Automatically Finding Good Clusters with Seed K-Means

Miyoung Shin shinmy@etri.re.kr
Eun Mi Kang emkang@etri.re.kr
Seon Hee Park shp@etri.re.kr

Bioinformatics Research Team, Electronics and Telecommunications Research Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Korea.

Keywords: gene expression data analysis, k-means clustering, seed extraction

1 Introduction

In finding biologically relevant groups of genes with gene expression data obtained by microarray technologies, the k-means clustering method is one of the most popular approaches due to its easiness to use and simplicity to implement. However, the randomness of k-means clustering method in choosing initial points to start with makes it impossible to obtain reliable results without much iteration of the entire clustering process [2]. Our goal here is to introduce a novel clustering method, which we call it *seed k-means clustering*, where a novel algorithm is employed to automatically find good initial seeds for k-means clustering.

2 Seed K-Means Clustering Method and Its Evaluation

2.1 Clustering Algorithm

The seed k-means clustering method is basically a two-phase process: seed extraction and cluster generation. Given the number of clusters, k, the first phase of seed extraction is to automatically select k good initial seeds of genes by analyzing their expression patterns. The goodness of initial seeds in our method is defined by how well they are distributed to capture the entire input space while not being very close to each other, i.e., far away enough to remove the information redundancy. The detailed description about seed extraction is presented in [4]. In the second phase, cluster generation is performed by employing k-means clustering method where the selected seeds are used as initial centroids of clusters. Thus, once the seeds are chosen, each gene in the dataset is assigned to the cluster whose centroid is the closest to the current gene. After all the assignments are made, the current centroids of k clusters are refined by taking the averaged expression profiles of the genes assigned to the respective clusters. This process is iterated until each centroid of clusters reach local minimum. The cluster memberships based on these centroids become our final clustering results.

2.2 Evaluation Method

To evaluate our clustering method, we used two datasets of which target clusters are already known. One dataset is the 5-cluster synthetic dataset generated by adding random noise with $N(0, 0.1^2)$ to the predefined five different patterns of log expression measures for 10 experiments, which are partially taken from the 9-cluster synthetic gene expression dataset of [3]. The other dataset is the real gene expression data regarding yeast cell cycles obtained by Cho *et al.* [1]. Here we used only 384 genes [5] whose expression levels peak at different time points corresponding to the five phases of cell cycle. For validation of the clustering results, the adjusted rand index [5] is computed which has been used to assess the degree of agreement between two partitions of clustering results and target clusters. As the value of the adjusted rand index is closer to 1, clustering results mean to be closer to target clusters.

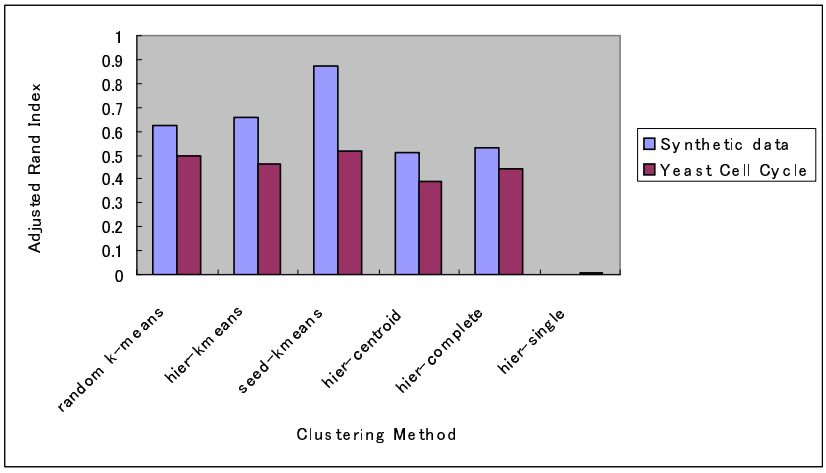

Figure 1: The summary of clustering results for the six clustering algorithms.

3 Experiment Results

For both datasets, we generated five clusters, respectively, by seed k-means clustering methods while using Euclidean distance as a distance metric. To compare with other initial selection methods for k-means clustering, we also generated the clusters by k-means clustering using randomly selected genes and the cluster means of hierarchical centroid-linkage clustering results, respectively, as initial points. In addition, for comparison with other clustering methods, we also generated the clusters by three hierarchical clustering methods using centroid-linkage, complete-linkage and single-linkage, respectively. Figure 1 is the summary of the clustering results for the six clustering methods mentioned earlier. Here you can see that the seed k-means clustering method performs best on two different datasets.

4 Discussion

Here we presented the usefulness of the seed k-means clustering method in identifying biologically relevant groups of genes based on microarray gene expression data. Our method of automatically finding good initial seeds for k-means clustering seems to provide better performance of k-means clustering than other initial selection methods popularly used for gene expression analysis. Further, note that clustering results obtained by the seed k-means clustering method are reproducible and consistent.

References

[1] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., and Davis, R.W., A genome-wide transcriptional analysis of the mitotic cell cycle, *Molecular Cell*, 2:65–73, 1998

[2] Datta, S. and Datta, S., Comparisons and validation of statistical clustering techniques for microarray gene expression data, *Bioinformatics*, 19(4):459–466, 2003.

[3] Quackenbush, J., Computational analysis of microarray data, *Nat. Rev. Genet.*, 2:418–422, 2001.

[4] Shin, M. and Park, S.H., Automatic detection of good initial seeds for clustering gene expression data, in preparation.

[5] Yeung, K.Y. and Ruzzo, W.L., Principle component analysis for clustering gene expression data, *Bioinformatics*, 17(9):763–774, 2001.