# A New Measure of Validity in Cluster Analysis

Seohoon Jin
*eCRM Team, OpenTide Korea*
*739-1, Hannam-Dong, Yongsan-Gu*
*Seoul, 140-210, Korea*
*shjin@opentide.com*

Myoungshic Jhun
*Department of Statistics, Korea University*
*Anam-Dong, Sungbuk-Ku*
*Seoul, 136-701, Korea*
*jhun@mail.korea.ac.kr*

## 1. Introduction

The "validity" in cluster analysis is used as the problem of an investigation on whether clusters resulting from a clustering procedure are genuine. Davies and Bouldin (1979) suggested a measure of separation degree, which is based on within-cluster dispersions and distances between centroids, in assessing the validity of clusters. We examine the limitation of the separation measure and propose a different one for evaluation of the validity.

## 2. Validity of clusters

Davies and Bouldin (1979) suggested a separation measure of the $k$-cluster partition as following. Suppose that independent multi-dimensional objects $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n$ are made on a distribution $F$ on $R^p$. These objects are partitioned into $k$ clusters $C_1, C_2, \cdots, C_k$ by using a certain clustering method. The $i$-th cluster $C_i$ is composed of $n_i$ objects $(1 \leq i \leq k)$. Let $S_i = \{\frac{1}{n_i} \sum_{\underline{x}_j \in C_i} ||\underline{x}_j - \underline{m}_i||^q\}^{1/q}$ and $M_{ij} = \{\sum_{l=1}^{p} ||m_{il} - m_{jl}||^t\}^{1/t}$ where $\underline{m}_i = (m_{i1}, \cdots, m_{ip})$ is the $i$-th cluster centroid. Let $R_{ij} = \frac{S_i + S_j}{M_{ij}}$ be the similarity measure between the cluster $i$ and the cluster $j$. The suggested cluster separation measure is $\overline{R} = \frac{1}{k} \sum_{i=1}^{k} R_i$ where $R_i = \max_j R_{ij}$, $(i \neq j)$. The best choice of the partition, then, will minimize this average separation degree. However, the separation measure $\overline{R}$ can lead us to improper conclusions for non-spherical structure. Let us consider the following data set, which consists of 60 objects where each of 30 points forms an elongated structure, for $t = 2$ and $q = 1$. Figure 1 shows the results of $k$-spatial medians clustering (Jhun, 1986) for $k = 2$ and $k = 3$. The value of $\overline{R}$ is computed as 1.035 for $k = 2$ and

0.708 for $k = 3$. The result makes us judge that the partition composed of three clusters has higher validity than that composed of two clusters. However, we can be aware of inadequacy of three clusters from Figure 1.

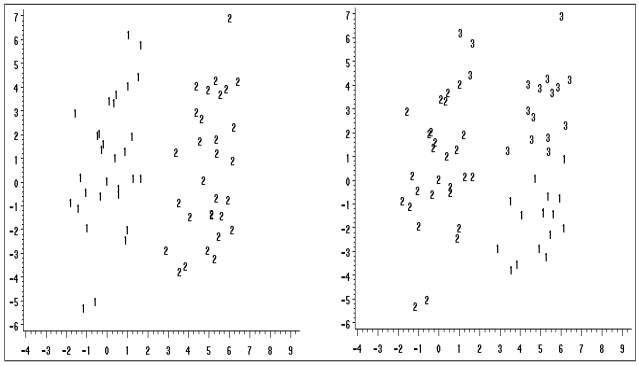

**Figure 1** Results of $k(= 2, 3)$-spatial medians clustering

Now, we suggest an alternative measure of separation degree of clusters. The separation measure $Q_i$ of $i$-th cluster can be computed as following steps. First, compute Euclidean distance $\delta_{jl}^{(i)} = ||\underline{x}_j - \underline{x}_l||$ for every $\underline{x}_j, \underline{x}_l \in C_i$, and calculate the average of $\binom{n_i}{2}$ distances $\overline{\delta}_i = \sum_{j<l} \delta_{jl}^{(i)} / \binom{n_i}{2}$. Next, among the objects except for the members of the $i$-th cluster, find the nearest object $y_{(i)}$ from the $i$-th cluster centroid and compute distances $d_1^{(i)}, d_2^{(i)}, \cdots, d_{n_i}^{(i)}$ from this object $y_{(i)}$ to every object which belongs to the $i$-th cluster (i.e. $d_j^{(i)} = ||\underline{x}_j - \underline{y}_{(i)}||$, $\underline{x}_j \in C_i$). If $\overline{\delta}_i$, which represents dispersions of the $i$-th cluster, is sufficiently small comparing to $d_1^{(i)}, d_2^{(i)}, \cdots, d_{n_i}^{(i)}$, then the $i$-th cluster is well separated from other objects or clusters. Thus, we propose a separation measure of $i$-th cluster as $Q_i = \sum_{j=1}^{n_i} I(d_j^{(i)} < \overline{\delta}_i)/n_i$. In other words, the separation degree of the $i$-th cluster is measured by counting the number of $d_j^{(i)}$, $(j = 1, \cdots n_i)$ which is less than $\overline{\delta}_i$. When the value $Q_i$ is small, the $i$-th cluster is well separated from other objects and vice versa. To incorporate the size of the clusters, we propose a weighted average of $Q_i$'s $\tilde{Q} = \frac{\sum_{i=1}^{k} n_i Q_i}{n}$ as a separation measure of the whole partition. The values of $Q_1$ and $Q_2$ from the result of $k$-spatial medians clustering for $k = 2$ are computed as $Q_1 = 0.233$, $Q_2 = 0.200$ and $\tilde{Q} = 0.217$. Meanwhile, it gives $Q_1 = 0.250$, $Q_2 = 0.185$, $Q_3 = 0.470$ and $\tilde{Q} = 0.283$ for $k = 3$. It seems that $k$-spatial medians clustering for $k = 2$ gives a relatively better result, comparing with $k = 3$.

## REFERENCES


Davies, D. L. & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on pattern analysis and machine intelligence PAMI-1*, 224-227.

Jhun, M. (1986). Bootstrap Method for $k$-Spatial Medians. *Jouranl of the Korean Statistical Society*, **15**, 1-8.