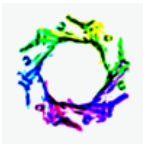
IJCB
International Journal of Computational Biology
ISSN: 2229–6700, E-ISSN: 2229–6719, Vol. 2, Issue 1, 2011, pp-24-31
Available online at http://www.bioinfo.in/contents.php?id=96



AN IMPROVED METHOD OF UNSUPERVISED SAMPLE CLUSTERING BASED ON INFORMATIVE GENES FOR MICROARRAY CANCER DATA SETS

TAJUNISHA N .[1*], SARAVANAN V.[2]
[1] Department of Computer Science, Sri Ramakrishna college of arts and Science for women Coimbatore, India
[2]Department of Computer Application, Dr.N.G.P Institute of Technology, Coimbatore, India
*Corresponding author. E-mail: [1]tajkani@gmail.com, [2]tvsaran@hotmail.com

Received: January 01, 2011; Accepted: February 22, 2011

Abstract— Microarrays have become the effective, broadly used tools in biological and medical research to address a wide range of problems. Many statistical methods are available for analyzing and systematizing the complex data into meaningful information, and one of the main goals in analyzing gene expression data is the detection of samples or genes with similar expression pattern. In microarray cluster analysis, there is a challenging problem; the dimension of gene expression is much larger than the sample size. The 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 clustering algorithm is used for much practical application. But its output is quite sensitive to initial position of cluster centers and the dimension of data. In this paper, we present a new framework for unsupervised sample based clustering using informative genes for microarray data. We propose the 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. We have applied our proposed method to cancer data sets. By comparing the results of original and new approach, it was found that the results obtained are more accurate.
Keywords: k-means; informative gene; dimension reduction; initial centroid; microarray gene data

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.

In Sample based clustering, samples are treated as objects and genes are features for clustering. 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 [11] 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 [10]. 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) [5] 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

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 (1) for the k-means algorithm is minimizing the SSE (Sum Squared Error).

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

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.

Before the k-means algorithm converges, distance and centroid calculations are done while loops are executed a number of times, say l, where the positive integer l is known as the number of k-means iterations. The precise value of l varies depending on the initial starting cluster centroids even on the same dataset. So the computational complexity of the algorithm is O(nkl), where n is the total number of objects in the dataset, k is the required number of clusters and l is the number of iterations. The time complexity for the high dimensional data set is O(nmkl) where m is the number of dimensions.

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) [8] and Independent component analysis (ICA) [6, 7] 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, 13]. Originally used in blind source separation (BSS) problems [9], 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. [19] applied ICA in gene dimension reduction and identified informative genes prior to the clustering. In our previous work [16], the new approach was proposed to find the initial centroid and reduce the dimension using PCA.

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. [18] 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 discriminant genes. It then applies a graph-theoretical clustering algorithm to generate initial partition for the samples and the selected genes. Tang et al. [17] 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.

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. To improve the efficiency of our approach, k-means is modified by using heuristic approach to assign the data point to cluster. Our algorithm is described as follows:

Algorithm 1: The proposed method

Steps:

1. Find the initial centroids c_j (1<=j<=k) for k-means using Algorithm 2.
2. Informative gene selection using Algorithm 3.
3. Cluster the data-points by algorithm 4.
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 i^{th} 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 ($\vec{v}_1, \ldots \vec{v}_s$) 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 $\vec{v}_1,\ldots\vec{v}_s$ may reveal the most variance in the data, the genes "similar" to $\vec{v}_1,\ldots\vec{v}_s$ should be relevant to the cancer. we use the cosine similarity measure[20] 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 one of the measure commonly used to compare the similarity of objects. Here in our work, 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

$$D_{ij} = \arccos\left(\frac{g_i v_j'}{\sqrt{(g_i g_i')(v_j v_j')}}\right) \quad (2)$$

Where i=1,2…,n genes , j=1,2,….s

Seen from (2), a large value of D_{ij} indicates more similarity between i^{th} gene and the j^{th} eigenvector. Therefore, we can rank genes based on the similarity values for each eigenvector. For j^{th} 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 $\vec{v}_1,\ldots\vec{v}_s$, we can obtain a set of genes with largest values of the cosine measure. Top selected genes with high variance are used for clustering.

The next algorithm is an iterative process which makes use of a heuristic method [4, 14] to improve the efficiency. During the iteration, the data-points may get redistributed to different clusters. The method involves keeping track of the distance between each data-point and the centroid of its present nearest cluster. At the beginning of the iteration, the distance of each data-point from the new centroid of its present nearest cluster is determined. If this distance is less than or equal to the previous nearest distance, that is an indication that the data point stays in that cluster itself and there is no need to compute its distance from other centroids. This result in the saving of time required to compute the distances to k-1 cluster centroids. On the other hand, if the new centroid of the present nearest cluster is more distant from the data-point than its previous centroid, there is a chance for the data-point getting included in another nearer cluster. In that case, it is required to determine the distance of the data-point from all the cluster centroids. This method improves the efficiency by reducing the number of computations.

Algorithm 4: Assigning data-points to clusters

Steps:

1. Compute the distance of each data-point x_i (1<=i<=n) to all the centroids c_j (1<=j<=k) using Euclidean distance formula..
2. For each data object x_i, find the closest centroid c_j and assign x_i to the cluster with nearest centroid c_j and store them in array Cluster[ ] and the Dist[ ] separately.
   Set Cluster[i] = j, j is the label of nearest cluster.
   Set Dist[i]= d(x_i, c_j), d(x_i, c_j) is the nearest Euclidean distance to the closest center.
3. For each cluster j (1<=j<=k), recalculate the centroids;
4. Repeat
5. for each data-point
   5.1 Compute its distance from the centroid of the present nearest cluster
   5.2 If this distance is less than or equal to the previous nearest distance, the data-point stays in the cluster
   Else
   For every centroid c_j
   Compute the distance of each data object to all the centre.
   Assign the data-point x_i to the cluster with nearest centroid c_j.
6. For each cluster j (1<=j<=k), recalculate the centroids;

Until the convergence criteria is met.

This algorithm requires two data structure Cluster [ ] and Dist [ ] to keep the some information in each iteration which is used in the next iteration. Array cluster [ ] is used for keep the label if the closest centre while data structure Dist [ ] stores the Euclidean distance of data object to the closest centre. The information in data structure allows this function to reduce the number of distance calculation required to assign each data object to the nearest cluster.

4.2 Cluster Validation

The Rand Index [15] 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

$$RI = \frac{a+d}{a+b+c+d} \qquad (3)$$

The Rand Index lies between 0 and 1. Higher values of the Rand Index indicate better performance of the algorithm. In this paper, Rand Index measure is used to find the clustering accuracy. In our experiments, we calculate rand index value between the ground truth and the results of predicted cluster to evaluate the quality of the proposed method.

Experimental Results

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

Wisconsin Diagnostic Breast Cancer (WDBC) data [12]. It contains 569 samples and each sample is measured over 30 genes. Alizadeh et. Al. (2000) expression profiles of 62 lymphoma samples were produced with 4026 genes. The samples represents the following types of malignancies: chronic lymphocytic leukemia (CLL,11 samples) , follicular lymphoma (FL,9 samples) and diffuse large B-cell lymphoma (DLBCL, 42 samples). SRBCT dataset has 2308 genes and 63 experimental conditions, 8 Burkitt Lymphoma (BL), 23 Ewing Sarcoma (EWS), 12 neuroblastoma (NB) , and 20 rhabdomyosarcoma (RMS) (Khan et al., 2001).

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.

The data sets used for testing the accuracy and efficiency of the proposed method and the value of k, given in Table 1, Table 2 shows the accuracy of k-means clustering with full gene space on the given datasets. Table 3 shows the informative genes selected for cancer datasets. Informative genes are selected based on cosine similarity value. The largest similarity value of each gene indicates the strong similarity between ith gene and jth eigenvector. We obtained two groups (Group1 & Group2) of genes according to D_{i1} and D_{i2} respectively (hence we have set s to be two for WDBC and SRBCT). 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 and Table 6 shows that our proposed method achieves higher accuracy than the existing methods on WDBC and SRBCT data.

In Table 6, top l=20 genes are selected in each group and after removing the duplication only thirty two genes are used for clustering to achieve higher accuracy. Table7 shows accuracy on lymphoma data set. Here the top l=49 genes are selected in each group. We have set s=5 for this data set. After removing the duplication, 223 genes are used for clustering to achieve higher accuracy.

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. Our proposed method converges quickly and number of iteration is reduced for the convergence than the existing methods. 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 with better accuracy in the low dimensional space to overcome the curse of dimensionality. Our proposed method with the heuristics approach to assign the data points to cluster reduces the running time with the same accuracy. This approach improves the efficiency of our method.

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. Our experimental results show that the quality of clustering results of k-means algorithm using informative genes and with fixed initial centroid is better than that of pure k-means. Though the proposed method gave better quality results in all cases, over random initialization, still there is a limitation associated with this, i.e. the number of clusters(k) is required to be given as input. Evolving some statistical methods to compute the value of k, depending on the data distribution is suggested for future research.

References

[1] Alter O., Brown P.O. and Bostein D. (2000) *Proc. Natl. acad.sci. USA*, vol. 97(18): 10101-10106..

[2] Chris Ding and Xiaofeng He (2004) *In proccedings of the 21st international conference on machine learning Banff, Canada.*

[3] Daxin Jiang, Chun Tang, Aidong Zhang (2004) *IEEE Transactions on Knowledge and Data Engineering,*16(11),1370-1386.

[4] Fahim A.M., Salem A.M., Torkey F. A., Saake G. and Ramadan M.A. (2009) *Georgian Electronic Scientific Journal: Computer Science and Telecommunications*, Vol.2, No. 19,pp. 47-57.

[5] Golub T.R., Stonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M., Downing J., Caligiuri M., Bloomfield C. and Lander E. (1999) *Science* 286(5439),531-537.

[6] Hyvarinen A. and Oja E. (2000) *Neural network*, 13(4-5):411-430.

[7] Hyvarinen A. (1986) *Neural Computing Surveys*,2:94-12,1999.

[8] Jolliffe, I.T. (*2002) Principal Component Analysis, second edition, New York: Springer-Verlag New York, Inc*

[9] Jutten C. and Herault J. (1991) *Signal processing*, 24: 1-10.

[10] Kumar Dhiraj, Santanu Kumar Rath, Korra Sathya Babu (2009) FCM for Gene Expression Bioinformatics Data. IC3: 521-532.

[11] Margaret H. Dunham (2006) *Data Mining-Introductory and advanced concepts, Pearson education*.

[12] Merz C. and Murphy P. (1997) *UCI Repository of Machine Learning Databases.*

[13] Misra et al. (2002) *Genome Res* , 12:1112-1120.

[14] Nazeer K. A., Abdul and Sebastian M.P. (2009) *Proceedings of the World Congress on Engineering*, 1, 308-312.

[15] Rand W.M. (1971) *Journal of the American Statistical Association,* 66, 846 -850.

[16] Tajunisha N., Saravanan V. (2010) *Proceedings of the IEEE fist international conference on integrated intelligent computing,* 17-21.

[17] Tang C., Zhang L., Zhang A. and Ramanathan (2001) *In Proceeding of BIBE2001: 2nd IEEE international symposium on Bioinformatics and Bioengineering*, 41-48.

[18] Xing E.P. and Karp R.M. (2001) *Bioinformatics*, Vol. 17(1):306-315.

[19] Zhu L. and Tang C. (2006) *Proceedings of the 2006 IEEE/SMC international conference on system of systems engineering.*112-117.

[20] http://www.mathwork.fr/matlabcentral/newsreader/view-thread/103251

Table 1- Dataset Description

Data Sets	#Samples	#Dimensions	#Number of class(k)
WDBC	569	30	2
SRBCT	63	2308	4
LYMPHOMA	62	4026	3

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

Data Sets	Accuracy on Full Gene space (%)		
	Min	Max	Average
WDBC	75.04	75.04	75.04
SRBCT	58.22	64.26	61.14
LYMPHOMA	70.44	97.36	71.92

Table 3- Informative genes for Cancer datasets

Data Sets	Original data size	Informative gene space
WDBC	30 * 569	6*569
SRBCT	2308*63	32*63
Lymphoma	4026*62	223*62

Table 4- Informative gene selection for WDBC data set

Top l=5 genes 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

Table 6- Performance comparison on SRBCT data with informative genes

Algorithm	Initial Centroid	Number of Run Times	Gene Space	Accuracy (%)
k-means	Random Selection	50	2308	61.14
k-means+ PCA	Random Selection	50	32	64.26
Proposed Method	Computed by Program	1	32	70.25

Table 7- Performance Comparison on Lymphoma Data with Informative Genes

Algorithm	Initial Centroid	Number of Run Times	Gene Space	Accuracy (%)
k-means	Random Selection	50	4026	71.92
k-means+ PCA	Random Selection	50	223	74.92
Proposed Method	Computed by Program	1	223	97.36

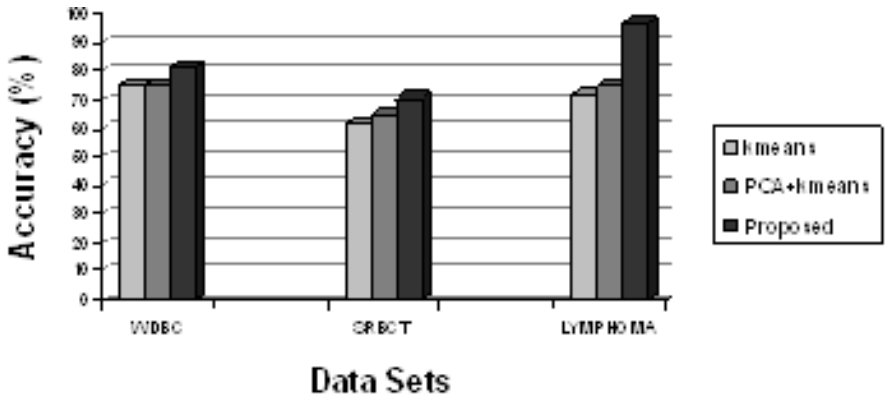


Fig. 1- Accuracy on data sets: WDBC, SRBCT, LYMPHOMA