Tight Clustering: a method for extracting stable and tight patterns in expression profiles

George C. Tseng[1] and Wing H. Wong[2]

Keywords: microarray, cluster analysis, scattered genes, data mining

Abstract

We propose a method for clustering that produces tight and stable clusters without forcing all points into clusters. Many existing clustering algorithms have been applied in microarray data to search for gene clusters with similar expression patterns. However, none has provided a way to deal with an essential feature of array data: many genes are scattered randomly and do not belong to any of the significant biological functions (clusters) of interest. In fact, most current algorithms have to assign all genes into clusters. For many biological studies, however, we are mainly interested in the most informative, tight and stable clusters with sizes of, say, 20-60 genes for further investigation. Tight Clustering has been developed specifically to address this problem. The tightest and most stable clusters are identified in a sequential manner through an analysis of the tendency of genes to be grouped together under repeated resampling. We validated this method in the expression profiles of the Drosophila life cycle and mouse embryonic development. The result is shown to better serve biological needs in microarray analysis.

1. Methods

1.1 Algorithm A

The following algorithm is used to select candidates of tight clusters when the number of clusters *k* in the *K*-means algorithm is pre-specified. The subsampling procedure is used to create variabilities so that a pair of points stably clustered together can be distinguished from those clustered by chance.

(a) Take a random subsample X' from the original data X, say with 70% of the original sample size. Apply *K*-means with the pre-specified *k* on X' to obtain cluster centers $C(X', k)=(C_1, C_2,\ldots,C_k)$.

(b) Use the clustering result C(X',k) as a classifier to cluster the original data X according to the distances from each point to the cluster centers. Following the convention of Tibshirani et al. [1], the resulting clustering is represented by a co-membership matrix D[C(X',k), X] where $D[C(X',k), X]_{ij}$, the element of the matrix in row *i* and column *j*, takes value 1 if point *i* and *j* are in the same cluster and 0 otherwise.

(c) Repeat independent random subsampling *B* times to obtain subsamples $X^{(1)}, X^{(2)},\ldots,X^{(B)}$. The average co-membership matrix is defined as $\overline{D}$ =mean($D[C(X^{(1)},k), X],\ldots, D[C(X^{(B)},k), X]$).

(d) Search for a set of points $V=\{v_1,\ldots,v_m\}\in\{1,\ldots,n\}$ such that $\overline{D}_{v_i v_j} > 1-\alpha\ \forall i,j$, where α is a constant close to 0. Order sets with this property by size to obtain $V_{k1}, V_{k2},\ldots$. These V sets are candidates of tight clusters.

1.2 Sequential identification of tight clusters

[1] Department of Biostatistics and Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, 15260. Email: ctseng@pitt.edu

[2] Department of Statistics and Department of Biostatistics, Harvard University, Cambridge, MA, 02138. Email: wwong@hsph.harvard.edu

The following algorithm is used to identify a tight cluster that is stably chosen by consecutive *k*. We first define a similarity measure of two sets V_i and V_j to be $s(V_i, V_j)=|V_i \cap V_j|/|V_i \cup V_j|$ where $|V|$ is the size of set V.

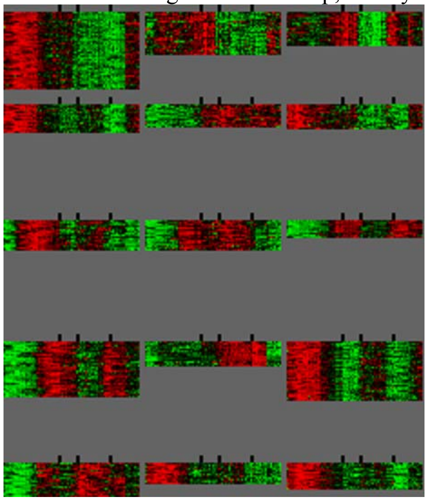
(a) Start with a suitable k_0. Apply Algorithm A on consecutive *k* starting from k_0. Choose the top 3 tightest clusters for each k, namely $\{V_{k0,1}, V_{k0,2}, V_{k0,3}\}$, $\{V_{k0+1,1}, V_{k0+1,2}, V_{k0+1,3}\},\ldots$

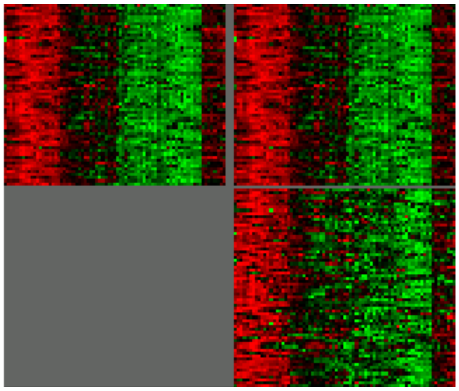
(b) Stop when $s(V_{k',l}, V_{(k'+1),m})>\beta$. Here β is a constant close to 1, $k'\geq k_0$ and $l,m\in\{1,2,3\}$. Identify $V_{(k'+1)m}$ as the tightest and most stable cluster. Remove it from the whole data.

(c) Decrease k_0 by 1 and repeat step (a) and (b) to identify the next tightest cluster.

2. Result

We applied our algorithm to a cDNA microarray data [2]. In Figure 1., the heat map [3] of 15 tight clusters when α =0.1, β =0.6, B=10 and k_0=25 is presented. The four life cycle periods are separated by black marks above the heat map. Figure 2. gives a side-by-side comparison of Tight Clustering and *K*-means algorithm. The left cluster is the first cluster identified by Tight Clustering in Figure 1. The right cluster is the corresponding cluster in *K*-means clustering when *k*=15. The two clusters have 61 common genes that were ordered and shown in the upper region. *K*-means, however, includes additional 67 genes with more variable patterns in the cluster and is likely to introduce many more false-positives. This figure shows the ability of Tight Clustering to produce tight and informative clusters for biologists to follow up, mainly because it does not require assigning all genes into clusters.





The method is further applied to a set of mouse embryonic development expression profile (data not yet published). Tight Clustering identifies a cluster of 26 genes containing seven mini chromosome maintenance (MCM) deficient genes. When using *K*-means with *k*=30, 50, 70, the resulting clusters containing these MCM genes are much larger (96, 60, 77 respectively). For *k*=100, MCM genes were distributed in two different clusters (size 31 and 15), making it harder to detect the co-regulation of the MCM genes.

3. References

[1] R. Tibshirani, G. Walther, D. Bostein, and P. O. Brown (2001). “Cluster validation by prediction strength.”, Technical report, Department of Statistics, Stanford University.

[2] M. Arbeitman, E. Furlong, F. Imam, E. Johnson, B. Baker, R. Davis. and K. White (2002). “Gene expression during the life cycle of Drosophila melanogaster.” *Science* **297**, 2270-2275.

[3] M. B. Eisen (2000). “TreeView (version 1.5).” Software download: http://rana.lbl.gov/