A Clustering Algorithm for Similar Pattern

Young-Hyun Ko
POSTECH, Industrial Engineering
San 31 Hyoja-dong
Pohang 790-784, Korea
notime@postech.ac.kr

Hyeseon Lee, Si-Young Park, Chi-Hyuck Jun
POSTECH, Industrial Engineering
San 31 Hyoja-dong
Pohang 790-784, Korea
{hyelee, parkc0, chjun}@postech.ac.kr

1. Introduction

Many algorithms for clustering such as K-means clustering method depend on the initial clustering prototype and the specified number of subgroups. Those clustering algorithms assume that every observation is assigned to one of subgroups, but it is unreasonable when the purpose is to find groups with similar patterns being distinguished from ordinary patterns. This study proposes the method for carrying out pattern clustering without *a priori* assumption on the number of clusters in the data set. This algorithm finds patterns sequentially and stops when there is no specific pattern since the unassigned observations do not have useful information.

2. Theory and algorithm

In order to cluster patterns without *a priori* assumptions on the number of clusters in the data set, we modified clustering problem to the problem of fining specific patterns. This algorithm consists of following procedure.

Procedure

Step 1 : Generate a new pattern using correlation matrix.
Step 2 : Obtain the centroid of the pattern.
Step 3 : Assign observations to the pattern.
Step 4 : Remove the assigned observations from the data set if condition meet.
Step 5 : If there is another specific pattern, go to step 1 and stop, otherwise.

In step 1, use a correlation matrix to generate a new initial pattern c_1 and select the one with the maximum correlation coefficients.

$\arg\max(sum(corr))$, where $corr$ is correlation matrix(N*N)

In step 2, use a fuzzy concept of membership function, and iterate this step until c_k is converged.

$$c_k = \frac{\sum_{i=1}^{N} u_i^m x_i}{\sum_{i=1}^{N} u_i^m}$$

where $u_i = corr(c_{k-1}, x_i)$ and x_i is the i th observation, m is weighting exponent.

In step 3, assign highly correlated observation with the converged c sequentially if $correlation(c, x_i)$ is greater than [mean(corr) - 2 s] of previously assigned observations in this cluster where s is the standard deviation of the mean(corr). In step 4, remove the assigned observation from data set. In step 5, investigate the remained observation if there is a specific pattern by the following criterion. If $\max(sum(corr)) >$ spec is go to step1, otherwise stop.

3. Simulation and Discussion

Here is an example that shows this algorithm to find successfully the patterns in the data set and assign observations to each pattern. We generated the data with 140 observations and 60 variables (x_{ij}, i=1, ..,140; j=1, ..,60) as follows: $x_{ij} = b_{ij}+z_{ij}$, where for $1 \le i \le 10$ b_{ij} = -1 if $j \le 30$, b_{ij} = +1 if j>30, for $11 \le i \le 20$, b_{ij} = -1 if j is odd, b_{ij}=+1 if j is even, for $21 \le i \le 100$ b_{ij} = 0, and $z_{ij} \sim N(0,1)$. And for $101 \le i \le 120$, $x_{ij} = \sin(j^2) + N(0,1/\sqrt{10})$ and for $121 \le i \le 140$, $x_{ij} = \cos(j) + N(0,1/\sqrt{10})$. This algorithm correctly finds four patterns even with noise in the data set, and correctly assigned observations to the its pattern except observations 15 and 18 which have smaller correlation with the center. Figure 1 shows the result of the clustering with this data set.

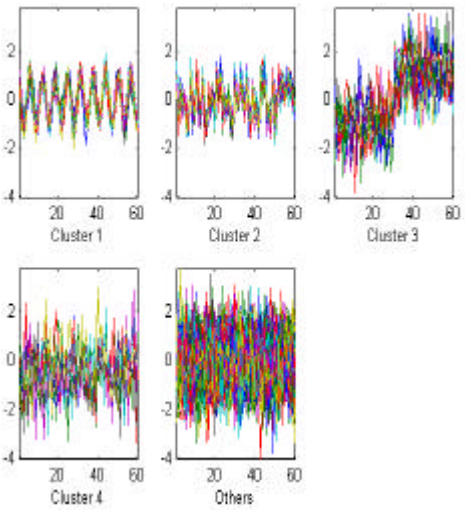

Figure 1. the result of clustering

REFERENCE

[1] I. Gath, A.B. Geva, "Unsupervised Optimal Fuzzy Clustering", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 2, no 7, 1989, pp. 773-781.

[2] C.A. Rosen, C.W. Stork, Pattern Classification, *John Wiley & Sons, Inc.*, second edition, 2001.

[3] T. Hastie, R. Tibshirani *et al.*, "Gene Shaving as a method for identifying district sets of genes with similar expression patterns", *Genome Biology*, vol 2, no 2, 2000, 1-21.