새로운 모형기반 군집분석 알고리즘[1]

박정수[2], 황현식[3]

Abstract

A new model-based clustering algorithm is proposed. The idea starts from the assumption that observations are realizations of Gaussian processes and so are correlated. With a special covariance structure, the posterior probability that an observation belongs to each cluster is computed using the ECM algorithm. A preliminary result of small-scale simulation study is given to compare with the k-means clustering algorithms.

Key words : Gaussian processes, covariance function, posterior probability, ECM algorithm, k-means algorithm

1. Introduction

Cluster analysis is the identification of groups of observations that are cohesive and separated from other groups. Interests on the clustering method have been increased recently due to the emergence of several new areas of applications. These include character recognition, tissue segmentation, classification of astronomical data, image analysis and data-mining (Fraley and Raftery, 2002).

One widely used class of methods involves hierarchical agglomerative clustering, in which two groups chosen to optimize some criterion are merged at each stage of the algorithm. Another common class of methods (for example, the k-means algorithm) is based on iterative relocation, in which data points are moved from one group to another until there is no further improvement in some criterion.

Banfield and Raftery(1993) introduced a new approach (so-called model-based clustering) based on parsimonious geometric modeling of the within-group covariance matrices in a mixture of multivariate normal distributions, using hierarchical agglomeration and iterative relocation. The well-known EM(Expectation and Maximization) algorithm is used for computing the probabilities that an observation belongs to each cluster.

We propose a model-based clustering technique using Gaussian processes which is a new clustering method for multivariate data. In general, the Gaussian process we are applying has been used in nonlinear regression, in discriminant analysis, in spatial statistics (Cressie, 1991), and in design and analysis of computer experiments (Park, 1994).

[1] This research was supported by Korea Research Foundation Grant funded by the Korean Government(MOEHRD) (KRF-2002-041-C00048).

[2] 전남대학교 통계학과 교수, jspark@chonnam.ac.kr

[3] 통계청 통계연수부 교수

2. Model-based Clustering using Gaussian Processes

The idea starts from the assumption that observations are realizations of Gaussian processes and so are correlated. The correlation between two observations $\underline{t}=(t_1,t_2,\cdots,t_d)$ and $\underline{u}=(u_1,u_2,\cdots,u_d)$ is specified by the following covariance function:

$$V(t,u)=\sigma_z^{\,2}\exp\{-\sum_{j=1}^{d}\theta_j\mid t_j-u_j\mid^{p}\},\quad 0<\theta_j,\quad 0<p\leqq 2.$$

In this preliminary work, the independence between variables is assumed for simplicity. With the above model assumptions, we compute the posterior probability that an observation belongs to each cluster using the ECM(Expectation, Clustering, and Maximization) algorithm which is a modification of EM algorithm. The maximum likelihood estimates of θ_j, the Quasi-Newton optimization routine and Cholesky decomposition are used in the following algorithm.

[Outline of Proposed ECM algorithm]

1) Determining the number of cluster
2) Partitioning the initial cluster by random or ordered method
3) Randomly determining the initial mean of each cluster and initial values of covariance function's parameter
4) Compute the probability that an observation belongs to each cluster (E step)
5) Classifying the observation to the cluster with maximum probability (C step)
6) Compute mean of cluster again using the relocated clusters (M step)
7) Repeat 4) to 6) until no more relocating process
8) For determining number of cluster, compute BIC (Baysian Information Criteria)

3. Simulation study

A small scale simulation study is presented for the evaluation of the performance and usefulness of our suggested clustering algorithm. The performance of this algorithm is almost same with the k-means algorithm. In case of the mean differences are relatively small, suggested algorithm works well to represent the connectedness of observations.

The following table shows the simulation setting of the correlated data with relatively small mean difference.

	initial cluster(n)	variable 1	variable 2	variable 3
ρ=7.0	cluster 1(25)	μ = 2	μ = 2	μ = 2
	cluster 2(25)	μ = 5	μ = 5	μ = 5

The following figures (in next page) show how the clusters differ between our method and the k-means algorithm.

4. Summary and Discussion

The clustering method proposed here turned out to be good compared to the k-menas algorithm for the case of correlated observations. When the mean differences are relatively small, our algorithm works well to represent the connectedness of observations. When the variables are composed of factor scores and principle component scores in which variables are independent, our method is directly applicable while Banfield-Raftery's method is not.

One disadvantage is that the speed of convergence can be slow because of the intrinsic problem of EM depending on how to choose a initial cluster. To overcome this problem and to select the initial clusters, one can proceed hierarchical clustering method before applying EM algorithm. We think the propose method can be applied many research areas (Papageorgiou, et.al, 2001, Li, et.al, 2005, and Reverter, et.al, 2003, for example).

References

Banfield, J. D. and Raftery, A. E. (1993). Model-Based Gaussian and non-Gaussian Clustering. *Biometrics*, 17, 803-821.

Cressie, N. A. (1991). *Statistics for Spatial data.* Wiley, New York.

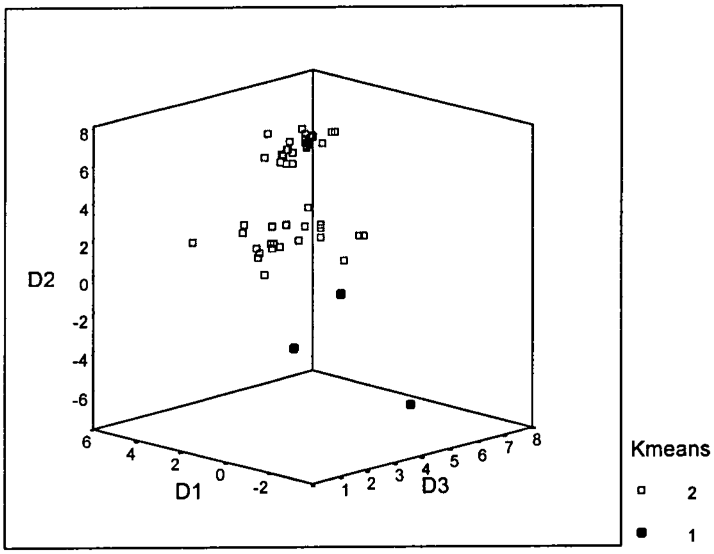
Fraley, C, and Raftery, AE. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of Amer Stat Assoc* 97 (458): 611-631.

Li, QH, Fraley, C, Bumgarner, RE, et.al.(2005). Donuts, scratches and blanks: robust model-based segmentation of microarray images, *Bioinformatics* 21 (12): 2875-2882.

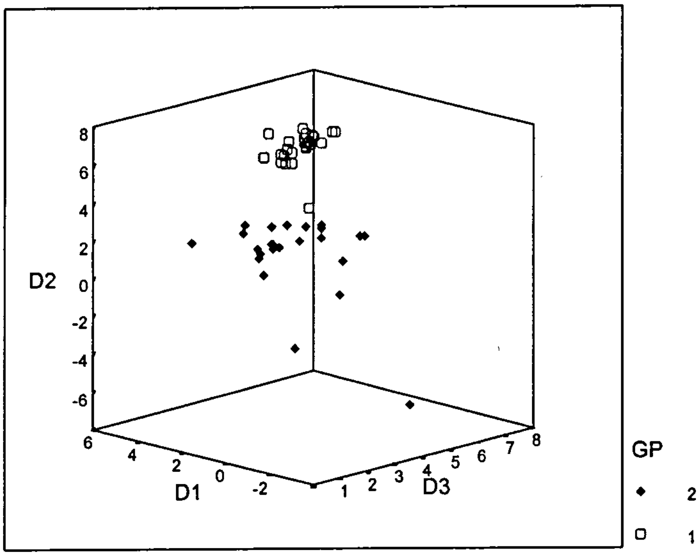
Papageorgiou, I, Baxter, MJ, Cau, MA (2001). Model-based cluster analysis of artefact compositional data, *Archaeometry* 43, Part 4: 571-588.

Park, J. S. (1994). Optimal Latin-hypercube designs for computer experiments. *Journal of Statistical Planning and Inference*, 39, 95-111.

Reverter, A, Byrne, KA, Bruce, HL, et al. (2003). A mixture model-based cluster analysis of DNA microarray gene expression data on Brahman and Brahman composite steers fed high-, medium-, and low-quality diets, *Journal of Animal Science* 81 (8): 1900-1910.



Cluster analysis using k-means algorithm



Cluster analysis using GP method