FEATURE DIMENSION REDUCTION FOR MICROARRAY DATA ANALYSIS USING LOCALLY LINEAR EMBEDDING

SHI CHAO AND CHEN LIHUI

School of EEE, Nanyang Technological University,
Republic of Singapore 639798
shichao@pmail.ntu.edu.sg, elhchen@ntu.edu.sg

Cancer classification is one major application of microarray data analysis. Due to the ultra high dimensionality nature of microarray data, data dimension reduction has drawn special attention for such type of data analysis. The currently available data dimension reduction methods are either supervised, where data need to be labeled, or computational complex. In this paper, we proposed to use a revised locally linear embedding(LLE) method, which is purely unsupervised and fast as the feature extraction strategy for microarray data analysis. Three public available microarray datasets have been used to test the proposed method. The effectiveness of LLE is evaluated by the classification accuracy of a SVM classifier. Generally, the results are promising.

1. Introduction

Cancer is a group of diseases characterized by uncontrolled growth and spread of abnormal cells.[4] In most cases, the early detection and treatment can substantially improve the survival rates of cancer patients. Traditionally, cancer diagnosis has been morphological and phenotype based, which maybe complex and deceivable. Cancer genetics, based on analysis of cancer genotypes, provides a valuable alternative in both theory and practice.

Gene expression datasets contain the genotype of many genes relevant or irrelevant to cancer development. Many classification and clustering algorithms have been proposed and tested on gene expression datasets. The results reported in the literature have confirmed the effectiveness of mining cancer information from gene expression data. However, the ultra high dimensionality of gene expression data makes the mining still a non-trivial task. Effective feature reduction tools are in great needs. We are particularly interested in the ability of using unsupervised methods to select features in high dimensional datasets. In this paper, we describe the proposed method using LLE and SVM for gene expression data analysis.

This paper is organized as follows: Section 2 provides the literature of machine learning techniques used in gene expression mining and necessity of feature reduction; Section 3 introduces our proposed feature extraction method; Section 4 demonstrates the performance of our proposed feature extraction method through experiments on three public available datasets. Finally conclusions are made in Section 5.

2. Background and review

Clustering and classification are extensively studied problems in statistics and machine learning domain. Many algorithms, such as decision tree, linear discriminant analysis, neural network, and the Bayesian network have been proposed and widely applied in practical problems.

Recently years, researchers have paid attention to tumor clustering and classification using gene expression data. Golub(1999) has analyzed leukemia dataset using weighted voting to classify predominant cancer types,[6] Alizadeh et al.(2000) has studied lymphoma dataset using hierarchical clustering methods,[2] U.Alon has applied Two-way data clustering(CTWC) on colon dataset to classify genes and samples interactively.[3] Their work has generally given positive support to gene expression data analysis as both an exploratory and diagnosis tool.

Data dimension reduction is a commonly applied preprocessing step in data mining applications. It is especially useful in mining gene expression datasets, which are usually of very high dimension, i.e, in the range of thousands, and contain very few samples, usually less than 100.

With such a huge attribute space, it is almost certain that all classifiers built upon it would be prone to *overfitting*. The small sample size makes it even worse.[8] Since most genes are known to be irrelevant for class distinction, their inclusion would not only introduce noise and confuse the classifiers, but also increase the computation time. Gene selection prior to classification would help in alleviating these problems. With the noise from the irrelevant genes removed, the biological information hidden within will be less obstructed.

Also, experiments have shown that gene selection prior to classification improves the classification accuracy of most classifiers.[8] Besides performance, the reduction from the range of thousands of features to tens will greatly reduce the running time of most of the classifiers.

3. Methods and theory

For classification purpose, the expression data samples are normally labeled and divided into training set T and test set S. The classifier is a function *Class* with two arguments, T and s, with T denotes the training samples and s is a testing sample from the test set S. The function *Class* returns a class prediction for sample s. The classification *accuracy* is measured by the number of correct predictions made by the classifier over the test set S using the function *Class* trained on the training samples.

By employing the feature reduction tools, we target to improve the performance on microarray data analysis, while also reduce the computational burden of the classifiers.

In our designed experiment, LLE is first applied to the expression data to reduce the dimensionality from several thousands to a reasonable small number. Then SVM classifier is applied, and the Leave-one-out classification accuracy is used to evaluate the effectiveness of the feature reduction performance.

In this section, the three key components used in the proposed method namely, LLE,

SVM and the similarity measure are discussed in details.

3.1. *Fractional metrics*

Consider a dataset consisting of n points, where each point is described by a d-dimensional vector, define distance measurement

$$dist = \left[\sum_{i=1}^{d} (x_i - y_i)^{1/k}\right]^k \tag{1}$$

between any two data points x and y.

$$relative\ contrast \qquad r_j = \frac{Dmax_{d,j}^k - Dmin_{d,j}^k}{Dmin_{d,j}^k} \tag{2}$$

for each point j.
where $Dmax_{d,j}^k$ and $Dmin_{d,j}^k$ denote the farthest and nearest distance from all other points in the dataset to the point j we consider respectively. When k=1/2, $dist$ is the Euclidean distance, which is a special case of fractional metrics.

For most of the data mining algorithms, the distance measurement metric is necessary and crucial. Unless otherwise specified, most of the time Euclidean distance is chosen. However, in the high dimensional space, the contrast nature of distance measurement become shattered with increase of dimensionality.
It is proofed that the relative contrast r will degrade as increase of dimensionality d.[1] To make the distance metrics a more meaningful proximity measure, we have to take a practical approach, which relaxes the k to take value other than $1/2$ as for Euclidean distance. The relative contrast r defined in Eq 2 on the specific dataset can be used as a guide for choice of specific k.

$$k = \arg\max_k \sum_{j=1}^{n} \frac{Dmax_{d,j}^k - Dmin_{d,j}^k}{Dmin_{d,j}^k} \tag{3}$$

In real practice, we can fine tune the fraction value k in the range of $0.1 \sim 0.5$ by a step size of 0.1. The one which reaches best performance is chosen.

3.2. *Locally linear embedding(LLE)*

Locally linear embedding was first proposed by Roweis[11] as an unsupervised learning algorithm that computes low dimensional, neighborhood preserving embeddings of high-dimensional inputs.

The LLE algorithm is based on simple geometric intuitions. Suppose the data consist of N real-valued vectors $\vec{X}_i$, each of dimensionality D, sampled from some underlying manifold. Provided there is sufficient data (such that the manifold is well-sampled), we expect each data point and its neighbors to lie on or close to a locally linear patch of the manifold. We characterize the local geometry of these patches by linear coefficients that

reconstruct each data point from its neighbors. Reconstruction errors are measured by the cost function

$$\epsilon(W) = \sum_{i=1}^{N} | \vec{X}_i - \sum_{j=1}^{k} W_{ij}\vec{X}_j |^2 = \sum_{i=1} \epsilon^i(W) \tag{4}$$

$$\epsilon^i(W) = | \sum_{j=1}^{k} W_j^i (x_i - x_j) |^2 = \sum_{j=1}^{k}\sum_{m=1}^{k} W_j^i W_m^i Q_{jm}^i \tag{5}$$

$$Q_{jm}^i = (x_i - x_j)^T (x_i - x_m) = (D_{i,j} + D_{i,m} - D_{j,m})/2 \tag{6}$$

which adds up the squared distances between all the data points and their reconstructions. The weights W_{ij} summarize the contribution of the jth data point to the ith reconstruction.The optimal weights W_{ij} are found by solving a least-squares problem.

The actual process of LLE is as follows:

(1) Assign neighbors to each data point $\vec{X}_i$ (for example by using the K nearest neighbors).
(2) Compute the weights W_{ij} that best linearly reconstruct X_i from its neighbors, solving the constrained least-squares problem in Eq. 4.
(3) Compute the low-dimensional embedding vectors $\vec{Y}_i$ best reconstructed by W_{ij}, minimizing embedding cost function by solving a linear algebra problem.

Seen from Eqn 6, the cost function can also be expressed as triangular relationship among pairwise distance, i.e., the LLE implementation does not rely on specific data sample similarity measurement metrics. In our implementation, the similarity measure $D_{i,j}$ is replaced with fractional metrics rather than Euclidean as in the original LLE implementation. The reason of using fractional metrics is discussed in section 3.1.

3.3. *Classification and evaluation*

SVM was originally introduced by Vapnik and widely used in data mining applications.[8] In this project, we use it as a classification tool to verify the effectiveness of our feature reduction algorithm.

Multi-class classification The SVM is special tailored for binary-class classification problem. In case of multi-class problem, we split the problem into k binary-classification problem, where k equals to the number of class, and each class is classified versus all other classes in the dataset. The classification accuracy is calculated by adding up the correctly classified samples over all classes.

Cross validation The gene expression dataset usually contains very few data samples, to analyze the performance of the classification, we need to estimate the generalized classification error. Cross validation is such a technique based on methods of resampling. In this paper, we propose to use *Leave-one-out* cross-validation.

Table 1. Gene expression datasets

Dataset	Classes	No of genes	No of Samples	Origin
Leukemia	2	7129	72	Golub et al.
Lymphoma	3	4026	62	Alizadeh et al.
Colon	2	2000	62	Alon et al.

In Leave-one-out cross-validation, the classifier is trained k (k equals the number of samples) times, each time leaving out one sample from training, but using only the omitted sample to compute whatever error criterion that interests you.

4. Experiment

4.1. *Software*

The whole program is implemented in Matlab(version 6.5) environment. The locally linear embedding is implemented using the coding provided by the original author.[9] The SVM we used is the SVM^{light} software package version 5.00 by Joachims.[7] It is available at http://svmlight.joachims.org/. The SVM is a executable program, we have used the script from Anton Schwaighofer to interface it with Matlab.[10]

4.2. *Dataset*

The proposed feature extraction strategy has been implemented and tested on three public available microarray datasets, namely Leukemia, Lymphoma, and Colon.

4.3. *Results*

We have compared the performance of the revised LLE with some other classical feature reduction techniques on three microarray datasets discussed in Sec 4.2. The results of other techniques are extracted from a survey reported by Sung(2003).[5] The results are listed in Table 2. The feature reduction techniques being compared include unsupervised methods, such as principal components(PC), and supervised methods, such as signal to noise ratio(SNR)and correlational coefficient(CC). It is observed that our proposed LLE methods are consistently better than the above methods in all three gene expression datasets.

We have also compared the performance of our revised LLE using fractional metrics with the original LLE implementation using the Euclidean distance. Fig. 1 shows the performance comparison for leukemia dataset. The improvement in classification accuracy is obvious when k takes value other than 0.5(as for Euclidean).

This LLE feature extraction process is purely unsupervised, and it does not need the existence of the class labels. Also, the feature extraction calculation is a simple linear algebra problem, and it does not involve any training or iteration process, such that the process can be extremely fast.

The feature reduction on leukemia and colon datasets clearly outperforms its on lymphoma, which may suggest the LLE feature extraction is more suitable for binary-class feature reduction than multiple(> 2) class problems. Here, we try to explain the difference

Table 2. Comparisons of classification accuracy between LLE and other supervised methods.

Dataset	Feature selection	Classifier	Accuracy	Fraction, k
Leukemia	LLE	SVM(RBF kernel)	95%	0.4
	PC	SVM(RBF kernel)	79%	N.A.
	SNR	SVM(RBF kernel)	59%	N.A.
	CC	SVM(RBF kernel)	85%	N.A.
Lymphoma	LLE	SVM(RBF kernel)	85%	0.4
	PC	SVM(RBF kernel)	60%	N.A.
	SNR	SVM(RBF kernel)	76%	N.A.
	CC	SVM(RBF kernel)	65%	N.A.
Colon	LLE	SVM(RBF kernel)	91%	0.3
	PC	SVM(RBF kernel)	65%	N.A.
	SNR	SVM(RBF kernel)	65%	N.A.
	CC	SVM(RBF kernel)	56%	N.A.

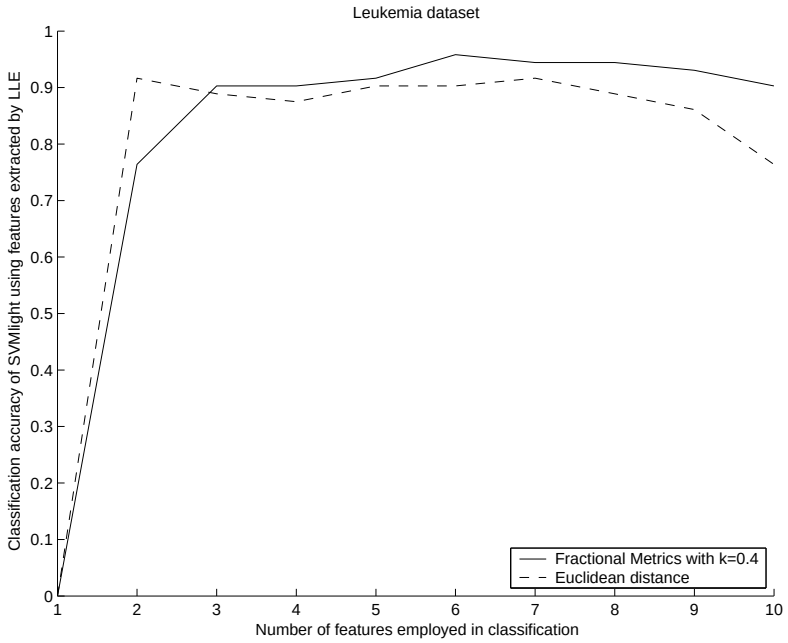

Figure 1. The performance of LLE feature extraction on leukemia dataset

from the working mechanism of LLE. LLE has made a general assumption that the samples lies on a smooth hyperplane which can be uniquely determined by the neighborhood coefficient W_{ij} among points, and for each data point, its nearest neighbors should be just its neighbors along the hyperplane. However, the distribution of data points maybe quite

random and the hyperplane formed maybe heavily twisted, so that the nearest neighbors of certain point may not be its neighbors along the hyperplane, but very distant apart. Intuitively, the hyperplane formed by data points will be far more complex for datasets with more classes. Therefore, LLE may perform better in binary-classification problems.

5. Conclusion

In this paper, we proposed to use Locally Linear Embedding methods for gene expression data dimensional reduction. The effectiveness is demonstrated with the help of the SVM classifier on three public available microarray datasets. The classification accuracy achieved with such feature extraction strategy is comparable to supervised feature reduction methods. The performance deviation on binary and multiple class classification is analyzed and tentatively justified.

References

1. Charu C. Aggarwal et al. On the surprising behavior of distance metrics in high dimensional spaces. In *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*, volume 1973, pages 420–434. Springer, 2001.
2. Alizadeh et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(503-511), 2000.
3. U. Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays. *PNAS*, 96(12):6745–6750, 1999.
4. Inc American Cancer Society. American cancer society homepage. *http://www.cancer.org*, 2004.
5. Sung-Bae Cho et al. Machine learning in DNA microarray analysis for cancer classification. In *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics*, volume 19, 2003.
6. T. R. Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, (286):531–537, 1999.
7. Thorsten Joachims. Svmlight support vector machine. In *http://svmlight.joachims.org/*, 2002.
8. Ying Lu and Jiawei Han. Cancer classification using gene expression data. *Information Systems*, pages 243–268, 2003.
9. Sam T. Roweis and Lawrence K. Saul. Locally linear embedding. In *http://www.cs.toronto.edu/ roweis/lle/code.html/*, 2004.
10. Anton Schwaighofer. Matlab interface to svm light. In *http://www.cis.tugraz.at/igi/aschwaig/software.html/*, 2004.
11. S.T.Roweis and L.K.Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2000.