A Framework for High Dimensional Data Reduction in the Microarray Domain

Ali Anaissi[#1], Paul J. Kennedy[#2], Madhu Goyal [#3]

Center of Quantum Computation and Intelligent Systems (QCIS), Faculty of Engineering and Information Technology (FEIT),
University of Technology, Sydney (UTS)
P.O. Box 123
Broadway, NSW 2007
Australia

[1] aanaissi@eng.uts.edu.au
[2] Paul.Kennedy@uts.edu.au
[3] madhu@it.uts.edu.au

***Abstract*— Microarray analysis and visualization is very helpful for biologists and clinicians to understand gene expression in cells and to facilitate diagnosis and treatment of patients. However, a typical microarray dataset has thousands of features and a very small number of observations. This very high dimensional data has a massive amount of information which often contains some noise, non-useful information and small number of relevant features for disease or genotype. This paper proposes a framework for very high dimensional data reduction based on three technologies: feature selection, linear dimensionality reduction and non-linear dimensionality reduction. In this paper, feature selection based on mutual information will be proposed for filtering features and selecting the most relevant features with the minimum redundancy. A kernel linear dimensionality reduction method is also used to extract the latent variables from a high dimensional data set. In addition, a non-linear dimensionality reduction based on local linear embedding is used to reduce the dimension and visualize the data. Experimental results are presented to show the outputs of each step and the efficiency of this framework.**

***Keywords*— Feature Selection; Linear dimension Reduction; Non-Linear Dimension Reduction.**

I. Introduction

Dimensionality reduction, as a significant and important tool in bioinformatics, has become an active research area [1, 2, 3]. The purpose of dimensionality reduction is to reduce, understand and visualize the structure of complex data sets. A gene expression microarray dataset is characterized by its high dimensionality with low numbers of observations. One reason for this is because microarray experiments are too expensive to produce many replications. With this kind of data set, analysis and visualization is difficult in practice and becomes an obstacle for the clinicians and biologists in the field of diagnosis and treatment of patients such as childhood leukaemia sufferers [4].

Feature selection and dimensionality reduction are two different technologies that will be involved in this study. Feature selection is an approach to selecting the most relevant features in a data set. The features might be selected based on a target class (ranking method) or based on a specific classifier (wrapper method). Dimensionality reduction is a way to transform a high-dimensional data set into a lower dimensional data set which represents the most important variables that underlying the original one. Many algorithms have been published for feature selection and dimensionality reduction. Some of these algorithms are related to feature selection algorithms such as f-statistic, t-statistic and mRMR [5]. Others are related to Linear Dimensional Reduction (LDR) like PCA and Non-Linear Dimensional Reduction (NLDR) such as MDS, ISOMAP [1] and Local Linear Embedding (LLE) [2]. These different techniques and algorithms might not be effective if used alone for a very high dimensional data set like microarray data set (thousands of features). For example, feature selection is an important tool to apply on microarray data in order to select the most important and relevant features, but it is not enough on its own to reduce a high dimensional data. The same is true for principal component analysis (PCA), which is considered as a linear method and very simple effective tool but it is also not efficient for high dimensional and complex data set. This is due to the fact that PCA can‛t retrieve precisely the true latent variables of complex and non-linear data sets [6]. With respect to the NLDR, it is known that these algorithms have been developed for high dimensionality data reduction and visualization, yet are not enough on their own for very high dimensional data sets [6]. For example, LLE is very efficient and powerful for dimensionality reduction among the other algorithms [6, 7, 8]. However, these types of algorithms are very heavy on the machine in terms of time consumption and memory. LLE is a complex calculation and memory consumption. For example, the three LLE steps require a complex computation which will be as $O(dn^2)$, $O(dnk^3)$ and $O(rn^2)$ respectively where d is the input data dimensionality, k the number of nearest neighbors, n the number of data points and r the output dimensionality [9].

Moreover, several papers have been published related to dimensionality reduction, yet few papers have been found that combine different technologies such as feature selection, LDR and NLDR. Bowman used Principal Component Analysis (PCA) alone for dimensionality reduction for bio-medical spectra [3], On the other hand, Quansheng [7] uses Local

Linear Embedding (LLE) for dimensionality reduction without using any feature selection algorithm in order to clean the data and remove noise which might affect the quality of LLE [7].

The goal of this study concerns a framework for dimensionality reduction. This framework is composed of a sequence of procedures that involve dimensionality reduction techniques for a Leukaemia microarray data set. The purpose of these procedures is to achieve the maximum reduction of attributes with the minimum noise and redundancy without losing any relevant and significant information from the original data. The framework applies feature selection followed by linear and nonlinear dimensionality reduction. The first procedure will be feature selection. The goal of this step is to remove the non-useful features and noise and ensures that only beneficial information and significant features will be delivered to the next step. LDR will be the next step because the dimensionality of the data is very high. Accordingly, LDR may be very useful to reduce and suppress a large number of useless features. The idea of LDR represented by PCA is to find the direction of maximum variance in the input space. Then, data is transformed into a linear combination of the original attributes. The last procedure is to apply non-linear dimensionality reduction in order to keep the most interesting variables which will represent the final data and provide a better understanding of the structure of the data.

The rest of this paper is organized as follows. Section II introduces the dimensionality reduction with the linear and non-linear dimension reduction. We will discuss feature selection and the used techniques in Section III. The structure of the framework which focuses on linking and uniting feature selection, LDR and NLDR will be discussed in Section IV with the used methods. Section V demonstrates the experimental results with the similarity measurement. In Section VI, we draw conclusions about the results and present some of the future work.

II. Dimensionality Reduction

Dimensionality reduction is an important tool in the proposed framework. It provides a way to reduce, understand and visualize the structure of complex data sets with very high dimensional data. There are two types of dimension reduction, Linear Dimensionality Reduction like Principal Component Analysis and Non-Linear Dimensionality Reduction such as LLE, ISOMAP and KPCA.

A. Linear Dimension Reduction

Principal Component Analysis (PCA) is one of the oldest and best known methods in the field of data analysis. It was introduced by Pearson [10]. PCA is characterized by its simplicity and it is a non-parametric method. It is used to extract the latent variables from a high dimensional data set. The effectiveness of this method is limited by its global linearity and simple matrix multiplication (covariance matrix). Nevertheless, it still provides a roadmap for a very high dimensional data set (dimension greater than 50) [6]. Consequently, PCA will act as a first step for dimensionality reduction to reduce the data set into lower number of dimensions in order to reveal the latent variables and simplified structure that often underlie it [6].

The idea behind PCA is to find the directions or components where the data has maximum variance. This is achieved by finding the eigenvalues with the corresponding eigenvectors for the covariance matrix of the data sets.

B. Non-Linear Dimension Reduction

Non-Linear Dimensionality Reduction methods are often more powerful than linear ones, because the connection between the latent variables and observed ones may be much richer than simple matrix multiplication [6]. The purpose of NLDR is to map high dimensional data to a lower dimensional space. Lee and Verleysen [6] define a taxonomy for NLDR. One method reduces the dimensionality of data by using distance preservation technique. Distance preserving is achieved by computing spatial distance (Euclidean distance) like MDS or by measuring the geodesic distance (graph distance) such as ISOMAP. Other methods reduce the dimensionality by preserving the topology of the data. Two types of topology preservation are identified, the predefined lattice (SOM) and model-derived lattice (LLE).

Some NLDR algorithms share the same basic approach, consisting of following three steps;

- Determining the neighbourhoods points in the input space;
- Constructing a square matrix with as many rows as elements in the input data set;
- Computing spectral embedding using the eigenvectors of this matrix.

III. Feature Selection

As previously said microarray data is very high dimensional data set and has abundant number of features. However, only a small number of these features are relevant for disease or genotype [11,12]. It can be implied that microarray has a large number of redundant features and noise which may affect the results of dimensional reduction algorithms especially the NLDR which is a highly sensitive to noise [7] and the results might be exposed to the loss of accuracy and quality.

Based on that, feature selection will be applied to the data at the first stage of the framework in order to remove the redundant and irrelevant features from the data set. Many potential benefits can be achieved by feature selection such as facilitating data visualization and data understanding, reducing training and utilization times, improving dimensionality reduction and similarity measurements [13].

There are two general approaches for feature selection. The first one is concerning feature selection regardless of classifier. This method is known as feature ranking or filtering [14]. The other approach participates in prediction and how to build a good classifier without caring about the features in themselves. This method is known as wrapper selection which aims to select subsets of features that are useful to build a good predictor [15].

IV. Feature Selection, Linear Dimensional Reduction and Non-Linear Dimensional Reduction Framework

Feature Selection, Linear Dimensional Reduction and Non-Linear Dimensional Reduction Framework is presented in this paper for high dimensional data reduction, time computations and better visualization. This framework is composed of three steps as illustrated in Fig 1. As previously said, the initial step in this framework is feature selection. Maximum Relevance Minimum Redundancy feature selection based mutual information has been used to rank the features and then select the top ranking ones which represent the most significant and correlated features. The next step is to apply LKPCA in order to find the maximum variance of the data and select the principal components with help of factor analysis technique. A non-linear dimension reduction algorithm LLE is finally used on the obtained data from the output of LKPCA. This algorithm is chosen because it is powerful in dimensional reduction and visualization. These steps will be presented and explained in the following paragraphs.

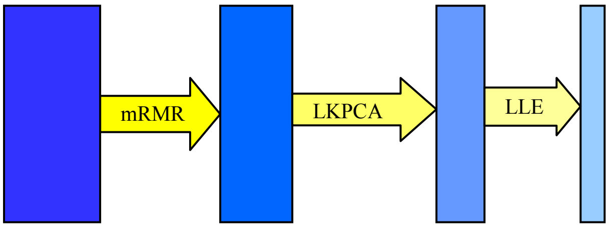

Fig. 1. Structure of the framework

A. Step 1: Feature Selection

Several considerations should be taken in feature selection. One of these is the correlation between features. Variables that are completely useless by themselves can provide significant performance and improvement when taken with others. Moreover, two variables that are useless by themselves can be useful together. Another consideration is the future feature awareness. Features might be useless in a given data set at a specific time and become useful with the addition of some extra points in the future.

A recent development of feature ranking is the Minimum Redundancy Maximum Relevance (MRMR) feature selection [5]. This algorithm will be considered for our data in order to achieve better feature selection that will act as a first step in the proposed framework. The goal of this algorithm is to minimize the redundancy and maximize the relevance of features in the subset. The selected algorithm is based on the mutual information between variables. This can be calculated based on their joint probability distribution *p(x,y)* and the respective marginal probabilities *p(x)* and *p(y)*:

$$I(x,y)=\sum_{i,j} p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)}. \qquad (1)$$

After the mutual information between the variables is calculated, the next step is to find the minimum redundancy and maximum relevance.

$$\text{Min W}_\text{I},\quad W_I=\frac{1}{|S|^2}\sum_{i,j\in S} I(i,j) \qquad (2)$$

where *S* is a set of features.

$$\text{Max V}_\text{I},\quad V_I=\frac{1}{|S|}\sum_{i\in S} I(h,i) \qquad (3)$$

where *h* is a target class.
Finally, the algorithm produces a ranked list of features. The top features represent the most significant and relevant features. Moreover, the combination of the top list also represents the best result for the target class than the combination of the bottom list where the algorithm look for the correlation between features which is required in our data.

B. Step 2: Linear Kernel Principal Component Analysis

In this Framework, we have applied a Linear Kernel Principal Component Analysis (LKPCA) instead of PCA. The goal of using the kernel method is to reduce the cost and time of the computations. In normal PCA, the size of covariance matrix (Y'*Y) is *d*d* where *d* is the number of variables. However, in LKPCA, the size of the covariance matrix (Y*Y') is n*n where n is the number of observation or points. Consequently, LKPCA is preferable when the dimensionality is high and the number of point is low.

Factor Analysis is a method can be used to precisely select the components from the obtained eigenvectors [20]. It aims to accumulate the eigenvalues of the kernel matrix until the incremented values become constant at one. This means that the rest of the eigenvalues have a zero value and can be ignored. The first step of factor analysis is to normalize the output values between zero and one. Then we start to accumulate the values till a predefined value if you do not need to take the whole information from data. For example, if we have 10 variables of data, and you found the first two factors represents 90% of the data, then the remaining eight factors represent just 10% of data and may be discarded.

C. Step 3: Local Linear Embedding

Local Linear Embedding (LLE) [2] is considered as one of most effective algorithms for non-linear dimensionality reduction. It has been used to solve various problems in information processing, pattern recognition, and data mining [16,17,18]. As LLE with topology preservation is more powerful than other methods which use distance preservation [6], this paper proposes the use of LLE for non-linear dimensionality reduction with the addition of factor analysis to the original algorithm in order to select the most significant low dimensional embedding vector space. The idea of LLE is to compute the K-nearest neighbors and then find the necessary weights in order to reconstruct each point using a linear combination of its neighbours. Finally, a low dimensional embedding is found which minimizes the loss of construction. Three major steps are involved in LLE algorithm: find the neighbours in X-space, determine the reconstruction weights which allow each point to be reconstructed from its neighbors and calculate the embedding coordinates using the reconstruction weights.

This algorithm takes an input X ($p \times n$ matrix where p is the number of attributes and n is the number of points) and outputs Y ($d \times n$ matrix) where $d < p$ is the dimensionality of the embedding input vector X in the low dimensional space (Y). The first step is to compute the neighbors for each data point. For that, we determine the K-nearest neighbors for each data point. The quality of dimensionality reduction is highly sensitive to the value of parameter (K) which should be carefully chosen; otherwise the result will be exposed to loss of quality. If this parameter is tuned with a very high value, the algorithm will loose its nonlinear character and act as a linear dimensional reduction. On the other hand, if the value is too small, the data points will be above each other and the mapping will not reflect any global properties [19].

The second step is to determine the reconstruction weights. This task is done firstly by constructing the weight for X_i only from its K nearest neighbours and set zero weights for the points which not neighbours X_i. Secondly, we enforce the sum of local weights to be equal to one.

The final step is to calculate the embedding coordinates Y using the construction weights and find the spectral embedding vector using the eigenvectors of this matrix.

V. Experiments

A Leukaemia dataset has been used to demonstrate this framework. The data is composed of 72 observations with 255 features. The 72 observations are divided into two clusters which separated between the diseased (-1) and healthy (1). After applying feature selection on our dataset, 198 features out of 255 have been selected from the data set which represent the most relevant and correlated variables. The images below show the result of the obtained data set from feature selection after applying LKPCA and PCA algorithm (Fig. 2 and 3).

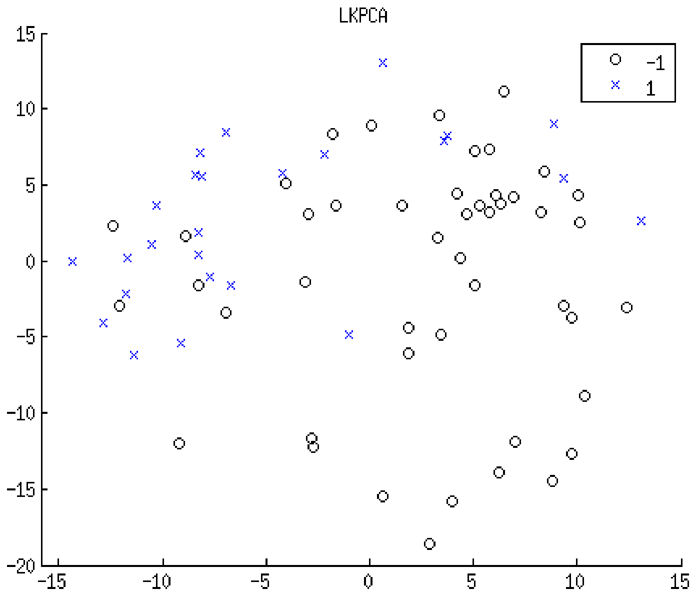

Fig. 2. Output of LKPCA algorithm

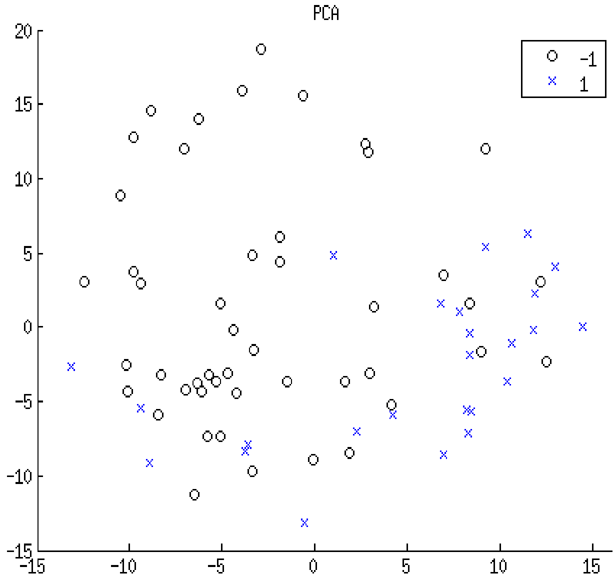

Fig. 3. Output of PCA algorithm

As can be seen from the figures (2 and 3), the outputs of the two algorithms are similar if we make some rotation for the image. However the computation time of the LKPCA is less than PCA because the matrix size of Y'*Y is less than Y*Y' in our data. The selected features from the LKPCA are 71 features. This number is selected based on the factor component analysis.

The next step in our framework is to apply the nonlinear dimension reduction to the dataset obtained from LKPCA. The data set size now is 72*71 which means that the nonlinear dimension reduction will have less time for computations. Like LKPCA, we also applied the factor component analysis on the LLE algorithm. Forty nine features have been selected from the 71 features. These 49 features represent 93% of the data which are obtained after applying factor component analysis. The results show that the clusters obtained after LLE are more separated than LKPCA with some interference between the two clusters because we visualize data in just two dimensions. The image below shows the output result of the LLE algorithm (Fig. 4).

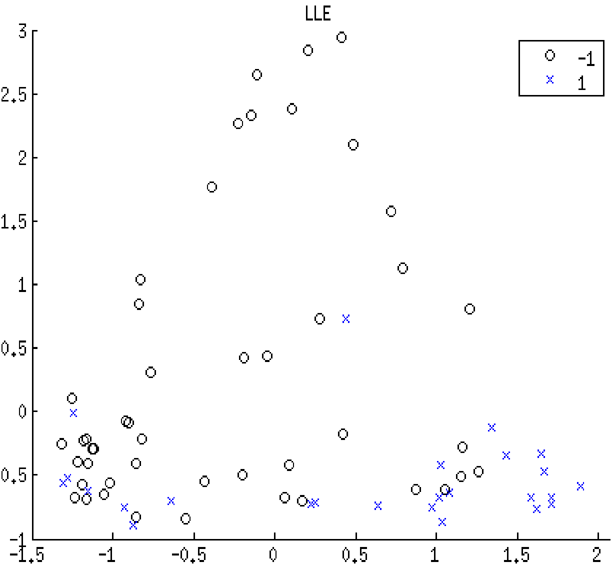

Fig. 4. Output of LLE algorithm

VI. Conclusion

In this paper we have proposed a framework for high-dimensional data reduction based on mRMR, LKPCA and LLE. We have discussed the usability of the features selection and the usage of LKPCA instead of PCA with respect the time calculation saving as the LKPCA and PCA produce the same output result. Moreover, we have shown that LLE provides a good dimensionality reduction in less time computation with the help of feature selection and LKPCA. In this framework, LLE operates on 72*49 instead of 72*255 (without doing feature selection) or 72 * 149 (without doing linear dimensionality reduction). This framework provides a way to visualize the data in order to see the position of a patient with respect to other patients. Our future work will include linking and joining this framework into a Case-Based Reasoning system and will apply it on a clinical Leukaemia data set. Developing an unsupervised feature selection algorithm will be also part of my future work in order to handle the new incoming case for the Case-Based Reasoning system

References

[1] Tenenbaum, V. de Silva, and J.C. Langford, A global geometric framework for nonlinear dimensionality reduction. *Science,* 290(5500):2319–2323, 2009.

[2] Roweis S.T. and Saul L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science, **290(5500):2323–2326, 2000**.*

[3] Bowman, C., R. Baumgartner et al, Dimensionality reduction for bio-medical spectra. Electrical and Computer Engineering, 2002. *IEEE* CCECE, 2002.

[4] Kennedy, P. J., Simoff, S. J., Skillicorn, D. and Catchpoole, D. *Extracting and Explaining Biological Knowledge in Microarray Data*. Proc. Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sydney. (eds) Dai, H., Srikant, R., and Zhang, C., LNAI 3056, pp 699-703, Springer-Verlag Berlin, 2004

[5] Chris Ding, and Hanchuan Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, Vol. 3, No. 2, pp.185-205, 2005

[6] Lee.J and Verleysen.M, *Nonlinear Dimensionality Reduction*, Springer, 2007.

[7] Quansheng, J., J. Minping, et al., New approach of intelligent fault diagnosis based on LLE algorithm. *Control and Decision Conference*, 2008. CCDC 2008. Chinese, 2008.

[8] Varini, C., T. W. Nattkemper, et al., Breast MRI data analysis by LLE. Neural Networks, 2004. *Proceedings. 2004 IEEE International Joint Conference*, 2004.

[9] Tian, H. and D.G. Goodenough. Nonlinear feature extraction of hyperspectral data based on locally linear embedding (LLE). *In Geoscience and Remote Sensing Symposium*, 2005. IGARSS '05. Proceedings. 2005 IEEE International. 2005.

[10] Pearson, K., On lines and planes of closest fit to systems of points in space . *Philosophical Magazine*, 2:559-572, 1901.

[11] Li W, Yang Y, How many genes are needed for a discriminate microarray data analysis?, *in Critical Assessment of Techniques for Microarray Data Mining Workshop*,pp. 137–150, 2000.

[12] Xiong M, Fang Z, Zhao J, Biomarker identification by feature wrappers, *Genome Res*11:1878–1887, 2001.

[13] Guyon.I and Elisseeff..A, An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 3 (2003) 1157-1182, 2002.

[14] Langley P, Selection of relevant features in machine learning, *in AAAI Fall Symposium on Relevance*,1994.

[15] Kohavi R, John G, Wrapper for feature subset selection, *Artificial Intelligence*, 1997.

[16] Zhang, C., Wang, J., Zhao, N., & Zhang, D., and analysis of multi-pose face images based on nonlinear dimensionality reduction. *Pattern Recognition*, 37(2), 325–336, 2004.

[17] Elgammal, A. M., & Lee, C. S., Separating style and content on a nonlinear manifold. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp 478–485, 2004.

[18] Mekuz, N., Bauckhage, C., & Tsotsos, J. K., Face recognition with weighted locally linear embedding, *In The second Canadian conference on computer and robot vision* pp. 290–296, 2005.

[19] De Ridder.D and Robert,.D, Locally linear embedding for classification. *In the Pattern Recognition Group Technical Report Series. ICIP*. 2005.

[20] Wang, F., K. Rob, et al., Factor Analysis and Principal-Components Analysis. *International Encyclopedia of Human Geography*. Oxford, Elsevier: 1-7, 2009.