

Gene Expression Analysis Using Fuzzy K-Means Clustering

Chinatsu Arima

arima@brs.kyushu-u.ac.jp

Taizo Hanai

taizo@brs.kyushu-u.ac.jp

Masahiro Okamoto

Okahon@brs.kyushu-u.ac.jp

Laboratory for Bioinformatics, Graduate School of Systems Life Sciences, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

Keywords: gene expression analysis, Fuzzy k-means, clustering

1 Introduction

The recent advances of array technologies have made it possible to monitor huge amount of genes expression data. Clustering, for example, hierarchical clustering, self-organizing maps (SOM), k-means clustering, has become important analysis for such gene expression data. We have applied the Fuzzy adaptive resonance theory (Fuzzy ART) [5] to the gene clustering of DNA microarray data and the clustering result using this method was more suitable for biological knowledge than those of the ordinary method including hierarchical clustering, SOM, and k-means clustering. In this study, therefore, Fuzzy k-means [2, 3] clustering method was applied to this data, since this method also have fuzziness as Fuzzy ART. We verified the clustering results using Fuzzy k-means clustering by comparing with those of hierarchical clustering, k-mean clustering, Fuzzy ART and SOM.

2 Method

2.1 Fuzzy K-Means Clustering

The fuzzy k-means clustering (Fig. 1) is done with based on following equation (1).

$$J(K, m) = \sum_{k=1}^K \sum_{i=1}^N (u_{ki})^m d^2(x_i, c_k) \quad (1)$$

K and N are the number of clusters and genes in the data sets, m is a parameter which relate to ‘fuzziness’ of resulting clusters, u_{ki} is the degree of membership of gene x_i in cluster k , $d^2(x_i, c_k)$ is the distance from gene x_i to centroid c_k . The parameters in this equation are the cluster centroid vector c_k and the components of the membership vectors u_{ki} . These unknown parameters can be optimized by Lagrange method. Calculated u_{ki} shows the belonging ratio to a cluster k and centroid c_k shows the representative gene expression profile of a cluster k . In this study, a parameter m was set to 2.0 and the number of clusters was set to 5. For the number of the clusters in the other clustering method, we selected the same number as that of clusters using Fuzzy k-means clustering in order to compare the clustering results.

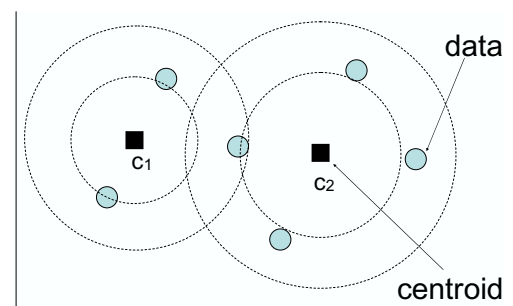


Figure 1: Fuzzy k-means clustering.

2.2 Data Preprocessing

In this study, we used expression data from a study of Chu *et al.* [1]. *Saccharomyces cerevisiae* was synchronized by transferred them to sporulation medium (SPM) at $t=0$ to maximize the synchrony of sporulation. RNA was harvested at time $t=0, 0.5, 2, 5, 7, 9$ and 11.5 hours after transfer to SPM. Polyadenylated RNA was prepared by purification with oligo(dT) cellulose column. Each gene's mRNA expression level just before transfer to SPM was used as control. About 6100 genes of expression profiles are included in this data [6]. Using them, we followed the same method [1] to extract the genes that showed significant increase of mRNA levels during sporulation. Among them, we finally selected 45 genes, whose functions are biologically characterized by Kupiec *et al.* [4].

3 Results and Discussion

The result of Fuzzy k-means clustering is shown in Fig. 2. This figure shows the representative time course data in each cluster and these values come from the centroid. 'Early', 'Middle', 'Mid-Late' and 'Late' genes, which were characterized in Mitchell, were used as 'index genes'. As the result, the cluster 1, 2 and 3 have only 'Early' genes, cluster 4 have only 'Mid-late' genes and cluster 5 have three 'Late' genes and two 'Middle' genes. In order to compare the result of clustering methods, we defined the correctness ratio for the clustering result based on index genes. The calculation for the correctness ratio was executed as follows. The majority of the index gene defined the character of the cluster. The correctness ratio was calculated by division by the number of minor genes in the total number of genes in the cluster. Table 1 shows the correctness ratios of five clustering method.

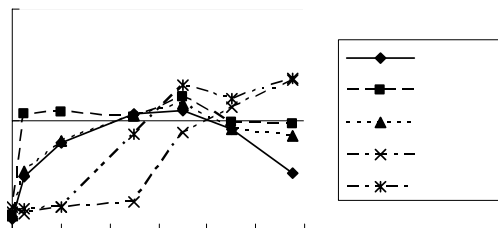


Figure 2: The result of Fuzzy k-means clustering.

Table 1: The correctness ratios five clustering algorithm.

Fuzzy k-means	Fuzzy ART	Hierarchical clustering	k-means clustering	SOM
0.90	0.90	0.81	0.86	0.86

References

- [1] Chu, S., Derisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I., The transcriptional program of sporulation in budding yeast, *Science*, 282:699–705, 1998.
- [2] Dembele, D. and Kastner, P., Fuzzy C-means method for clustering microarray data, *Bioinformatics*, 19:973–980, 2003.
- [3] Gasch, A. and Eisen, M., Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering, *Genome Biology*, 3(11):research0059.1–research0059.22, 2002.
- [4] Kupiec, M., Ayers, B., Esposito, R.E., and Mitchell, A.P., The molecular and cellular biology of the yeast *Saccaromyces*, *Cold Spring Harbor*, 889–1036, 1997.
- [5] Tomida, S., Hanai, T., Honda, H., and Kobayashi, T., Gene expression analysis using Fuzzy ART, *Genome Informatics*, 12:245–246, 2001.
- [6] The data set is available at <http://cmgm.stanford.edu/pbrown/sporulation/>