Effective Dimensionality Reduction Based Clustering in Gene Data

Mrs. B. Dharanya, Mrs. S. Maheshwari
Assistant Professor
Department of Computer Science
Dr. NGP Arts and Science College
Coimbatore-48

ABSTRACT

Clustering high dimensional data presents a major challenge. Some of the challenges are many irrelevant dimensions may mask clusters. Distance measure becomes meaningless due to equi-distance and some clusters may exist only in some subspaces.

It may be difficult in modeling the precise relationship between many number of feature variables and object variables. So to improve the accuracy and efficiency of the clustering, effective clustering approaches are taken and clustering can be done for the dimensional data. Micro array data now a day provides a new opportunities and challenges for data mining. So Micro array data is taken as the high dimensional data and the dimension is been reduced by using techniques of dimension reduction.

Density based clustering for gene expression data is been made and Dimensionality reduction based clustering is also been made for gene data. In the Noise removal Dimensionality reduction clustering outperforms Density Based clustering

In clustering gene expression data it contains noisy data, irrelevant data, missing data proper preprocessing is made by using k nearest neighbor and clustering by using K-MEANS produced an effective clustering.

Independent component analysis (ICA) is been taken to reduce the dimension by taking only the informative genes.

ICA to reduce the dimensions may seriously gain the quality and reliability of clustering results. ICA+KMEANS clustering give the best accuracy when compared DBSCAN which is the density based Clustering.

Keywords: Density based clustering, Dimensionality Reduction, Density Based Spatial clustering of application with noise, Independent component analysis, k-means clustering.

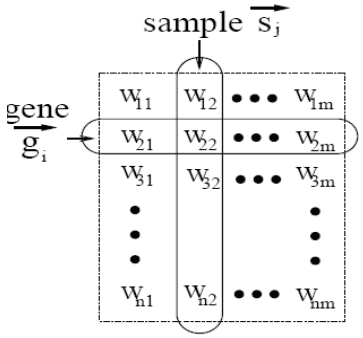
1. INTRODUCTION

In DNA Microarray technology, gene expression data can reveal many meaningful biological processes, for example, gene response to drug treatments, cancer diagnosis, etc. In gene expression data analysis, the data are arranged in a matrix form, where rows correspond to genes and the columns correspond to the genes' responses under different experimental conditions. Hence, one can examine the expression profiles of different genes by comparing rows in the expression matrix, or study the responses of genes to different experimental conditions by examining the columns of the expression matrix.

1.1 High dimensional Data

Gene Expression data is taken as the high dimensional data. A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags [ESTs]) under multiple conditions. These conditions may be a time series during a biological process (e.g., the yeast cell cycle) or a collection of different tissue samples (e.g., normal versus cancerous tissues)[4].

A gene expression data set from a microarray experiment can be represented by a real-valued expression matrix M= {Wij | 1≤i≤n, 1≤j≤m} (Figure1), where the rows *(G= {g1…gn})* form the expression patterns of genes, the columns*(S= {S1…Sm})* represent the expression profiles of samples, and each cell Wij is the measured expression level of gene *i* in sample j. Below the figure 1, some notations description is given.



(a)

Figure 1 Gene Expression matrix

n	number of genes
m	number of samples
M	a gene expression matrix
w_{ij}	each cell in a gene expression matrix
$\vec{g}_i$	a gene
$\vec{s}_j$	a sample
$G, G', G_0, \ldots$	a set of genes
$S, S', S_0, \ldots$	a set of samples

Notation 1

2. CLUSTER ANALYSIS

Cluster analysis is a powerful tool in the study of gene expression data. Clustering is the process of grouping data objects into a set of disjoint classes, called clusters, so that objects within a class have high similarity to each other, while objects in separate classes are more dissimilar.

Clustering algorithms are attractive for the task of class identification in spatial databases. However, the application to large spatial databases rises the following requirements for clustering algorithms: minimal requirements of domain knowledge to determine the input parameters, discovery of clusters with arbitrary shape and good efficiency on large databases.

The well-known clustering algorithms offer no solution to the combination of these above requirements. In our proposed work, the new clustering algorithm DBSCAN relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. DBSCAN requires only the input parameters and supports the user in determining an appropriate value for it while performed an experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data and the real data[6].

The results of proposed work demonstrate that ICA+KMEANS clustering discovers the clusters with noise removed more efficiently than DBSCAN which discover the clusters with detecting less noise.

3. DENSITY BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE

DBSCAN is a "**Density Based Spatial Clustering of Applications with Noise".** It is fundamentally a density based clustering. It is used to create clusters with a minimum size and density. Density is defined as a minimum number of points within a certain distance of each other. It handles the outlier problem by ensuring that an outlier will not create a cluster. One input parameter, **Minpts**, indicates the minimum number of points in any cluster. In addition, for each point in a cluster there must be another point in the cluster whose distance from it is less than a threshold input value, **Eps.** The neighborhood of a point is the set of points within a distance of Eps. It just imports the kdtree which is mainly used for calculating the distance between two points. It initializes the kdtree for the total number of points and it just inserts the data points into it by using kdtree it finds out the nearest neighbors of data points[5].

DBSCAN uses a new concept of density. Some definition it defines that is directly density-reachable. The first part of the definition ensures that the second point is "close enough "to the first point. The second portion of the definition ensures that there is enough **core points** close enough to each other [12]. These core points form the main portion of a cluster in that the points are all close to each other. A directly density-reachable point must be close to one of these core points, but it need not be a core point itself. In that case, it is called a **border point.** Refer Figure3.Let the distance between two sets of points Sl and S2 be defined as

dist (Sl, S2) = min {dist(p,q) I p E Sl, q E S2}

A point is said to be density-reachable from another point if there is a chain from one to the other that contains only points that are directly density-reachable from the previous point. It is possible that a border point could belong to two clusters. The stated algorithm will place this point in whichever cluster is generated first. It find out all the clusters and noise from the test dataset with less effectively.

Figure2. All Points

Figure 3. Border Points

4. DIMENSIONALITY REDUCTION BASED CLUSTERING

Dimensionality reduction technique is been used for selecting the informative genes and the following have to be considered:

• Scientific: understand structure of data (visualization)

• Statistical: fewer dimensions allows better generalization
• Computational: compress data for efficiency (both time/space)
• Direct: use as a model for anomaly detection.

One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are important" for understanding the underlying phenomena of interest. While certain computationally expensive novel methods [3] can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modeling of the data.

4.1 Goal Of Dimensionality Reduction

- It is so easy and convenient to collect data
- Data is not collected only for data mining
- Data accumulates in an unprecedented speed
- Data preprocessing is an important part for *effective* machine learning and data mining
- Dimensionality reduction is an effective approach to downsizing data

4.2. Independent Component Analysis for high dimensional data

ICA (Independent component analysis): find subspace where sources are independent. This Paper introduces a method utilizing Independent Component Analysis (ICA) For feature selection and informative genes identification for microarray sample clustering.

ICA aims to find a transformation that decomposes an input dataset into components so that each component is statistically as independent from the others as possible. ICA has advantage over DBSCAN because ICA exploits higher order statistics and has no restriction on its transformation, whereas DBSCAN deals only with the density. It typically based on dense region and data space which are separated by regions of low density. So each component is statistically as independent from the others.
In this paper, ICA is taken as the dimensionality reduction technique and analyze the quality of informative genes by conducting ICA-based investigation.

Independent component analysis (ICA) [2, 1] is a relatively new statistical and computational technique that recovers a set of linearly mixed hidden independent factors from a set of measurements or observed data. A typical ICA model assumes that the source signals are not observable, statistically independent and non-Gaussian, with an unknown, but linear, mixing process. The latent variables are assumed non-Gaussian and mutually independent and they are called the independent components of the observed data. These independent components (ICs), also called sources or factors, can be found by ICA.

K-means clustering algorithm is used for clustering the gene data using informative genes that have been selected. K-Means is a partitioned based algorithm that takes number of input clusters as input parameters. First the algorithm randomly selects the objects. Each object represents a cluster. This object represents the mean of the cluster.

Secondly it assigns each object to exactly one cluster based on a distance measure. Each object is assigned to those cluster that is the nearest to the given object. The distance is checked for each object and the object is placed into another cluster than to the cluster it is already.

5. RESULT ANALYSIS

Table 1. Original data size and Informative gene space size by using ICA

Microarray data	Original data size	Informative gene space size
Leukemia	7129*38	62*38
Breast cancer	2000*62	80*62

Figure 4. Clustering Performance Of Density And Dimensionality Reduction Based Clustering

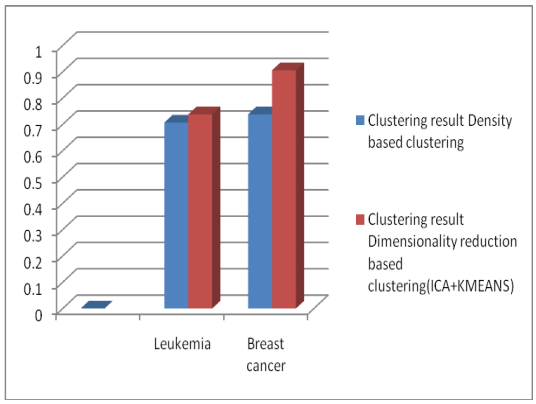


Figure 4 shows the clustering result for dbscan and ICA+kmeans clustering

Results have been obtained on other two datasets: Leukemia Dataset and Breast Cancer Dataset. Generally, K-means has gained the largest performance improvement on the ICA-based informative gene space. We have studied various informative gene selection methods and various unsupervised algorithms and analyzed that ICA based informative gene selection produces better quality cluster.

6. CONCLUSION

In this paper, we studied about density and dimensionality reduction based clustering methods for high dimensional data and we analyzed that by reducing the dimension and clustering, produced the best clustering. The output of the system [11] produced the quality of clustering done with ICA-based investigation is better than DBSCAN clustering. So ICA outperforms better than DBSCAN.

Various clustering algorithms have achieved higher performance based on the new and reduced informative gene space and improving the quality of informative genes by conducting ICA-based investigation more thoroughly. They have applied K-means clustering algorithm on the informative gene space to verify its effectiveness. In future work, we can apply other dimensionality reduction algorithms and other clustering algorithm.

REFERENCES:

1. A. Hyv¨arinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*,13(4-5):411–430, 2000.

2. A. Hyv¨arinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

3. Alter O., Brown P.O. and Bostein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, Vol. 97(18):10101–10106, Auguest 2000.

4. Daxin Jiang and Chun Tang Cluster analysis of gene expression data: A survey

5. Ester.M, Kriegel H.P , Sander. J, and X. Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD96), 1996.

6. Garcia J.A., Fdez-Valdivia J., Cortijo F. J., and Molina R. 1994. "A Dynamic Approach for Clustering Data. Signal Processing", Vol. 44, No. 2,1994, pp. 181-196.

7. Getz G., Levine E. and Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA*, Vol. 97(22):12079–12084, October 2000.

8. Golub T.R., Slonim D.K., Tamayo P., Huard C., Gassenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield D.D., and Lander E.S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, Vol. 286(15):531–537, October 1999.

9. Halkidi, M., Batistakis, Y. and Vazirgiannis, M. On Clustering Validation Techniques. *Intelligent Information Systems Journal*, 2001.

10. Lei Zhu, Chun Tang, An ICA-based Feature Selection Method for Microarray Sample Clustering, 2006.

11. Thomas J.G., Olson J.M., Tapscott S.J. and Zhao L.P. An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles. *Genome Research*, 11(7):1227–1236, 2001.

12. Zhang.T, Ramakrishnan.R, and M. Livny, "BIRCH: an efficient dataclustering method for very large databases," pp. 103-114, 1996