Sample Based Algorithm for $k$-Spatial Medians Clustering

Seohoon Jin$^1$ · Byoung Cheol Jung$^2$

$^1$Department of Informational Statistics, Korea University
$^2$Department of Statistics, University of Seoul

Abstract

As an alternative to the $k$-means clustering the $k$-spatial medians clustering has many good points because of advantages of spatial median. However, it has not been used a lot since it needs heavy computation. If the number of objects and the number of variables are large the computation time problem is getting serious. In this study we propose fast algorithm for the $k$-spatial medians clustering. Practical applicability of the algorithm is shown with some numerical studies.

Keywords: Cluster analysis, partitioning method, $k$-spatial medians clustering.

1. Introduction

Cluster analysis is a branch of statistics that, in the past three decades, has been intensely studied and successively applied to many applications. There are two main categories of clustering algorithms: hierarchical method and partitioning method. The most widely used partitioning method is the $k$-means clustering which minimizes within-cluster sum of squares. However, the $k$-means clustering is sensitive to outliers because an object with an extremely large value may substantially distort the distribution of data. This effect is particularly exacerbated due to the use of the square error function. To avoid the influence of outliers, actual objects can be picked to represent the clusters instead of taking the mean value of the objects in a cluster as a reference point. The $k$-mediods method, which uses actual objects in a cluster as a reference point, is performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point (Kaufman and Rousseuw, 1990). However, the medoid is also not fully free from effect of outliers.

The $k$-spatial medians clustering was proposed as an alternative to the $k$-means clustering in some previous studies such as Butler (1986), Butler (1988), Jhun (1986) and Jhun and Jin (2000). For verification of superiority of the $k$-spatial medians clustering let us consider an example. Suppose...
we measure two variables for each of 11 objects. The data consist of two the same structure clusters and one outlier (the point (7,3)). Figure 1.1, 1.2 and 1.3 show the results of \( k \)-means, \( k \)-medoids and \( k \)-spatial medians clustering sequentially. Each point is plotted by its cluster identification number. For the \( k \)-means clustering, the outlier perturbs the genuine cluster structure. The \( k \)-medoids clustering also doesn’t overcome the outlier’s influence. However, the \( k \)-spatial medians clustering gives the most preferable result under existence of an outlier. Since the spatial median is less sensitive to outliers, the centroid of each cluster is affected little by the outlier.

Even though the \( k \)-spatial medians clustering clearly has advantages under the situation of being outliers and special cluster structures, it has not been used a lot because of its computational difficulties. This study is about fast algorithm of the \( k \)-spatial medians clustering. We tried to improve applicability of the \( k \)-spatial medians clustering by modification of the clustering algorithm.