Fractal Clustering for Microarray Data Analysis

Lu-yong Wang*, Ammaiappan Balasubramanian, Amit Chakraborty, Dorin Comaniciu
Integrated Data Systems Department
Siemens Corporate Research
Luyong.Wang@siemens.com

Abstract

DNA microarray experiments generate a substantial amount of information about global gene expression. Gene expression profiles can be represented as points in multi-dimensional space. It is essential to identify relevant groups of genes in biomedical research. Clustering is helpful in pattern recognition in gene expression profiles. Some clustering techniques have been introduced. However, these traditional methods mainly utilize shape-based assumption or distance metric to cluster the points in multi-dimension linear Euclidean space. Poor consistence with the functional annotation of genes is shown in their validation study.

We propose fractal clustering method to cluster genes using intrinsic (fractal) dimension from modern geometry. Fractal dimension is used to characterize the degree of self-similarity among the points in the clusters. The main idea of fractal clustering is to group points in a cluster in such a way that none of the points in the cluster changes the cluster's intrinsic dimension radically. We computed Hausdorff fractal dimension through the means of the box-counting plot algorithm, since it is the fastest and also robust enough.

We assess this method using validation assessment using public microarray dataset. It shows that this method is superior in identifying functional related gene groups than other traditional methods.

1. Introduction

A critical aspect in the analysis of gene expression data is identification of clusters of genes that have similar expression patterns. Clustering techniques transform a large matrix of expression levels for different genes in different conditions into a more organized and informative collection gene sets, which are expected to share similar biological properties. Clustering techniques are predominantly influential in tissue classification, function annotation, and other biomedical applications[1]. Clustering has been studied for a long time in statistical learning. There are many algorithms that was carried out in gene expression profiling, *i.e.*, hierarchical clustering method, k-means clustering, self-organizing map (SOM) [1], principle component analysis[1], fuzzy c-means clustering [2], CLICK [3], adaptive quality-based clustering [4], quantum clustering[5], mean-shift [5], bagged clustering [6] and Gustafson-Kessel method[7], etc. However, these clustering methods have their inherent limitations in shape-sensitive, Euclidean-based and dubious linear relationship assumption.

In this paper, we propose a fractal clustering algorithm in gene expression profiling. It provides a very natural way of defining clusters that is not restricted to any particular cluster shape. This algorithm is based on the introduction of the concept of the fractal dimension in clusters. This method clusters points in such a way that data points in the same cluster are more self-affine among themselves than to the points in other clusters, although the clusters do not have to be ideal fractal themselves[8].

2. Fractal Clustering for Microarray analysis

For a cluster set of n points in a D-dimensional space, the traditional way to compute fractal dimension by the means of the box-counting plot algorithm, which was described in details in physics literature [8]. Briefly, for a set of n points, each of D dimensions, the space is divided in grid cells of size ε (hypercubes of dimension D). If N(ε) is the number of cells occupied by points in the data set, the plot of N(ε) versus ε in log-log scale is called box-counting plot. The negative value of the slope of that plot is called Hausdorff fractal dimension. The following formula only gives a brief concept of Hausdorff fractal dimension *d* (detailed maths derivation and computation is in [8]). :

$$d = -\lim_{\epsilon \to 0} logN(\epsilon)/log(\epsilon)$$

Initial clusters are required as the input for the fractal clustering as the consequence. *K*-means can be utilized as the initialization step. Each cluster found by the initialization step can be represented as C= $\{C_1,C_2,\ldots, C_f\}$, where C_i is the set that represent cluster

i. Let $F_d(C_i)$ be the fractal dimension of cluster i. The incremental step brings a new set of points to main memory and proceeds to take each point and add it to each cluster, computing its new fractal dimension. The algorithm computes the fractal dimension for each modified cluster after adding a point to it. Then, find the appropriate cluster to place the point by computing the minimal fractal impact, *i.e.*, the minimal change in fractal dimension generated by any of current clusters when the new point is added. We use Hausdorff fractal dimension, as it is the fastest in the fractal dimension definitions and also robust enough for clustering task.

3. Experiments and Results

Figure 1 shows the fractal clustering results on these published Mizuguchi's ATP regulation microarray data set [9]. In this figure, we utilized k-means as the initialization method. The number of clusters is 4. We take 3108 out of 6215 genes randomly, and carried out k-means for cluster initialization for the subsequent fractal clustering. Intuitively, there is no apparent tendency to have more than one shape clusters from traditional point of view in the figure. However, Mizuguchi's ATP regulation study indicated that there are around 3~4 functional clusters. It shows the limitation of traditional methods in identifying functional related clusters. Fractal clusters rely on intrinsic relationship in the cluster set, rather than geometric shape or distances in traditional methods.

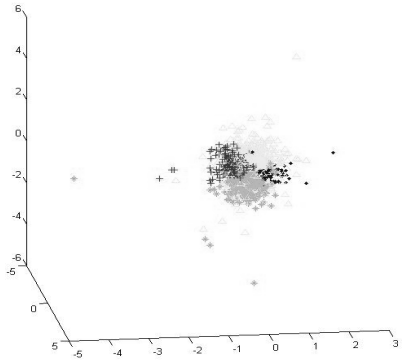

Fig 1. Fractal clustering results on Mizuguchi's microarray dataset

Moreover, we carried out this method on yeast microarray dataset, and use knowledge-driven assessment technique to compare our method with other microarray clustering techniques. Table 1 summarized the optimal clustering results by different clustering methods for the yeast ATP-regulation data set. It shows the highest *z*-score for each method. The results show that fractal clustering has the highest *z*-score, which indicated most consistent with gene annotations.

TABLE 1. EVALUATION OF DIFFERENT CLUSTERING METHODS.

METHOD	Z-SCORE	NUMBER OF CLUSTERS
Fractal Clustering	23	4
GK(ρ=1)	20.3	3
GK(ρ=5)	20.4	6
k-means*	3.97	5
SOM*	1.84	3
BagClust*	0.87	10
Adaptive quality-based clustering	3.71	3
Fractal clustering with FCM	27.4	4

References

[1] J. Quanckenbush, "Computational analysis of microarray data," *Nature Review Genetics*, vol. 2, pp. 418-427, 2001.

[2] D. Dembele, Kastner, P., "Fuzzy C-means method for clustering microarray data," *Bioinformatics*, vol. 19, pp. 973-980, 2003.

[3] R. Sharan, Maron-Katz, A., and Shamir, R., "Click and expander: a system for clustering and visualizing gene expression data," *Bioinformatics*, vol. 19, pp. 1787-1799, 2003.

[4] F. D. Smet, Mathys J., Marchal K., THijs G., Moor BT and Moreau Y, "Adaptive quality-based clustering of gene expression profiles," *Bioinformatics*, vol. 2002, pp. 735-746, 2002.

[5] D. Barash, Comaniciu,D., "Meanshift clustering for DNA Microarray Analysis," *Proceeding of the 2004 IEEE Computational Systems Bioinformatics Conference*, 2004.

[6] S. F. Dudoit, J., "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics*, vol. 19, 2003.

[7] D.-W. Kim, Lee,K.H., Lee,D, "Detecting clusters of different geometrical shapes in microarray gene expression data," *Bioinformatics*, vol. Advanced Access Published, pp. 1-11, 2005.

[8] L. Liebovitch, and Toth, T., "A Fast algorithm to determine fractal dimensions by box counting," *Physics Letters*, vol. 141A, 1989.

[9] G. Mizuguchi, Shen, X., Landry J., "ATP-driven exchange of histone h2az variant catalyzed by swvl1 chromatin remodelling complex," *Science*, vol. 303, pp. 343-348, 2004.