Networks*, 19(6-7), p.p. 889-899.
18. C. Shi and L. Chen, (2005)Feature Dimension Reduction for Microarray Data Analysis Using Locally Linear Embedding, Proc. *Third Asia Pacific Bioinformatics Conf.(APBC’05)*, Singapore, p.p. 211-217.
19. J. Tenenbaum, V. de Silva, and J.C. Langford, (2000)A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500), p.p. 2319-2322.
20. S. Roweis and L. Saul, (2000)Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science*, 290(5500), p.p. 2323-2326.
21. M. Belkin and P. Niyogi (2003) Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, *Neural Computation*, 15(6), p.p. 1373-1396.
22. Lee G, Rodriguez C, Madabhushi A. (2008) Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3), p.p.368-384.
23. Shental N, Hertz T, Weinshall D, et al. (2002) *Adjustment learning and relevant component analysis.* Springer: Berlin Heidelberg, p.p.776-790.
24. Bar-Hillel A, Hertz T, Shental N, et al. (2005) Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6(6), p.p.937-965.
25. Yeoh E J, Ross M E, Shurtleff S A, et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer cell*, 1(2), p.p.133-143.
26. Ramaswamy S, Tamayo P, Rifkin R, et al. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26), p.p.15149-15154.
27. Hedenfalk I, Duggan D, Chen Y, et al. (2001) Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344(8), p.p.539-548.
28. Alizadeh A A, Eisen M B, Davis R E, et al. (2000)Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769), p.p.503-511.
29. Armstrong S A, Staunton J E, Silverman L B, et al. (2001) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, 30(1), p.p.41-47.
30. Ross D T, Scherf U, Eisen M B, et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics*, 24(3), p.p.227-235.
31. Kononenko I. (1994) Estimating attributes: analysis and extensions of RELIEF. *Machine Learning: ECML-94.* Springer Berlin Heidelberg, p.p.171-182.