Combining LPP with PCA for Microarray Data Clustering

Chuanliang Chen, Rongfang Bie, and Ping Guo, Senior Member, IEEE

Abstract—DNA Microarray technique has produced large amount of gene expression data. To analyze these data, many excellent machine learning techniques have been proposed in recent related work. In this paper, we try to perform the clustering of microarray data by combining the recently proposed Locality Preserving Projection (LPP) method with PCA, i.e. PCA-LPP. The comparison between PCA and PCA-LPP is performed based on two clustering algorithms, K-means and agglomerative hierarchical clustering. As we already known, clustering with the components extracted by PCA instead of the original variables does improve cluster quality. Moreover, our empirical study shows that by using LPP to perform further process the dimensions of components extracted by PCA can be further reduced and the quality of the clusters can be improved greatly meanwhile. Particularly, the first few components obtained by PCA-LPP capture more information of the cluster structure than those of PCA.

I. INTRODUCTION

Monitoring tens of thousands of genes in parallel under different experiment environments or across different tissue types provides a systematic genome-wide approach to solve the problems such as gene functions in various cellular processes, gene regulations in various cellular signaling pathways and gene expression differentiation in various diseases or drug treatments [1]. DNA microarray, one of the major methods of measurement for gene expression data, has become a popular technique for collecting large amount of gene expression data that is required to study the behavior of a cell [2, 3].

Clustering technique manifests its crucial power as the first step in extracting information from the mass of gene expression data set, and plays an important role in the discovery and understanding of various classes and subclasses of cancers [1, 2]. Clustering technique has played a major role in analyzing DNA microarray gene-expression data. There have been various clustering techniques used for microarray data analysis, most notably by hierarchical clustering [15], K-means clustering [16] and Self-Organizing Maps (SOM) [17]. For improving our understanding of the underlying biological phenomena, these clustering techniques have made great contributions. However, one main challenge in this task is the high dimensionality of the gene expression data. Principal Component Analysis (PCA), a classical feature extraction that has been widely used in the area of pattern recognition, can increase the overall performance of clustering [5]. Locality Preserving Projections (LPP) as an alternative to PCA is a recently proposed method for dimensionality reduction [6]. The primary consideration of LPP is to preserve the neighborhood structure [6] of the data by building a graph, which may be in favor of clustering. Since LPP can optimally preserve the neighborhood structure, it may increase the performance of clustering much more by extracting features from components extracted by PCA. To find out the potential dimension reduction ability of LPP for microarray data analysis is the motivation of our work.

In this paper, a novel dimension reduction method, namely PCA-LPP, is proposed to extract features from high dimensional microarray data and the potential of it for microarray data clustering is also illustrated.

The remainder of this paper is organized as follows. Section 2 describes the above two feature extraction method in brief, LPP and PCA, and the method PCA-LPP is also proposed in this section. Two clustering algorithms and two performance measures are presented in section 3. Section 4 describes the benchmark microarray dataset and presents the results of our experiments. Finally, the main conclusions are presented in section 5.

II. FEATURE EXTRACTION

In this section, we briefly review Principal Component Analysis (PCA) and Locality Preserving Projection (LPP) at first. And then we propose PCA-LPP based on these two feature extraction methods.

At first, the generic problem of linear dimensionality reduction problem is formally described as follows:

1. Given the original data $X = \{x_1, x_2, \ldots, x_m\}$ in high-dimensional space $R^n$.

2. Find a matrix $A$ that transforms the original data points into a new set of data points $Y = \{y_1, y_2, \ldots, y_m\}$ in a low-dimensional space $R^l (l \ll n)$, such that $y_i$ “represents” $x_i$, where $y_i = A^T x_i$.

A. PCA and LPP

PCA: Principal Component Analysis, or Karhunen-Loeve transform, is a powerful technique for extracting structure from possibly high-dimensional data sets [8]. It reduces the dimensionality of the data set by transforming the data to a new set of variables (the principal components) to summarize the features of the data. In fact, PCA is equivalent to applying Singular Value Decomposition (SVD) on the covariance matrix of the data and the procedure of PCA is
The graph can also be constructed by an edge if a sparse symmetric matrix with choosing connected by an edge. We can construct the graph by their eigenvalues, similarity function may also be used). Let parameter edge). Heat kernel similarity is used to set the weight (other similarity function may also be used). Let parameter $t\in R$. If nodes $i$ and $j$ are connected, put $W_{ij} = e^{-\frac{1}{2}(i-j)^2}$.

Step 3: Eigenmaps. Eigenvectors and eigenvalues are calculated for the generalized eigenvector problem:

$$XLX^T a = \lambda XDX^T a$$

(1)

where $D$ is a diagonal matrix whose entries are column(or row, since $W$ is symmetric) sums of $W$. Let the column vectors $a_1, \cdots, a_l$ be the solutions of (1), ordered according to their eigenvalues, $\lambda_1 < \cdots < \lambda_l$. Thus the embedding is as follows [6]:

$$x_i \rightarrow y_i = A^T x_i, A = (a_1, a_2, \cdots, a_l)$$

(2)

where $y_i$ is a $l$-dimensional vector, and $A$ is a $n \times l$ matrix.

B. The method PCA-LPP

Although LPP can preserve locality information, which may be in favor of clustering, the problem is that when $X$ is in the high dimensional manifold, which is the case for microarray data analysis, LPP will fail due to the curse of high-dimension. To utilize the ability of preserving locality information of LPP and get better subspace $Y$, we propose that the data, which LPP deals with, is the components extracted by PCA, i.e. the novel method PCA-LPP. The procedure of PCA-LPP is stated below:

Step 1: Extract all eigenvectors whose eigenvalues are greater than $\varepsilon$ by PCA, where $\varepsilon \in R$ and is set to be $10^{-12}$ in this paper. These eigenvectors compose a matrix $V$.

Step 2: Transform $X$ to $T$: $T = V^T X$.

Step 3: Extract features from $T$ by using LPP and then get $l$-dimensional subspace $Y$.

Since PCA can preserve global information and LPP can preserve locality information, clustering with $Y$ obtained by performing LPP-PCA possibly improves cluster quality further. Our empirical study shows that $Y$ does capture much more information of the cluster structure than the same dimensional subspace transformed by PCA.

III. CLUSTERING ALGORITHMS AND PERFORMANCE MEASURES

A. Clustering Algorithm

Hierarchical Clustering: Hierarchical clustering is a common method used to determine clusters of similar data points in multidimensional spaces [9]. Hierarchical clustering techniques produce a nested sequence of partitions, with a single, all-inclusive clusters at the top and singleton clusters of individual points at the bottom [10]. The algorithm combines two clusters from the next lower level (or splits a cluster from the next higher level) in each intermediate level. One common basic approach to generate a hierarchical clustering is agglomerative algorithm, which is used in our experiments as well. The traditional agglomerative hierarchical clustering procedure is described as follows [10]:

1. Compute the similarity between all pairs of clusters, i.e., calculate a similarity matrix whose $i$th row gives the similarity between the $i$th and $j$th clusters.

2. Select the most similar pair of clusters and merge them into a single cluster, i.e., the total number of clusters is reduced by one.

3. Update the similarity matrix between the new cluster and each of the original clusters.

4. Repeat steps 2 and 3 until a single cluster remains or desirable number of clusters is achieved.

Hard Partitioning Clustering: To construct $k$ clusters, in contrast to hierarchical techniques, a partitioning method creates exactly $k$ clusters at once and then iteratively improves the partitioning by moving data objects from one group to another. $K$-means is the most well-known partitioning method and the basic $K$-means is stated below:

1. Select $k$ initial centroids.

2. Assign all points to their nearest centroid.

3. Update the centroid of each cluster.

4. Repeat steps 2 and 3 until the centroids do not change or change little.

Euclidean Distance: The Minkowski distances $D_p(X,Y) = \left(\sum_i |x_i - y_j|^p\right)^{\frac{1}{p}}$ are metrics commonly used in clustering algorithms. Euclidean distance is a Minkowski distance with $p = 2$. In our experiment, euclidean distance is used as the similarity metric.

B. Performance Measure

We use two performance measures altogether in our experiments.

Mutual Information: Though entropy and purity are suitable for measuring a single cluster’s quality, they are both biased to favor smaller clusters. Instead, we use a
symmetric measure called mutual information to evaluate the overall performance. The mutual information is a measure of the additional information known about one expression pattern when given another [11]. The definition of Mutual Information is shown in Eq. 3:

$$MI(A, B) = H(A) + H(B) - H(A, B)$$  \hspace{1cm} (3)

where \(H(A)\) is the entropy of the gene expression pattern \(A\) and can be calculated by using Eq. 4:

$$H(A) = -\sum_{x \in A} p(x) \log_2(p(x))$$  \hspace{1cm} (4)

Yet entropy and mutual information are computed using discrete probabilities. Thus, to calculate entropy, we use a histogram technique [11]. Because mutual information criterion can successfully capture the relation between labels and categorizations without a bias towards smaller clusters, it is often used to evaluate the performance of the clustering.

**Accuracy**: The evaluation metric in [12] is also used. The goal of accuracy is to construct a partition that correctly identifies the underlying classes in the given data, creating one cluster for each class [12]. It is possible to view a partition as a relation between two instances, either in the same cluster or in different clusters. For a data set with \(n\) instances, there are \(n(n-1)/2\) unique pairs of instances, and thus there are \(n(n-1)/2\) pairwise decisions reflected in any partition. As a result, we evaluate a partition i.e. the correct partition using Eq. 5.

$$\text{accuracy} = \frac{\text{num(correct decisions)}}{n(n-1)/2}$$  \hspace{1cm} (5)

### IV. Experiment Results

**A. Dataset Description**

Four popular datasets are used in our experiments to test the capability of PCA-LPP. The details about these four datasets are described as follows.

**The Leukemia dataset**: This microarray dataset comes from a study of gene expression in two types of acute leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) [18]. Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing 6,817 human genes. The dataset comprises 47 cases of ALL (38 ALL B-cell and 9 ALL T-cell) and 25 cases of AML.

**The SRBCT dataset**: This microarray dataset comes from a study of four different childhood tumors [19], including Ewing's family of tumors (EWS), neuroblastoma (NB), non-Hodgkin lymphoma (in our case Burkitt's lymphoma, BL) and rhabdomyosarcoma (RMS). The dataset contains 2,308 human genes, and comprises 29 cases of EWS, 11 cases of BL, 18 cases of NB and 25 cases of RMS.

**The Brain Tumor dataset**: This microarray dataset is the version of [14], coming from a study of brain tumor. It consists of 5 types of brain tumor: Medulloblastoma, Malignant glioma, AT/RT, Normal cerebellum, and PNET. The dataset contains 5,921 genes and 90 samples. There are 60 cases of Medulloblastoma, 10 cases of Malignant glioma, 10 cases of AT/RT, 4 cases of Normal cerebellum, and 6 cases of PNET.

**The 9 Tumors dataset**: This microarray dataset contains 9 various human tumor types, NSCLC, Colon, Breast, Ovary, Leukemia, Renal, Melanoma, Prostate, and CNS. There are 5,726 genes, and 60 samples. The dataset contains 9 cases of NSCLC, 7 cases of Colon, 8 cases of Breast, 6 cases of Ovary, 6 cases of Leukemia, 8 cases of Renal, 8 cases of Melanoma, 2 cases of Prostate, and 6 cases of CNS. More details can be found in [13].

**B. Results and Discussion**

The overall results from our empirical study can be summarized as follows:

1. As a whole, the performance of K-means clustering on data transformed by PCA-LPP is much better than on those transformed by PCA and on the original data, using both evaluation metrics.

2. The performance of agglomerative hierarchical clustering, measured by both of the two evaluation metrics, on few dimensions data transformed by PCA-LPP is much better than on data transformed by PCA and on the raw data.

Fig. 1 shows that the best performance, using either of the two clustering algorithms or either of these two evaluation metrics, is achieved by PCA-LPP when the subspace is six dimensions. When using K-means and the dimension of the transformed data is six, the mutual information of the result on data transformed by PCA-LPP is about 0.678, but the mutual information on data transformed by PCA is about 0.420, on the original data is only 0.118. Similarly, the accuracy of results obtained by K-means on the six dimensional data transformed by PCA-LPP and PCA are 85% and 68%, comparing with 57% achieved on the original data. When the dimension of the transformed data is less than eighteen, the performance of agglomerative hierarchical clustering (Agglom) on the data transformed by PCA-LPP is much higher than that on the data transformed by PCA and on the original data. The best results given by agglomerative hierarchical clustering is also achieved when the subspace is six dimensions. However, when the dimension of subspace is more than eighteen, the results on the data transformed by PCA-LPP is approximately the same as those on the other kinds of data.

In Fig. 2, the best clustering performance is also achieved by clustering data transformed by PCA-LPP. The result of K-means on the data transformed by PCA-LPP is all better than on the two other data except a few points. The highest mutual information of K-means based on the PCA-LPP is as high as 0.730, comparing with 0.604 of it on data transformed by PCA and 0.166 on the original data. The highest accuracy of results of K-means on the three data are 85.4%, 80.4% and 64.2% respectively. As for agglomerative hierarchical clustering, when the dimension of the transformed data is less than thirteen, the performance on the data transformed by PCA-LPP is much higher than those on the data transformed by PCA and the original data. When the dimension of subspace is more than thirteen, similarly, the results based on the data transformed by PCA-LPP are approximately the same as results on the other kinds of data.
The clustering results of other two datasets, 9 Tumors and Brain Tumor, are showed in Fig. 3 and Fig. 4. These results are similar to the results on Leukemia and SRBCT datasets: the PCA-LPP method performs much better than PCA. Furthermore, clustering on components extracted by PCA-LPP and PCA can achieve higher mutual information and accuracy than on the original data.
In the rest of this section, we analyze the possible reasons of some observation from the results of our experiments briefly.

Why the performance of clustering algorithms on the first few components extracted by PCA-LPP and PCA is so excellent? PCA can capture global information of the cluster structure. The components extracted by PCA expose clearer cluster structure to K-means and agglomerative hierarchical clustering. The reason why these components extracted by PCA improve agglomerative hierarchical clustering very little when the number of components is bigger is because the process of agglomerating from bottom to top needs more
local information than global information. However, the first
few components extracted by PCA-LPP improve the
performance of agglomerative hierarchical clustering very
much. This is because, as a graph-based algorithm, the
procedure of constructing graph of LPP captures most useful
local information. Furthermore, since LPP of PCA-LPP runs
on the components extracted by PCA, the PCA-LPP method
captures global information of cluster structure meanwhile.
Therefore, clustering methods running on components
extracted by PCA-LPP perform better. Through above
discussion, the excellent performance of agglomerative
hierarchical clustering on first few components extracted by
PCA-LPP is possible because the local information of cluster
structure desired by agglomerative hierarchical clustering is
provided by the procedure of LPP.

Why there are very few points in Fig. 1, Fig. 2, Fig. 3 and
Fig. 4 where the performance of K-means based on data
transformed by PCA-LPP becomes worse than those on the
data transformed by PCA or even on the original data? The
reason of it may be that K-means easily converges to a local
minimum point of the error function.

Overall, the quality of clusters can be greatly improved on
the data transformed by PCA-LPP, and PCA-LPP is a
promising dimension reduction technique for extracting
useful features from microarray data.

V. CONCLUSIONS

In this paper, we propose a novel method PCA-LPP and
use it to perform clustering microarray data. Two different
clustering algorithms, agglomerative hierarchical clustering
and K-means, are used to compare the performance of
PCA-LPP and PCA. The performances of the two clustering
methods are evaluated by using two different performance
metrics, mutual information and accuracy. The results of our
experiments show that clustering based on PCA-LPP
performs much better than clustering based on PCA and on
the original data.

Future work of ours will focus on introducing some other
excellent dimension reduction methods to improve
performance of clustering methods and meanwhile to reduce
complexity. What’s more, we will consider additional
clustering methods in our feature work.

VI. ACKNOWLEDGE

The research work described in this paper was supported by
grants from the National Natural Science Foundation of
China (Project Nos. 10601064, 60675011).

REFERENCES

gene expression data: a survey,” IEEE Trans. on Knowledge and Data
and I. Herskowitz. “The transcriptional program of sporulation in
analysis and display of genome-wide expression patterns,” in Proc.
approach to the clustering of microarray expression data,”
Independent Component Analysis for Gene Expression Data
Advances in Neural Information Processing Systems 16, Vancouver,
Techniques for Embedding and Clustering,” Advances in Neural
Information Processing Systems 14, Vancouver, British Columbia,
Canada, 2002.
analysis as a kernel eigenvalue problem,” Neural Computation, vol. 10,
[9] Clark F. Olson, “Parallel algorithms for hierarchical clustering,”
Agglomerative Clustering Methods for Document Retrieval,” the
networks: functional genomic clustering using pairwise entropy
415–426.
[12] Kiri Wagstaff and Claire Cardie, “Clustering with Instance-
Level Constraints,” in Proc. the Seventeenth International Conference on
Ladd, C., Beheshti, J., Bueno, R., Gillette, M. et al., “Classification of
human lung carcinomas by mRNA expression profiling reveals
13790–13795.
[14] Pomeroy S.L., Tamayo P., Gaasenbeek M., Sturla L.M., Angelo M.,
McLaughlin M.E., Kim J.Y., Goumnerova L.C., Black P.M., Lau C. et
al., “Prediction of central nervous system embryonal tumour outcome
display of genome-wide expression patterns,” In Proc. of National
[16] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM,
“Systematic determination of genetic network architecture,” Nature
Lander ES, Golub TR, “Interpreting patterns of gene expression with
self-organizing maps: methods and applications to hematopoietic
differentiation,” In Proc. of the National Academy of Sciences, USA
JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD,
Lander ES, “Molecular Classification of Cancer: Class Discovery and
[19] Javed Khan, Jun S. Wei, Markus Ringner, Lao H. Saal, Marc Ladanyi,
Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R.
Antonescu, Carsten Peterson, and Paul S. Meltzer, “Classification and
diagnostic prediction of cancers using gene expression profiling and