FMC, a Fuzzy Map Clustering algorithm for microarray data analysis

Li-Min Fu and Enzo Medico

Istitute for Cancer Research and Treatment (IRCC), University of Torino, 10060, Candiolo, Italy.

limin.fu@ircc.it
enzo.medico@ircc.it

Keywords. Fuzzy clustering, Fuzzy membership, Locally Linear Embedding (LLE), microarray

Introduction

As the microarray technology is emerging as a widely used tool to investigate gene expression and function, laboratories over the world have produced and are producing a huge amount of data, which demand advanced and specialized computational tools to process them. Clustering methods have been successfully applied to such data to reorganize the data and extract biological information from them. But the classical clustering methods [1] such as k-means and hierarchical clustering have some intrinsic limits such as the linear, pair-wise nature of the similarity metrics (which fail to highlight non-linear substructures of the data) and the univocal assignment of each gene to one cluster (which may fail to highlight cluster-to-cluster relationships) [2]. Here we introduce a novel method for clustering microarray data, named *Fuzzy Map Clustering (FMC),* which may partly overcome these limits.

Basically, the clustering process of FMC starts from identification of an initial set of clusters by calculating the "density" around each data point (object), that is, the average proximity of its *K* nearest other objects (*K* neighbours) and choosing the ones that have the highest density among all their *K* neighbors. *K* can be a fixed number of choice or the number of neighbors within a distance threshold.

Then, each object in the dataset is assigned a fuzzy membership to all the defined clusters (a vector containing a percentage of membership to all the clusters). Membership is assigned so that similar objects have similar fuzzy membership vectors. Membership assignment is optimized by measuring how the fuzzy membership vector of one object can be approximated by the vectors of its neighbors.

Finally, a process based on the merging of adjacent clusters and fuzzy membership reassignment is reiterated until the number of clusters is reduced to a fixed one decided by the operator.

Our computational experiments have shown that FMC can correctly reveal the true cluster structure of the dataset if such structure exists, even if the clusters contained in the dataset have arbitrary shape. And perhaps the basic idea underlying FMC points out a new way to develop novel clustering methods with good mathematical foundation.

Detailed description of the clustering procedure

Mathematically speaking, FMC maps data points from their feature space to a fuzzy membership space in such a way that the local structure of the original dataset is preserved, and essentially the dataset is clustered according to its lower dimensional structure.

1. Identification of the initial set of clusters

For each object, its "density" is approximated by *K/d*, where *d* is the average distance between that object and its *K* neighbors. Any distance metric can be used (we used Euclidean and Pearson correlation) to estimate the density. Then the initial set of clusters is determined as the objects with local maximum "density", which means that its density is higher than all its neighbors. These objects with local maximum densities are called *cluster supports*, which are the prototypes of clusters.

2. Neighborhood Approximation of Fuzzy Memberships

Membership assignment is optimized by measuring how the fuzzy membership vector of one object can be linearly approximated by the vectors of its neighbors. This idea of linearly approximating the fuzzy memberships of neighboring objects resembles that of linearly approximating the lower dimensional embedding coordinates in a non-linear embedding method call Locally Linear Embedding (LLE) [3].
A simple iterative procedure is applied to find the optimal fuzzy memberships which the overall deviation of the fuzzy membership of one object from a linear combination of its neighbors' fuzzy memberships. In each step of this iterative procedure, the fuzzy memberships are updated as

$\mathbf{p}^{t+1}(\mathbf{x}) = \sum_{\mathbf{y}\in G(\mathbf{x})} w_{\mathbf{x},\mathbf{y}} \mathbf{p}^{t}(\mathbf{y})$, where the sum is over the nearest neighbors of object x, and $w_{x,y}$ is the linear combination coefficient with $\sum_{\mathbf{y}\in G(\mathbf{x})} w_{\mathbf{x},\mathbf{y}} = 1$.

3. Merging similar clusters

After the first two steps, we get an initial set of clusters and fuzzy memberships. There will be some objects with high membership degrees in more than one cluster. Two clusters sharing a large number of such objects are very close to each other, they must be merged. However, to choose which two clusters to merge, simply counting the number of their common objects will not work. So we proposed an alternative way to merge clusters.
We defined *Dentropy* of a cluster as the average Shannon entropy of one cluster weighted by the density of each object. High Dentropy of a cluster indicates that the set of objects assigned to that cluster could also belong to one or more other clusters. So we identify the cluster with the highest Dentropy, and merge it to its nearest neighbor. The nearest neighbor cluster is found by calculating the density-weighted fuzzy membership centroid of the cluster to be merged, and choose the one with second largest value. The merging of two clusters is done not by putting objects from the two clusters together, but by adding up their fuzzy membership degrees. The cluster merging is stopped when a pre-defined number of clusters is reached.

4. Computational experiments

We have applied FMC to some synthetic datasets and to real gene expression datasets. In the synthetic, 2D dataset experiments, where the results can be clearly visualized, FMC shows much better performance than the classical clustering methods, since it can identify most clusters correctly even when the clusters are irregular. In higher dimension dataset, the most effective way to compare the performance of FMC with that of classical methods has still to be defined. For example, by analyzing the mean variance of each cluster, which indicates how objects belonging to the same cluster are really similar to each other, we found that FMC is at least as good as classical clustering methods. Probably more complex ways to estimate how the FMC approach yields more informative clusters need to be implemented. However the basic idea of FMC points out a new way to develop novel clustering methods with good mathematical foundation.

References

[1] A.K. Jain, M.N. Murty and P.J. Flynn, Data Clustering: A Review. *ACM Computing Surveys*, Vol.31, No.3,1999.
[2] Yidong Chen, Michael L. Bittner and Edward R. Dougherty, Issues associated with microarray data analysis and integration. *Nature Genetics,*1999.
Information supplementary to article by Michael Bittner, Jeffrey Trent and Paul Meltzer (*Nature Genet.* **22**, 213–215; 1999)
[3] Sam T. Roweis and Lawrence K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323-2326, 2000.