Novel Hybrid Hierarchical-K-means Clustering Method (H-K-means) for Microarray Analysis

Bernard Chen[1*], Phang C. Tai[2], R. Harrison[1] and Yi Pan[1]

1. Department of Computer Science, Georgia State University, Atlanta GA 30303, USA
2. Department of Biology, Georgia State University, Atlanta GA 30303, USA

[*] *Contact email:bchen3@student.gsu.edu*

Abstract

Hierarchical and k-means clustering are two major analytical tools for unsupervised microarray datasets. However, both have their innate disadvantages. Hierarchical clustering cannot represent distinct clusters with similar expression patterns. Also, as clusters grow in size, the actual expression patterns become less relevant. K-means clustering requires a specified number of clusters in advance and chooses initial centroids randomly; in addition, it is sensitive to outliers. We present a novel hybrid approach to combined merits of the two and discard disadvantages we mentioned above. It is different from existed method: carry out hierarchical clustering first to decide location and number of clusters in the first round and run the K-means clustering in another round. The brief idea is we cluster around half data through hierarchical clustering and succeed by K-means for the rest half in one single round. Also, our approach provides a mechanism to handle outliers. Comparing with existed hybrid clustering approach and K-means clustering in 2 different distance measure on Eisen's yeast microarray data, our method always generate much higher quality clusters.

1. Introduction

Advances in microarray technology have made it possible to simultaneously monitor the expression of thousands of genes in genomes. The challenge is to effectively analyze and interpret this large volume of information. Exploration of gene expression data sets is always problematic due to its inherent dispersion and missing values (Quackenbush, 2001). Thus, dealing with outliers is also an important issue toward analysis methods. Clustering methods can be divided into two general classes, designated supervised and unsupervised clustering. In this paper, we focus on unsupervised clustering which may again be separated into two major categories: partition clustering and hierarchical clustering. There are many algorithms for partition clustering category, such as k-means clustering (MacQueen 1967), k-medoid clustering, genetic k-means algorithm (GKA), Self-Organizing Map (SOM) and also graph-theoretical methods (CLICK, CAST). Among those methods, K-means clustering is the most popular one because of simple algorithm and fast execution speed. However, there are three major parts that require improvements: First, the number of k (clusters) must be decided before execution. Second, random choosing of the initial start points makes it impossible to obtain reliable results without much iteration of the entire clustering process (Shin et al, in preparation). Third, it's sensitive to outliers. Although hierarchical clustering nests and represents the clusters as a dendrogram that provides an easy understanding of the data, the quality of clusters often degrades as more data are joined. It is becoming increasingly clear that none of the approaches alone are sufficient and that the application of various techniques will allow different aspects of the data to be explored. Therefore, we seek to combine the strength of both approaches and discard the disadvantages. Initially, we tried to generate starting points for k-means clustering from the hierarchical method, but the results are not good enough. In the end, a novel concept is to carry out the hierarchical clustering as the first step, and then not only generates the required information for K-means, but makes good use of clusters created in the first step. In this paper, we illustrate how this technique can be carry out by using Eisen *et al* yeast microarray data (79samples × 2467genes).

Section 2 gives the introduction of two distance measurement for DNA chip data. Section 3 discusses both K-means clustering and hierarchical clustering. Section 4 shows how we create our algorithm and how it works. Section 5 compares H-K-means with the old methods and draw conclusions. We have our discussion in the last part, section 6.

2. Distance Measurements

When clustering data, we want to group together observations that are similar. Thus, we need to be able to compute the distance between two data objects, but it can be defined in many forms. In the following, we introduce two major measurements that we used in this paper. (Because the gene chips datasets are always in high dimensional space, let us assume that we are working on n-dimensional data)

2.1 Euclidean distance

In two dimensions, it is the length of the straight line connection to two points x and y. In n dimension the Euclidian distance is defined as:

$$|X - Y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

One problem with this method is that Euclidean distance is sensitive to large values; in other words, it is sensitive to outliers. Besides, it will miss negative correlations since they will give large distance.

2.2 Pearson Correlation Coefficient

It is a widely used measurement of microarray gene expression data defined as follows:

$$P_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

Here, $\overline{x}$ and $\overline{y}$ is the mean of x and y, respectively. This looks very daunting but in a sense it just computes the ratio between the co-variance of x and y and the product of their standard deviations. The correlation can be anything from 1 to -1. It is invariant to scale and location of the data points, unlike Euclidean distance. However the measurement is still somewhat sensitive to outliers.

3. Clustering Methods

Once we have a set of gene expression or samples, we would like to group them together so that genes with similar expressions or samples having similar conditions are in the same group. This procedure is called clustering. Two major categories in this field are partition methods and hierarchical methods.

K-means algorithm is the most widely used method in partition category due to its fast speed and easy understanding. However, the major disadvantage of this method is that the number k is often not known in advance. Furthermore, randomly chosen initial points may cause two points that are close in distance to be determined as two centroids. Also, it is sensitive to outliers. Tamayo *et al* explain that K-means clustering is "a completely unstructured approach, which proceeds in an entirely local fashion and produces an unorganized collection of clusters that is not conductive to interpretation." (Tamayo 1999)

After Eisen *et al* successfully applied hierarchical clustering into microarray gene expression data in graphic form (Eisen 1998); it has become the most popular clustering method in this area. In this paper, we apply agglomerative hierarchical clustering (bottom-up) method as our analysis tool. The drawback of agglomerative hierarchical clustering is best described by Quackenbush: "one potential problem with many hierarchical clustering methods is that, as clusters grow in size, the expression vector that represents the cluster might no longer represent any of the genes in the cluster. Consequently, as clustering progresses, the actual expression patterns of the genes themselves become less relevant" (Quackenbush 2001). As a result, an active area of research in agglomerative hierarchical clustering is in detecting when to stop the merging of elements.

4. Our Approach

Since the weak point of hierarchical clustering is its termination, and the problem of K-means is its initiation, it is intuitive to combine two methods together.

4.1 Naïve Approach

At the beginning, we tried to run agglomerative hierarchical clustering to get initial information (number of cluster and initial seeds location) for k-means clustering. Every time when bottom-up clustering joined two clusters together and computed the mean value of each of their attributes as their new attributes, we considered the new attribute as our initial point for k-means clustering. If the distance of a new point compared with the existing initial points is not close to any one of them, then we had a new initial centroid; if it is near any one of existing centroids, then we computed the mean value of the new attribute with the nearest initial point to adjust the initial seed's attribute values. After attaining certain termination conditions for hierarchical clustering, we then began k-means clustering with initial number of clusters and locations calculated from hierarchical clustering.

We tried this method in all possible different stop points of hierarchical clustering and then complete the following k-means clustering. Then, we got many different clustering results with different number of clusters and initial centroids. Afterward, we computed the within-cluster distance between any point in the cluster and the centroid in correlation coefficient (better result if closer to 1), as shown in Figure 1. The results in this method were not always better than the original algorithm. We analyzed the member of clusters to figure out why good results were not generated. We realized some samples of gene expression data clustered together by hierarchical clustering did not group together after the following k-means clustering. It was because at the beginning two or more centroids were far from each other, but after adjusting their locations they may become closer. Thus, a means to merge centroids or some better methods may be required to improve this hybrid approach

4.2 New Algorithm

First, we carried out agglomerative hierarchical clustering and let the program stop at a certain terminal point. From the clusters generated from hierarchical clustering, we computed the mean value of each cluster as the initial point for k-means to obtain the initial centroid. Also, the number of clusters generate from hierarchical clustering is k-mean's number of clusters. After that, we worked on k-means clustering with which every cluster MUST at least contain the same objects generated from hierarchical clustering. This is due to the fact that hierarchical clustering had already put objects that were very close with one another into clusters, and the goal of k-means clustering is to put close objects together, which is in the same direction. Therefore, we can trust the results of hierarchical clustering. Besides, in order to deal with outliers, we also set a threshold for k-means: after counting the distance with existing clusters, if the shortest distance is not further than threshold, we assign the dataset to its closest cluster. If the shortest distance exceeds the threshold, it may consider belongs to minor group.

The objects in the minor group are defined as objects that did not belong to any major groups. It is not always outliers, because if the terminal point of hierarchical clustering were set too early, then we can only get few numbers of clusters. Thus, there may be some objects which were not outliers that did not belong to one of the small number of clusters. We can also do cluster again in minor group to get more information if necessary.

The advantage of this method was that people didn't have to choose an arbitrary number of k; instead, the user only had to choose the percentage for execution of hierarchical clustering (the stop point for the first step). The initial centroids were also generated in a much better way. Besides, points close to one another wouldn't be chosen as different centroids since they were already clustered together.

```
(1)repeat
(2)    find two objects(clusters) with closest distance
         among all, and cluster them together.
(3)    the value of attributes of new cluster are the
         average of attributes of two old objects(clusters);
(4)until the percentage of hierarchical clusters
     requested by user is done;
(5)calculate average attribute values of members of
     clusters that generate from step (1) to (4) as initial
     cluster centroids;
(6)repeat
(7)     for all objects
(8)       if the object already appeared in step(2)
             then the object remain  in original cluster;
(9)       else
             calculate distances between the object and
              existed clusters
(10)          if the shortest distance lower than
              threshold
(11)              then the object are assigned to the
                  closest cluster
(12)          else
                  the object belongs to minor group
(13)    end for loop
(14)    update the centroid attribute value;
(15)until no member changes belonging cluster;
```

Figure 2 H-K-means algorithm

5. Results and Conclusions

We applied our new method to Eisen's yeast gene expression data to cluster samples, and we set 0.3 correlation distance as our outlier threshold. We chose Pearson correlation coefficient as our distance measure because it is more meaningful in DNA chips data and easier to set a threshold to control outliers without using any normalization on the distance measured. We also worked on normal k-means clustering with two different distance measurements. Comparison was shown in Figure 3. From the figure, our approach had better within-cluster distance most of time. In addition, the true power of our algorithm is that objects which are close to one another would not be separated. Because if objects were close, they would be chosen in the hierarchical clustering steps to merge as a cluster. The reason our method was slightly worse than k-means in the end part was that we almost finished hierarchical clustering which meant the drawback of hierarchical clustering was revealed. Figure 3 is the relation between quality of cluster and percentage that hierarchical cluster had completed. Figure 4 gives the relation between numbers of cluster generated and percentages that hierarchical clustering had completed.

Our approach is very flexible to terminal conditions of hierarchical clustering, but there were two things of concern: quality of cluster and number of objects that belong to minor group. If hierarchical clustering terminates too early or too late, as shown in Figure 3, the quality of cluster can not be good. Besides, if hierarchical clustering complete in low percentage condition, it will generate only few number of clusters, possibly assigning many objects to minor group. Figure 5 presents the relation between percentages that hierarchical clustering has completed and the number of objects belong to minor group. Under this two condition and the results we got, we assert that our approach could generate better results if we terminate hierarchical clustering at around 40% to 60%. Within this criterion, we may obtain well cluster quality and few objects belong to minor group.

6. Discussion

In this paper, we have proposed a novel clustering method for micorarray gene expression data. Our method of automatically finding good initial centroids for K-means clustering and dealing with outliers seems to provide better performance and more meaningful results. In the future, we will continue working on combining hierarchical clustering with other clustering methods that required initial information to start, such as K-medoid algorithm, genetic algorithm. Automatically generate stop point for first step and obtain non-spherical shape clusters are two major improvable techniques. We will also apply our clustering method to other fields. Both finding sequence motif through our clustering method and clustering transmembrane prediction classification rules are two of our major tasks in the near future. We believe the potential for additional progress in this new method is strong.

References

[1] Eisen, M.B. (1998) *Cluster analysis and display of genome-wide expression pattern*. Proc, Natl. Acad. Sci. USA, 95:14863-14868.

[2] MacQueen, J. *Some methods for classification and analysis of multivariate observations.* Proc: 5th Berkeley Symp. Math. Statist, Prob, 1:218-297, 1967.

[3] Quackenbush, J. (2001) *Computation analysis of microarray data.* Nat. Rev Genet., 2, 418-427.

[4] S. Kwon and C. Han (2002) *Hybrid Clustering Method for DNA Microarray Data Analysis*, Genome Informatics 13: 258-259

[5] Tamayo, P. et al. (1999) *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differtiation*. Proc. Natl. Acad. Sci. USA 96, 2907-2912.

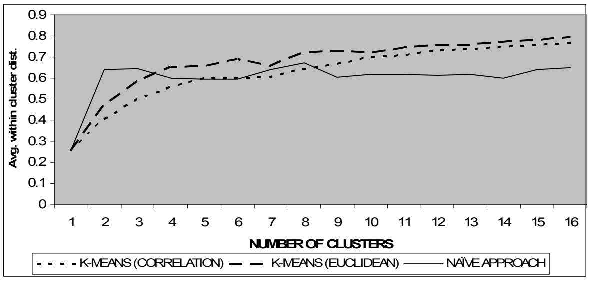

Figure 1 Comparison of Naïve approach with normal k-means in two different distance measure.

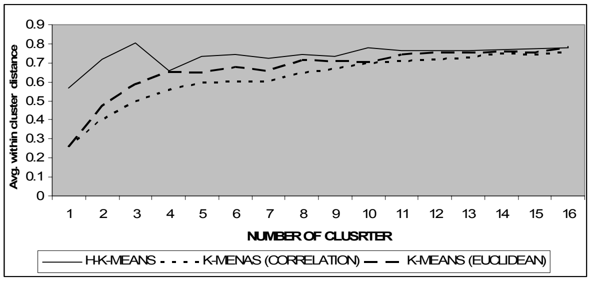

Figure 3 Comparison of our H-K-means with normal K-means in two different distance measure.

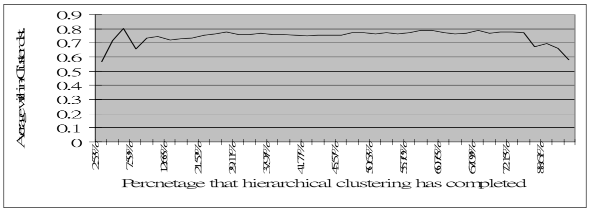

Figure 4 the relation between percentage hierarchical clustering has completed and cluster quality.

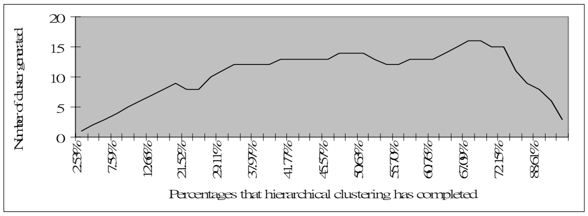

Figure 5 the relation between percentages hierarchical cluster has completed and generated number of clusters.

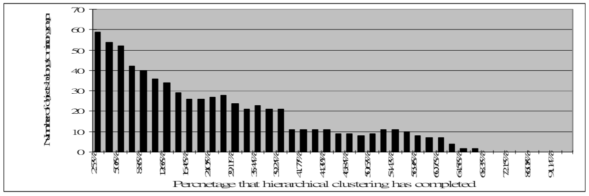

Figure 6 presents the relation between percentages that hierarchical clustering has completed and number of objects belong to minor group