An Information Theory Approach for Validating Clusters in Microarray Data

Sudhakar Jonnalagadda[1] and Rajagopalan Srinivasan[1]*

[1]Department of Chemical and Biomolecular Engineering, National University of Singapore,
10 Kent Ridge Crescent, Singapore 119260

ABSTRACT

Cluster validation is commonly used for evaluating the quality of partition produced by any clustering algorithm. In this paper, we present a novel method to assess the quality of clustering in gene expression data. In contrast to methods which are totally based on intra- and inter-cluster distances, our approach considers the dynamics and rearrangement of elements when a new cluster is introduced. Cluster quality is measured based on information change and the partition with the highest total information is selected. We illustrate the efficacy of the proposed method using two microarray datasets and two artificial datasets and discuss the advantages and limitations.

Key words: Cluster validation, Gene Expression, Information Theory.

*__Contact__: chergs@nus.edu.sg

Supplementary information available at http://cheed.nus.edu.sg/~chergs/ismb2004.htm

INTRODUCTION

DNA microarray technology facilitating the simultaneous monitoring of the expression levels of thousands of genes. Several clustering algorithms have been applied on gene expression data to identify the groups of co-expressed genes and correlated samples [Iyer et al (1999), Sharan et al. (2003)]. Unsupervised classification calls for the specification of number of clusters which is rarely known to the user. Various methods have been proposed [Bezdek, J.C., Pal, N.R. (1998) Dunn, J. (1974)] to resolve this, but very few have been applied to gene expression data. These indices evaluate different partitions independently without considering the relation between one partition to other. Besides that, these indices give higher importance for larger inter cluster distances which limits then to apply to the data consisting clusters which are close to each other. Here we propose a method which overcomes this and predicts optimal clusters.

MATERIALS AND METHODS

Datasets

Dataset 1: This two dimensional synthetic data consists of 300 objects grouped into 3 well defined clusters (see fig.1).

Dataset 2: This 10 dimensional artificial data consisting of 480 objects is developed by Doulaye Dembélé and Philippe Kastner (2003) http://www-igbmc.u-strasbg.fr/projets/fcm/ There are 14 clusters in this data.

Dataset 3: We used the gene expression data described by Iyer et al. (1999). This data set is available at: www.sciencemag.org/feature/data/984559.shl

Dataset 4: This gene expression data is the data used in Fig 1. of Alizadeh et al. (2000). This data set is available at: http://llmpp.nih.gov/lymphoma/data.shtml

The similarity measure used for synthetic datasets is Euclidean distance. For microarray data we used the same similarity metric used by the Authors who reported the data.

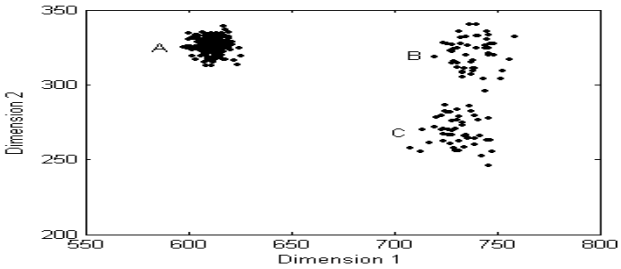

Fig. 1: Two-dimensional artificial Dataset 1 with 3 clusters

Method

Our approach is based on the evolutionary view of the clustering process. We start by considering the whole dataset as single cluster. We call this as generation 1(G_1). In each subsequent generation the data is reclustered using k-means algorithm with multiple replicates. The number of clusters N is incremented by one in each generation, i.e. k = N in generation G_N. The behavior of elements during the evolution from G_i to G_{i+1} provides the basis for evaluation of quality of the cluster. This evolutionary process is schematically showed in Fig. 2.

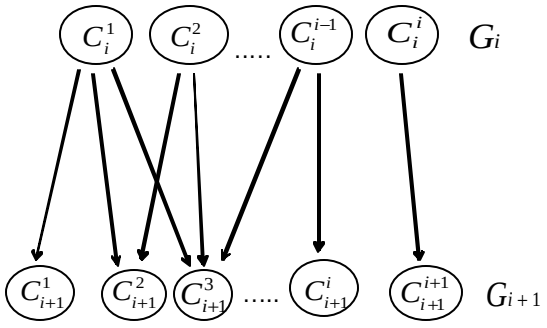

Fig 2: Migration of elements during evolution. A few clusters in G_i continue as single clusters in G_{i+1} while others disassociate or undergo leakage.

Let C_i^j be the j^{th} cluster $(1 \le j \le i)$ in G_i.

Three scenarios are possible during evolution from G_i to G_{i+1}: 1. All the elements in C_i^j may continued to be clustered together as a single cluster in G_{i+1}. We call this as stagflation. 2. Members of C_i^j may migrate to a small number (≥ 2) of clusters in G_{i+1} with each recipient cluster receiving a large fraction of the elements of C_i^j. This is called as disassociation. 3. Most of the elements of C_i^j may stay together as single cluster in G_{i+1} but a few escapes to other clusters. This is termed as leakage.

A generation characterized by compact (smaller intra-cluster distance) and non-overlap (large inter-cluster distance) clusters is considered to be the preferable one. Next, we quantify this phenomena and quality of clustering using information theory.

Net Information Gain (NIG)

We measure the Net Information Gain (NIG) during the evolution from G_i to G_{i+1}. The gain or loss of information on cluster j from G_i to G_{i+1} is given by:

$$g_i = d_i \times M_i \qquad (1)$$

Where d_i is the direction (increase or decrease) and M_i is the magnitude of change in information. If the offspring of cluster j overlap, information is deemed to have been lost and d_i = -1. In contrast, if offspring are clearly separated without overlap, information is deemed to have been gained and d_i = 1. While variety of methods can be used to measure the overlap we used centroid diameter and centroid linkage for this purpose.

Centroid diameter:

$$\Delta_A = 2\left(\frac{\sum_{x \in A} d(x, \overline{v_A})}{|A|} \right) \qquad (2)$$

Centroid Linkage:

$$d_{AB} = d(\overline{v_A}, \overline{v_B}) \qquad (3)$$

Where

A, B are clusters.

$\overline{v_A}$, $\overline{v_B}$ are centroids of clusters A, B

d (a, b) measures the distance between elements a, b.

|A| is the No. of elements in A

d_j is defined as follows:

$$d_j = \begin{cases} 1 \;\; if \;\; d_{AB} \ge \tfrac{1}{2}(\Delta_A + \Delta_B) \\ -1 \;\; if \;\; d_{AB} < \tfrac{1}{2}(\Delta_A + \Delta_B) \end{cases} \qquad (4)$$

Where A, B are the offspring of cluster j.

The magnitude of information is measured using information theory.

$$M_j = -\sum_k p_k \ln p_k \qquad (5)$$

Where k is the number of offspring of cluster j and p_k is the fraction of elements migrated from cluster j to k^{th} offspring. This equation is borrowed from information theory which satisfies our requirements: M_i should be zero for stagflation, a smaller value for leakage and larger value for disassociation.

The NIG from G_i to G_{i+1} is given by

$$NIG_{i+1} = \sum_j^i g_j \qquad (6)$$

The total information content of i^{th} generation I_i is the cumulative sum of all the NIGs till that generation. The generation with the largest information content is considered to be the optimal clustering.

RESULTS AND DISCUSSION

We compare our method with silhouette (Rousseeuw, P.J. 1987) and generalized Dunn's indices (Bezdek, J.C., Pal, N.R. 1998). For the later, specifically the normalized average of the Dunn's indices is considered (Bolshakova, N.., Azuaje, F. 2003) . We have used 500 replicates for all datasets. The time taken for dataset 1, 2, 3 & 4 are 0.6, 10, 5, and 18 mins respectively. All index values are further normalized such that they fall in the range from 0 to 1. In the following, we show the three indices for each dataset. For every index, the maximum value gives the optimal number of clusters.

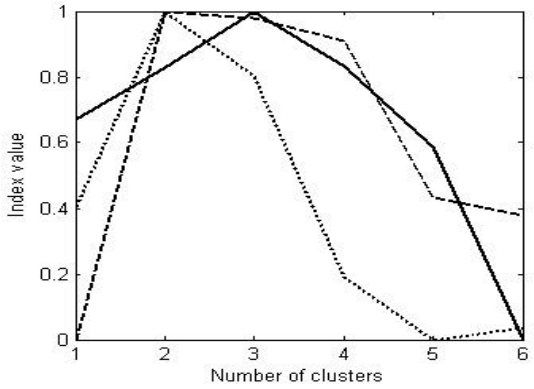

Fig 3: Results for dataset 1. Our method (solid line) predicts the correct number of clusters whereas silhouette (broken line) and Dunn's method (dotted line) predict two clusters.

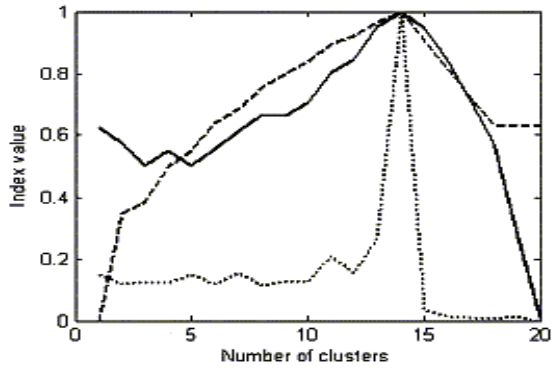

Fig 4: Results for dataset 2. All the three methods predict the correct number of clusters (14).

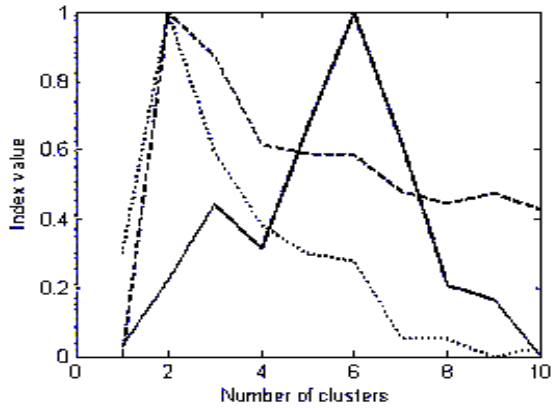

Fig 5: Results for dataset 3. The proposed method (solid line) predicts 6 clusters which agrees with Sharan et al (2003). Both Silhouette (broken line) and Dunn's method (dotted line) predict only 2 clusters.

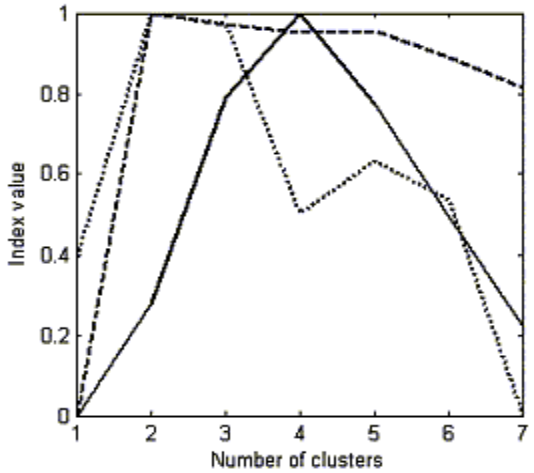

Fig 6: Results for dataset 4. Silhouette (broken line) and Dunn's method (dotted line) predict two clusters while our method (solid line) predicts 4 clusters.

As shown in Fig 3, Silhouette and Dunn's index predict 2 clusters in dataset 1 where as our method correctly predicts three – the exact number of clusters. Since clusters B and C are close to each other, at N = 3 the inter-cluster distance is very small and methods based on that cannot predict the correct number of clusters. In dataset 2, all methods predict the correct number of clusters. This shows applicability of our method to high-dimensional data. In serum data (set 3) our method estimates 6 clusters whereas silhouette and Dunn's methods predict 2 clusters. While Iyer et al. (1999) reported 10 clusters for this data; Sharan et al. (2003) showed that this data can be well described with 6 clusters. Our method agrees with this.

There are 2 classes in dataset 4 (DLBCL and non-DLBCL samples). Silhouette and Dunn's indices predict the correct number of clusters while our method predicts 4 clusters. A further exploration of the dataset reveals that there are actually three different types of samples in the non-DLBCL (normal, FL and CLL) and our results correspond to these 4.

CONCLUSIONS

A new method for evaluating quality of clusters in gene expression data is presented. This method is based on the distribution of the member of clusters from one generation of clustering to the next. Its effectiveness is assessed by applying this method to synthetic as well as real datasets. A comparison with silhouette and Dunn's methods reveal the superiority of the proposed approach. The proposed approach is also general and can be directly used to find the parameters of other clustering algorithms like Fuzzy c-means, DBSCAN, and BIRCH.

REFERENCES

Alizadeh et al., (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature., 403, 503-511.

Bezdek, J.C., Pal, N.R. (1998) Some new indexes of cluster validity, IEEE Transactions on Systems, Man Cybernet. B 28, 301–315.

Bolshakova, N.., Azuaje, F. (2003) Cluster validation techniques for genome expression data, Signal Processing 83, 825- 833.

Demb´ el ´e, D., Kastner, P. (2003) Fuzzy C-means method for clustering microarray data, Bioinformatics 19, 973-980.

Dunn, J. (1974) Well separated clusters and optimal fuzzy partitions, Journal of Cybernet. 4, 95–104.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of National Academy of Science USA 95, 14863–14868.

Iyer et al., (1999) The transcriptional program in the response of human fibroblast to serum. Science 283, 83–87.

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20, 53–65.

Sharan, R., Adi Moron-Katz, Shamir, R. (2003) CLICK and EXPANDER: a system for clustering and visualizing gene expression data, Bioinformatics 19, 1787-1799.