AN INCREASED PERFORMANCE OF CLUSTERING HIGH DIMENSIONAL DATA THROUGH DIMENSIONALITY REDUCTION TECHNIQUE

[1]P. VALARMATHIE, [2]DR MV SRINATH, [3]K. DINAKARAN

[1]Dept. of Computer Science and Engineering, Dr. MGR University, Chennai, India

[2]Dept. of Computer Science and Engineering, Mahendra Engineering College, Namakkal, India.

[3]Dept. of Computer Science and Engineering, RMK Engineering College, Chennai, India.

E-mail: valarmathi_p@rediffmail. com, dina_72@yahoo.com

ABSTRACT

With the incredible growth of high dimensional data such as microarray gene expression data, the researchers are forced to develop some new techniques rather than using existing techniques to meet their requirements. Microarray is a mechanism of measure the expression level of tens of thousands of genes simultaneously as a result, the data generated is very large. There must be an efficient technique to handle this huge amount of data thereby the researchers can able to do analyze and interpret them. The accuracy of the resultant value perhaps not up to the level of expectation when the dimensions of the dataset is high because we cannot say that the dataset chosen are free from noisy and flawless. So it is required to reduce the dimensionality of the given dataset in order to improve the efficiency and accuracy. Moreover the running time of an algorithm certainly has to be minimized to achieve the desired results. This is being done by apply the same data set to a same clustering technique with and without performed the dimensionality reduction technique principal component analysis on original data. The results of the two approaches are compared and it is proved that the results of clustering using PCA are more accurate, easy to understand and above all the time taken to process the data was substantially reduced.

Keywords: *Microarray, Principal Component Analysis, Dimensionality Reduction, Running Time.*

I. INTRODUCTION

The researchers are now a day is inundated with large data after the advent of microarray technique which forced them to devise some new techniques to manipulate those data effectively. Microarray is a mechanism of measure the expression level of tens of thousands of genes simultaneously as a result, the data generated is very large [1]. In this context clustering is inevitable exploratory technique for analysis of gene expression data. Even though, diverse of techniques were introduced and applied by the researchers to solve many biological problems, still there are some limitations inherent with them. Some popular and widely used data mining clustering techniques such as hierarchical and k-means clustering techniques are statistical techniques and can be applied on microarray gene expression data [2]. The problem associated with the above said techniques are, the results are indistinct if the data set was large and the number of clusters defined may be uncertain for hierarchical and k-means clustering techniques respectively.

In general, handling high dimensional data using clustering techniques obviously a difficult task in terms of higher number of variables involved. In order to improve the efficiency the noisy and outlier data may be removed and minimize the execution time, we have to reduce the no. of variables in the original data set. To do so, we can choose dimensionality reduction methods such as principal component analysis (PCA), Singular value decomposition (SVD), and factor analysis (FA). Among this, PCA is preferred to our analysis and the results of PCA are applied to a popular model based clustering technique [3][4].

PCA is a classical technique; the central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables [5]. This is achieved by transforming to a new set of variables (Principal Components) which are uncorrelated and, which are ordered so that the first few retain the most of the variants present in all of the original variables [3]. It is a statistical technique for determining key variables in a

high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set. The main advantage of PCA is that once we have found these patterns in the data and we can compress the data i.e., by reducing the number of dimensions without much loss of information.

PC's (Principal Components) can be defined in two ways they are correlation and covariance matrices. Here, we preferred covariance matrix wherein the variances of variables are very high compared to correlation with respect to microarray gene expression data in different time points. It would be better to choose the type correlation when the variables are of different types [6].

Numerous clustering algorithms have been developed to deal with different data sets but none of them is suitable for all kind of data and applications [7]. A clustering method pertinent to probability model and become very popular in very short period for clustering data is model based clustering [2]. It has become an essential one in microarray gene expression data in order to determine the number of clusters and provides a statistical framework to model the cluster structure of gene expression data. A proposed approach uses this method to determine the exact number of clusters to the given microarray gene expression data and offer better interpretability since the resulting model for each cluster directly characterizes that cluster [8].

II. PROPOSED METHOD

In traditional data analysis techniques such as k-means, hierarchical etc, that were developed for low dimensional data often do not work well for high dimensional data like microarray gene expression data and the results may not accurate most of the time due to noisy and outliers associated with original data. Also, for some data analysis algorithms, the computational complexity increases rapidly as the dimensionality increases.

In order to improve the efficiency, we proposed a new approach that PCA is applied on original data set. The resultant uncorrelated data is being clustered by using a popular method model based technique which determine the precise number of clusters and widely used on multidimensional data in order to cluster the data and to facilitate the researchers to simplify the analysis and visualization of multidimensional data. In this method, PCA is applied before the given data set is to be clustered and the resultant value is being compared with the results of original data applied on the same clustering technique. Moreover the running time of an algorithm is calculated and compared for both the approaches i.e., after PCA and for the original data. In order to enhance the performance of an algorithm, it is mandatory to minimize the processing time.

III. RESULTS AND DISCUSSIONS

In this proposed method, we discussed about dimensionality reduction of a dataset before data analysis. The sample dataset taken for this analysis is publically accessible web site [9]. The data set has 245 genes and 11 attributes which represents the expression of genes in 11 different series of time points which is shown in the fig. 1. The Eigen values and the percentage of variances for corresponding principal components are shown in the table 1. Among them the first three PC's having maximum variances therefore the first three PC's were taken for our analysis.

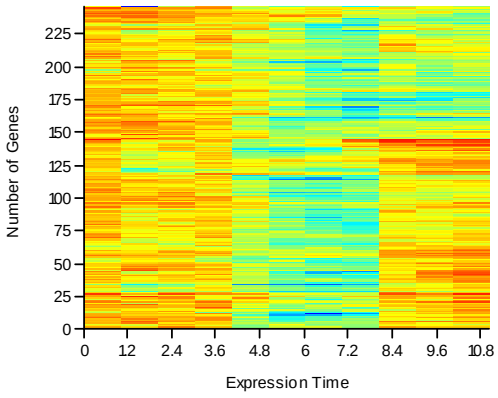

Figure 1. Shows how the genes are expressed in different time points

Choosing the number of PC's also an important issue i.e., the number of PC's used in data analysis. This can be done using scree graph as shown in Fig. 2 which illustrates the no. of components to be chosen. The plot is actually drawn between the Eigen values and components. The no. of components is chosen from the figure is to be the point at which the line is steep on the left. Thus, the no. of components is three according to scree graph.

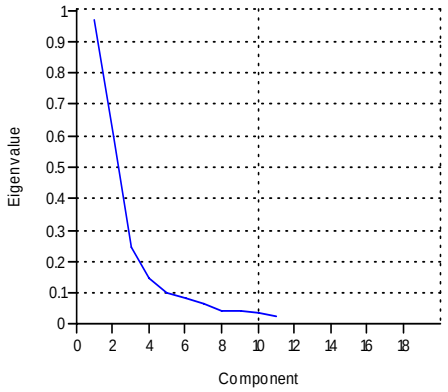

Figure 2. The scree plot illustrates the no. of valid components that can be taken for analysis. The no. of components chosen in the figure is to be the point at sswhich the line is steep on left.

The main objective of applying PCA on original data before clustering is to obtain accurate results so that the researchers can do analysis in better way. Secondly, minimize the running time of a system because time taken to process the data is a significant one. Normally it takes more time when the number of attributes of a data set is large and sometimes this dataset not supported by all clustering techniques hence the number of attributes are directly proportional to processing time.
This experiment shows a substantial improvement in running time by reducing the dimensions which is given in table 2. Log likelihood value 4.13 clearly shows the accuracy of the results while applying PCA. According to the value of log likelihood the correct number of clusters is 3. So, the results of the scree plot and the outcome of the clustering technique used in this article are same.

Table 1. shows the first three PCs are high variance of Eigen value

PC	Eigen value	% Variance
1	0.964663	41.356
2	0.595385	25.525
3	0.244404	10.478
4	0.145674	6.2452
5	0.0999173	4.2836
6	0.0791146	3.3917
7	0.064267	2.7552
8	0.0418039	1.7922
9	0.0383639	1.6447
10	0.0343383	1.4721
11	0.0246295	1.0559

Table 2. Comparison of results between original data and PCA transformed data which clearly shows the time taken to cluster the data are directly proportional to number of variables of an instance.

	No. of Clusters	Running Time (in seconds)	Log Likelihood
Original Data	5	58	-4.14
PCA Applied	3	36	4.13

REFERENCES:

[1]. Michel B Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proc, Natl. Acad. Sci. USA*, Vol. 95, pp, 14863-14868, December 1998.

[2]. RM Suresh, K Dinakaran, P Valarmathie, "Model based modified k-means clustering for microarray data", *International Conference on Information Management and Engineering*, Vol.13, pp 271-273, 2009, IEEE

[3]. Pang-Ning Tang, Michal Steinbach and Vipin Kumar, " *Introduction to Data Mining*", Pearson Education, Third edition, 2009.

[4]. I.T Jolliffe, "*Principal Component Analysis*", Springer, second edition.

[5]. Chris Ding and Xiaofeng He, "K-Means Clustering via Principal Component Analysis", In proceedings of the 21st *International Conference on Machine Learning*, Banff, Canada, 2004

[6]. Ka Yee Yeung, Walter L. Ruzzo, " An empirical study on principal component analysis for clustering gene expression data", *University of Washington*, 2000.

[7]. Anil K. Jain, Richard C. Dubes, "*Algorithms for Clustering Data*", Prentice Hall, 1988.

[8]. Shi Zhong, Joydeep Ghosh, "A unified framework for model based clustering", *Journal of Machine Learning Research* (2003) 1001-1037

[9]. http://genomics-stanford.edu/serum