A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets

Tajunisha N[1], Saravanan V[2]
[1] *Associate Professor in Computer Science,*
Sri Ramakrishna college of arts and Science for women,Coimbatore, India.
tajkani@gmail.com
[2] *Director in Dept. of Computer Application,*
Dr.N.G.P Institute of Technology,Coimbatore, India.
tvsaran@hotmail.com

Abstract

DNA microarray technology can be used to measure expression levels for thousands of genes in a single experiment across different samples. Within a gene expression matrix there are usually several particular Macroscopic Phenotypes of samples related to some diseases or drug effects such as diseased samples, normal samples or drug treated samples. The goal of sample based clustering is to find the phenotype structure or substructure of the samples. Currently most of research work focuses on the supervised analysis, relatively less attention has been paid to unsupervised approaches in sample based analysis which is important when domain knowledge is incomplete or hard to obtain. The standard k-means algorithm is effective in producing clusters for many practical applications. But the computational complexity of the original k-means algorithm is very high in high dimensional data and the accuracy of the clustering result depends on the initial centroid. In this paper, we present a new framework for unsupervised sample based clustering using informative genes for microarray data. We proposed a method to find initial centroid for k-means and we have used similarity measure to find the informative genes. The goal of our clustering approach is to perform better cluster discovery on sample with informative gene.

Keywords: k-means; informative gene; dimension reduction; initial centroid; microarray gene data

1. Introduction

Mining microarray gene expression data is an important research topic in bioinformatics with broad applications. Microarray technologies are powerful techniques for simultaneously monitoring the expression of thousands of genes under different sets of conditions. Gene expression data can be analyzed in two ways: unsupervised and supervised analysis. In supervised analysis, information about the structure/groupings of the object is assumed known or at least partially known. This prior knowledge is used in analysis process. In unsupervised analysis, prior knowledge is not known.

Clustering of gene expression data can be divided into two main categories. Gene-based clustering and sample-based clustering [3]. In gene based clustering, genes are treated as objects and samples are features or attributes for clustering. The goal of gene-based clustering is to identify differentially expressed genes and sets of genes or conditions with similar expression pattern or profiles, and to generate a list of expression measurements.

Sample based clustering can be used to reveal the phenotype structure or substructure of samples. Applying the conventional clustering methods to cluster samples using all the genes as features may degrade the quality and reliability of clustering results.

The standard k-means algorithm [9] is effective in producing clusters for many practical applications. But the computational complexity of the original k-means algorithm is very high in high dimensional data. A clustering algorithm typically considers all features of the data in an attempt to learn as much as possible about the objects. However, with high dimensional data, many features are redundant or irrelevant. The redundant features are of no help for clustering; even worse, the irrelevant features may hurt the clustering results by hiding clusters in noises. There are many approaches to address this problem. The simplest approach is dimension reduction techniques including principal component analysis (PCA) and random projection. In these methods, dimension reduction is carried out as a preprocessing step. Different methods have been proposed [2] by combining PCA with k-means for high dimensional data.

Golub et al, (1999) [4] have demonstrated that the phenotypes of samples can be discriminated by employing only a small subset of genes whose expression levels strongly correlate with the class distinctions. These genes are called informative genes. The remaining genes are irrelevant to the classification of samples of interest and thus are regarded as noise. To select informative genes, neighborhood analysis approach, supervised learning method and ranking based methods are to be included. Here we will focus on sample based clustering using k-means.

The accuracy of the k-means clusters heavily depending on the random choice of initial centroids. If the initial partitions are not chosen carefully, the computation will run the chance of converging to a local minimum rather than the global minimum solution. The initialization step is therefore very important. To combat this problem it might be a good idea to run the algorithm several times with different initializations. If the results converge to the same partition then it is likely that a global minimum has been reached. This, however, has the drawback of being very time consuming and computationally expensive.

In this paper, we propose the new approach to unsupervised sample based clustering by selecting informative genes. Here we also proposed the method to find the initial centroid for k-means algorithm.

2. K-means clustering algorithm

K-means is a prototype-based, simple partition clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice. The algorithm consist of two phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to nearest centroid. The k-means algorithm works as follows:

a) Select initial centroid of the k clusters. Repeat steps b through c until the cluster membership stabilized.

b) Generate a new partition by assigning each data to its closest cluster centroid.

c) Compute new cluster centroid for each cluster.

The most widely used convergence criteria for the k-means algorithm is minimizing the SSE.

$$SSE = \sum_{j=1}^{k} \sum_{x_i \in c_j} \left\| x_i - \mu_j \right\|^2 \qquad (1)$$

Where $\mu_j = \frac{1}{n_j} \sum_{x_i \in c_j} x_i$ denotes the mean of cluster c_j and n_j denotes the no. of instances in c_j.

The k-means algorithm always converges to a local minimum. The particular local minimum found depends on the starting cluster centroids. The k-means algorithm updates cluster centroids till local minimum is found.

3. Existing Methods

In sample clustering problems, it is common to come up against the challenges of high dimensional data due to small sample volume and high feature dimensionality. High dimensional data not only bring computational complexity, but also degrade a classifier's performance. In addition traditional clustering techniques may not be effective in detecting the sample patterns because the similarity measures used in these methods are based on the full gene space and cannot handle the heavy noise existing in the gene expression data. Therefore it is necessary to conduct feature selection on the gene dimension and identify informative genes prior to the clustering on the samples.

Two general strategies have been employed to address the problem of unsupervised clustering.

1. unsupervised gene selection
2. Interrelated clustering

The first strategy difference the gene selection and sample clustering as independent process. First the gene dimension is reduced then the conventional clustering is applied.

Linear transformation methods transform the data into some new space that has some desirable properties. Principal component analysis (PCA) [7] and Independent component analysis (ICA) [5, 6] are two linear transformation methods widely used in microarray analysis. PCA projects the data into a new space spanned by the principal components. Each successive principal component is selected to be orthogonal to the previous one, and to capture the maximum information that is not already present in the previous components. Applied to expression data, PCA finds principal components, the eigenarrays , which can be used to reduce the dimension of expression data for visualization , filtering of noise and for simplifying the subsequent computational analysis [1, 11]. Originally used in blind source separation (BSS) problems [8], ICA aims to find a transformation that decomposes an input datasets into components so that each component is statistically independent from the others as possible. ICA has advantage over PCA because ICA exploits higher order statistics and has no restriction on its transformation, whereas PCA exploits only second order statistics and is restricted to orthogonal transformation. In 2006, zhu et. Al. [15] applied ICA in gene dimension reduction and identified informative genes prior to the clustering.

The second strategy can be suggested to dynamically use the relationship between the genes and samples and iteratively combine the clustering process and gene selection process. Xing et al. [14] presented a sample-based clustering algorithm named CLIFF (Clustering via

Iterative Feature Filtering) which iteratively use sample partitions as a reference to filter genes. CLIFF first uses a two-component Gaussian model to rank all genes in terms of their discriminability and then select a set of most discriminate genes. It then applies a graph-theoretical clustering algorithm to generate initial partition for the samples and the selected genes. Tang et al. [13] proposed new framework for unsupervised analysis of gene expression data. This framework, dynamically use the relationship between the groups of the genes and samples while iteratively clustering through both gene-dimension and sample-dimension.

This paper proposed new approach for sample based clustering using informative genes selected by cosine measure. Clustering with informative gene dimension will benefit the accuracy improvement of class discovery.

4. Proposed Method

In unsupervised sample based clustering, once informative genes have been identified, then it is relatively easy to use conventional clustering algorithms to cluster samples. The standard k-means can be used for partition. But the accuracy of the clustering results heavily depends on the initial centroid and the dimension of the data.

In this paper, we have proposed a method to find the initial centroid for k-means algorithm and cosine measure is used to find the informative genes. Our algorithm is described as follows:

Algorithm 1: The proposed method

```
Steps:
  1. Find the initial centroid for k-means using Algorithm 2.
  2. Informative gene selection using Algorithm 3.
  3. Cluster the data-points by k-means with fixed initial centroid using informative genes
  4. Validating cluster results using Rand Index.
```

The raw data in many cancer gene-expression datasets can be arranged in a matrix. In this matrix, the rows and columns represent the genes and the different conditions (e.g. different patients), respectively. Then we carry out the data normalization. Since gene expression microarray experiments can generate datasets with multiple missing values. The k-nearest neighbor (KNN) algorithm is used to fill those missing values.

We can obtain the input matrix for cluster initialization. Then find the eigenvectors corresponding to the largest eigenvalues and find the initial centroid for k-means as we said in algorithm 2. In algorithm 2, the $\vec{v}_i$ is chosen as the eigenvectors corresponding to the largest eigenvalue for class partitioning. The main reason is that they can capture most of the variance in the data and provide the optimal partition.

Algorithm 2: Cluster initialization

```
Steps:
  1. Obtain the input matrix table
  2. subtract the mean
  3. calculate the covariance matrix
  4. calculate the eigenvectors and eigenvalues of the covariance matrix
  5. Choose $\vec{v}_i$ as the eigenvector corresponding to the largest eigenvalues.
```

6. Sort the ith vector column ($\vec{v}_i$) of the corresponding data column.
7. Divide it into k subsets where k is the number of clusters.
8. Find the median of each subset.
9. Use the corresponding data points of original data for each median to initialize the cluster for k-means algorithm.

Next informative genes are selected before clustering. To find the informative genes the eigenvectors ($\vec{v}_1,...\vec{v}_s$) are chosen corresponding to the largest eigenvalues. The main reason to select these eigenvectors is that the genes which are most relevant to the cancer should capture most variance in the data. Since $\vec{v}_1,...\vec{v}_s$ may reveal the most variance in the data, the genes "similar" to $\vec{v}_1,...\vec{v}_s$ should be relevant to the cancer. we use the cosine similarity measure[16] to compute the similarity between each gene profile(e.g., $\vec{g}_i$) and the eigenvectors(e.g., $\vec{v}_j$, j=1,2,………,s).

Algorithm 3: Informative gene selection

Steps:

1. Cosine measure is used to find the similarity between the eigenvector $\vec{v}_j$ and each gene $\vec{g}_i$.
2. Sort the genes based on similarity values.
3. Select top most genes as informative genes.

Cosine Measure

The Cosine measure is used to compute the similarity between each gene profile (e.g., $\vec{g}_i$) and the eigenvectors (e.g., $\vec{v}_j$, j=1,2,………,s) as

$$D_{ij} = acos\left(\frac{g_i v_j'}{\sqrt{(g_i g_i')(v_j v_j')}}\right) \quad i=1,2\ldots,n \text{ genes}, j=1,2,\ldots.s \qquad (2)$$

Seen from (2), a large value of *Dij* indicates more similarity between ith gene and the jth eigenvector. Therefore, we can rank genes based on the similarity values for each eigenvector. For jth eigenvector we can select the top l genes according to the corresponding D_{ij} value for each j=1,2,……,s. The value l can be empirically determined. Thus, for each eigenvector of $\vec{v}_1,...\vec{v}_s$, we can obtain a set of genes with largest values of the cosine measure. Top selected genes with high variance are used for clustering.

Cluster Validation

The Rand Index [12] between the ground-truth of phenotype structure P of the samples and the clustering result Q of an algorithm has been adapted to for the effectiveness evaluation. Let **a** represent the number of pairs of samples that are in the same cluster in P and in the same cluster in Q, **b** represent the number of pairs of samples that are in the same cluster in P but not in the same cluster in Q, **c** be the number of pairs of samples that are in the same cluster in Q but not in the same cluster in P, and **d** be the number of pairs of samples that are in different clusters in P and in different clusters in Q. the Rand Index is calculated as

$$RI = \frac{a+d}{a+b+c+d} \tag{3}$$

The Rand Index lies between 0 and 1. Higher values of the Rand Index indicate better performance of the algorithm. Here, in our work, Rand Index measure is used to find the accuracy of the clustering results.

5. Experimental Results

In this section, we will report performance evaluation of the proposed method on the following gene expression datasets:

We evaluated the proposed algorithm on the data sets from UCI machine learning repository [10]. We have taken iris dataset to explain our method and we have applied our method to Winconsin Diagnostic Breast Cancer (WDBC) data. It contains 569 samples and each sample is measured over 30 genes. We compared the clustering results achieved by the k-means, PCA+k-means with random initialization and the proposed algorithm.

The ground-truth of the partition, which includes such information as how many samples belong to each class and the class label for each sample, is only used to evaluate the experimental results. During the experiment, we compared the clustering results of k-means with full gene space, k-means with PCA-based informative gene space and proposed method with informative gene space.

Table 1. Dataset Description

Data Sets	#Samples	#Dimensions	#Number of class(k)
Iris	150	4	3
WDBC	569	30	2

The above data sets are used for testing the accuracy and efficiency of the proposed method. The value of k, given in Table 1. Table 2 shows the accuracy of k-means clustering with full gene space.

Table 2. Accuracy of K-Means Clustering With Full Gene Space

Data set	Full Gene space		
	Min	Max	Average
Iris	71.45	87.97	78.7
WDBC	75.04	75.04	75.04

In Table 3, the cosine measure is used to find the informative gene for iris dataset. The largest similarity value of each gene indicates the strong similarity between ith gene and jth eigenvector. We obtained two groups (G1 & G2) of genes according to D_{i1} and D_{i2}

respectively (hence we have set s to be two). Here, the 4th gene shows the high variance in the data. It achieves higher accuracy. Table 4 shows the gene selection on WDBC data. Here top l=5 genes are selected in each group and after removing the duplication only six genes are used for clustering to achieve higher accuracy. Table 5 shows that our proposed method achieves higher accuracy than the existing methods.

Table 3. Gene Ranking For Iris Dataset

Gene	Gene Rank		Top l genes	Gene Combination	Accuracy (%)
	G1	G2			
1	4	4	1	(4)	94.95
2	2	2	2	(4,2)	91.04
3	3	3	3	(4,2,3)	92.67
4	1	1	4	(4,2,3,1)	87.97

Table 4. Informative Gene Selection For Wdbc Data Set

Top l=5 gene in		Genes selected for clustering
Group 1	Group 2	
25	23	2, 6, 7, 22, 23, 25
23	22	
22	25	
6	6	
7	2	

Table 5. Performance Comparison On Wdbc Data With Informative Genes

Algorithm	Initial Centroid	Number of Run Times	Gene Space	Accuracy (%)
k-means	Random Selection	50	30	75.04
k-means+ PCA	Random Selection	50	6	75.04
Proposed Method	Computed by Program	1	6	81.38

Figure 1 shows the accuracy on data sets. Note that the proposed method provides better cluster accuracy than the existing methods. The clustering results of random initial center are the average results over 50 runs since each run gives different results. It shows the proposed algorithm performs much better than the random initialization algorithm.

The experimental results show the effectiveness of our approach. This may be due to the initial cluster centers generated by proposed algorithm are quite closed to the optimum solution and it also discover clusters in the low dimensional space to overcome the curse of dimensionality.

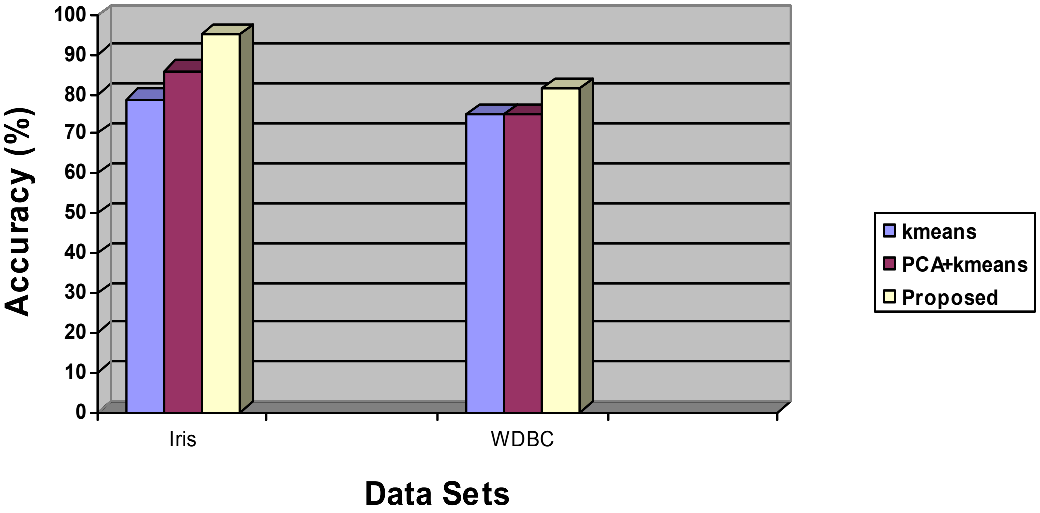

Figure 1. Accuracy on data sets: Iris, WDBC

6. Conclusion

In this paper, we have described the problem of sample clustering on high gene dimension datasets. We have proposed new approach to improve cluster accuracy for gene data. We have achieved higher performance by our proposed method when compared with the existing methods. In future work, we will apply this method to lymphoma and SRBCT datasets.

References

[1] 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, august 2000.

[2] Chris Ding and, Xiaofeng He,"K-means clustering via principal component analysis", In proccedings of the 21st international conference on machine learning Banff, Canada,2004.

[3] Daxin Jiang, Chun Tang, Aidong Zhang,"Cluster analysis for gene expression data: a survey, knowledge and Data Engineering",IEEE Transactions on , 2004;16(11),1370-1386.

[4] Golub T.R., D.K. Stonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller,M.Loh, J.Downing, M. Caligiuri,C.Bloomfield and E.Lander."Molecular classification of cancer class discovery and class prediction by gene expression", Science 286(5439),531-537,1999.

[5] Hyvarinen A. and E. Oja., "Independent component analysis: algorithms and applications", Neural network, 13(4-5):411-430,2000.

[6] Hyvarinen A., "survey on independent component analysis",Neural Computing Surveys,2:94-12,1999.

[7] Jollie I.,Principal Component Analysis. Springer-verleg, 1986.

[8] Jutten C. and Herault J. Blind Seperation of sources, part i: and adaptive algorithm based on neuromimetic architecture. Signal processing, 24: 1-10,1991.

[9] Margaret H. Dunham, Data Mining-Introductory and advanced concepts, Pearson education,2006.

[10] Merz C and Murphy P, UCI Repository of Machine Learning Databases.

[11] Misra et al. "interactive exploration of microarray gene expression patterns in a reduced dimensional space. Genome Res 2002, 12:1112-1120,2002.

[12] Rand, W.M. Objective criteria for evaluation of clustering methods. Journal of the American Statistical Association,1971.

[13] Tang C.,Zhang L., Zhang A., and Ramanathan"interrelated two-way clustering: An unsupervised approach for gene expression data analysis." In Proceeding of BIBE2001: 2nd IEEE international symposium on Bioinformatics and Bioengineering, pages 41-48,2001.
[14] Xing E.P. and Karp R.M. Cliff: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. Bioinformatics, Vol. 17(1):306-315,2001.
[15] Zhu, L. and C. Tang. 2006 microarray sample clustering using independent component analysis. Procedings of the 2006 IEEE/SMC international conference on system of systems engineering.112-117.
[16] http://www.mathwork.fr/matlabcentral/newsreader/view-thread/103251.

Authors

N. Tajunisha, received MCA degree from Madurai Kamaraj University and is pursuing her PhD in Computer Science at Mother Terasa Women's University. She is currently an Associate Professor at the Department of Computer Science, Sri Ramakrishna College for Women, Coimbatore. Her current research interest include Data mining. She has presented papers in International journals and IEEE conferences.

Dr. V. Saravanan is the Director of MCA Department at Dr.N.G.P Institute of technology, Coimbatore, India. He received his PhD in computer science from bharathiar university. He specialized on automated and unified data mining using intelligent agents. His research area includes data warehousing and mining, software agents and cognitive systems. He has presented many research papers in National, International conferences and Journals and also guiding many researchers leading to their PhD degree. He is the life member of Computer Society of India, Indian Society for Technical Education, and Indian Association of Research in Computing Sciences and and International Association of Computer Science and Information Technology.