A New Clustering Strategy with Stochastic Merging and Removing Based on Kernel Functions

Huimin Geng
Department of Pathology and Microbiology
University of Nebraska Medical Center
hgeng@mail.unomaha.edu

Hesham Ali
Department of Computer Science
University of Nebraska at Omaha
hali@mail.unomaha.edu

Abstract

With hierarchical clustering methods, divisions or fusions, once made, are irrevocable. As a result, when two elements in a bottom-up algorithm are assigned to one cluster, they cannot subsequently be separated. Also, when a top-down algorithm separates two elements, they can't be rejoined. Such greedy property may lead to premature convergence and consequently lead to a clustering that is far from optimal. To overcome this problem, we propose a new Stochastic Message Passing Clustering (SMPC) method based on the Message Passing Clustering (MPC) algorithm introduced in our earlier work [1]. SMPC, as a generalized version of MPC, extends the clustering algorithm from a deterministic process to a stochastic process, adding two major advantages. First, in deciding the merging cluster pair, the influences of all clusters are quantified by probabilities, estimated by kernel functions based on their relative distances. Secondly, clustering can be undone to improve the clustering performance when the algorithm detects elements which don't have good probabilities inside the cluster and moves them outside. The test results on colon cancer gene-expression data show that SMPC performs better than the deterministic MPC or hierarchical clustering method.

1. Introduction

Clustering is a fundamental technique which has numerous applications in expression-data analysis, phylogenetic tree building and regulatory network reconstruction. In this paper, we propose a generalized version of Message Passing Clustering (MPC) [1], MPC with stochastic merging (SMPC), which extends the clustering algorithm from a deterministic process to a stochastic process to include the following features that the deterministic algorithms cannot offer:

1). *Breaking ties*: Ties often occur when the distances are integers as in the case of protein interaction data. The ties in HC can be a serious problem because the final clustering solution depends on the index of the nodes. A better mechanism would be to choose the merging neighbor by coin tosses because this is independent of the arbitrary index of nodes.

2). *Ensemble influence*: By ensemble we mean the set of all nodes (clusters). We believe that the ensemble influence should be considered when deciding merging cluster pairs, because, besides the first nearest neighbor, the second, the third, etc., may also be good candidates if there are no significant differences in terms of distances between them and the target cluster. We use kernel functions generated by each cluster to estimate the ensemble probability distribution, and we can use a random switch to determine which cluster is to be selected based on the probabilities.

3). *Undoing clusters*: The probability estimation technique (similar to the stochastic merging mechanism) can be applied to removing outliers if their probabilities staying inside are no longer satisfactory, i.e. below a certain threshold.

2. Methods

From the definition of a probability density, if the random variable X has density f, then $f(x)=\lim_{w\to 0}\frac{1}{2w}P(x-w<X<x+w)$. For any given w, where w is known as the *bandwidth parameter,* a naïve estimator of $P(x-w<X<x+w)$ is the proportion of the observations X_1, X_2, ..., X_n falling in the interval

$(x-w, x+w)$. This could lead naturally to the kernel estimator defined as $\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}K(x,X_i)$, where $\hat{f}(x)$ is a probability density estimator at point *x*, and *K* is known as the *kernel function* which is a function of *x* and satisfies the condition $\int_{-\infty}^{\infty}K(x,X_i)dx=1$. R*ectangular* kernel function is a simple one, defined as: $W(x,X_i)=\begin{cases}\frac{1}{2w} & \text{if } |x-X_i|<w \\ 0 & \text{otherwise}\end{cases}$, and *Gaussian* kernel function is one of the most popular, defined as: $K(x,X_i)=\frac{1}{w\sqrt{2\pi}}e^{-\frac{1}{2w^2}(x-X_i)^2}$. SMPC can be reduced to MPC if the selected kernel function is rectangular kernel function and the selected bandwidth parameter is the minimum distance among distances between the target cluster and other clusters.

The kernel estimator tells us that an instance at X_i generates a kernel function *K(x, X_i)* that assigns a probability to each point *x* in the space; the density estimate as a whole is just the normalized sum of all of the little kernel functions. Figure 1 gives an example of the graphs of probability density estimates. The bandwidth parameter, *w*, affects the kernel density estimates. Figure 2 shows this effect with an example of two-dimensional data. Making *w* too small results in a very spiky estimate (Figure 2a), while making it too large results in losing the structure altogether (Figure 2d). In Figure 2 (b and c), a medium value of *w* gives a very good reconstruction.

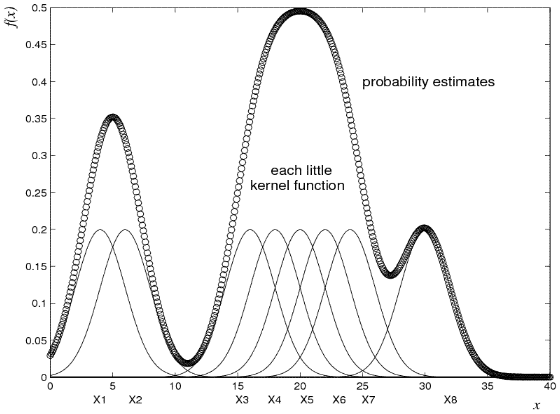

Figure 1. Probability density estimates based on little Gaussian kernel functions

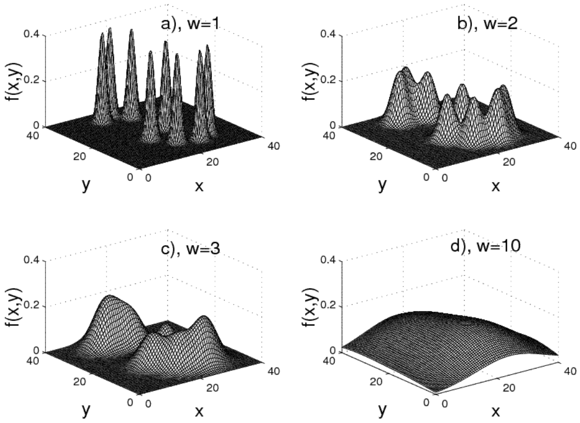

Figure 2. Kernel density estimates using Gaussian kernels with a) *w*=1, b) *w*=2, c) *w*=3, d) *w*=10.

3. Results

We use a set of artificial data to test SMPC. Clustering results obtained by SMPC give more possible solutions, providing more points of view about the relationship among data, as compare to deterministic methods. Also, we apply SMPC to colon cancer microarray data [2] to see the potential improvement of the clustering results, which shows SMPC improves the classification precision from 62.9% to 85.5% when classify the normal and tumor samples, as compared to MPC. SMPC is especially useful for identifying the relationship in the data where many distances are identical as in the protein interaction data. The supplementary information about the data and the results is available at http://bioinformatics.ist.unomaha.edu/~hgeng/.

4. Conclusion

We introduce the method of SMPC and its advantages over the deterministic MPC algorithms in clustering gene microarray data. SMPC generalizes MPC by involving stochastic processes and it can be reduced to MPC if we choose the particular kernel function and bandwidth parameter to estimate the probability. The analysis of the real colon cancer data with SMPC shows higher classification accuracy as compared to MPC.

References

1) Geng H., Bastola D. and Ali H. A New Approach to Clustering Biological Data Using Message Passing. In *Proceedings of 2004 IEEE Computer Society Bioinformatics Conference (CSB 2004)*, pp. 493-494.

2) Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D. and Levine A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *PNAS*, 1999, vol. 96, pp. 6745–6750.