Separability index for cluster analysis

Andrzej Sokołowski[1], Sabina Denkowska[2], Kamil Fijorek[3]

Abstract

A new separability index for groups obtained in cluster analysis has been proposed in the paper. It is assumed that the number of groups and object assignments are given. The main idea is based on the squared Euclidean distance between each object and the closest one belonging to different group. Sum of these distances should be normalized, e.g. by within group sum of squares. The proposed measure can be use also for overlapping or fuzzy clusters.

The results of simulation studies under different models of separability are given. The proposed measure is compared to Calinski-Harabasz, Krzanowski-Lai and Rousseeuw indexes.

Keywords: *cluster analysis, cluster validity indices, simulation studies*

JEL Classification: C38, C15

1 Introduction

There are many measures of the quality of clustering (cluster validity). Migdał-Najman (2011) gave comprehensive bibliographic survey of this subject (see also: Everitt, 1995; Arabie, Hubert, De Soete, 1998; Everitt et al., 2011). Her paper has 178 references and she proposed some classifications of cluster validity measures. In this context, it can look strange that we propose another measure. The main idea is based on the squared Euclidean distance between each object and the closest one belonging to different group.

The proposed measure is compared to three popular measures: Caliński-Harabasz (CH), Krzanowski-Lai (KL) and Rousseeuw (S) indexes.

2 The new separability index AS

The new index AS is based on the squared Euclidean distance between each data point $\mathbf{x}$ and the closest one belonging to different group, $S(\mathbf{x})$. Sum of these distances SS is normalized by $WGSS$ which denotes the within-group sum of squares (it is a sum of squared Euclidean distances from each point to the center of its cluster).

[1] Corresponding author: Cracow University of Economics, Department of Statistics, 27 Rakowicka St., 31-150 Cracow, Poland, e-mail: sokolows@uek.krakow.pl.
[2] Cracow University of Economics, Department of Statistics, 27 Rakowicka St., 31-150 Cracow, Poland, e-mail: sabina.denkowska@uek.krakow.pl.
[3] Cracow University of Economics, Department of Statistics, 27 Rakowicka St., 31-150 Cracow, Poland, e-mail: kamil.fijorek@uek.krakow.pl.

Let X denotes a set of n observations (which are *v*-dimensional points), divided into k separated clusters C_i such that: $X = \bigcup_{i=1}^{k} C_i$. Let $\boldsymbol{d}(\cdot,\cdot)$ denotes squared Euclidean distance between two vectors.

The AS index is defined as:

$$\mathrm{AS} = \frac{SS}{WGSS}. \tag{1}$$

The SS in the numerator of index AS is the sum of $S(\mathbf{x})$:

$$SS = \sum_{\mathbf{x} \in X} S(\mathbf{x}) \tag{2}$$

where $S(\mathbf{x})$ is defined as:

$$S(\mathbf{x}) = \min_{\mathbf{y} \notin C(\mathbf{x})} \boldsymbol{d}(\mathbf{x},\mathbf{y}) \tag{3}$$

where $\mathbf{x},\mathbf{y}$ are the data points $(\mathbf{x},\mathbf{y} \in X)$ and $C(\mathbf{x})$ is a cluster to which $\mathbf{x}$ belongs to.

The overall within-cluster variance $WGSS$ is defined as:

$$WGSS = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \boldsymbol{d}(\mathbf{x},\mathbf{m}_i) \tag{4}$$

where $\mathbf{x}$ is a data point $(\mathbf{x} \in X)$, $\mathbf{m}_i$ is the centroid of i-th cluster.

The optimal number of clusters is the solution with the highest AS index value.

3 Some other criteria for determining the number of groups in a data set

Let *k* denote the number of groups (clusters) in a data set. Assume that *n* denotes the number of cases (number of rows in a data set). Each case is described using *v* quantitative variables and each case is viewed as a *v*-dimensional point in the Euclidean space. Consequently for each pair of points a distance can be calculated.

3.1 Caliński-Harabasz criterion (1974)

Caliński and Harabasz (CH) in their 1974 paper introduced a criterion "to select the value of *k* at which the final partition appears to be best". They call this criterion the *VRC* (*Variance Ratio Criterion*) and define it as:

$$VRC = \frac{BGSS/k\text{-}1}{WGSS/n\text{-}k}. \tag{5}$$

CH calculate distances between data-points using squared Euclidean metric. The *WGSS* denotes the within-group sum of squares and it is a sum of distances from each point to the center of a cluster it belongs to. The *BGSS* denotes the between-group sum of squares and it is a sum of distances, weighted by the cluster sizes, from each cluster centroid to the overall centroid.

When the clusters are well separated the *BGSS* will tend to be large and the *WGSS* small causing the *VRC* to be large. Consequently the *k* that gives the largest value of *VRC* is preferable. However in case of multiple local maxima the one with the smallest value of *k* should be chosen.

CH point out that the *VRC* is a heuristic that has no evident probabilistic foundation. However the *VRC* has some interesting properties. (1) When all points in a data set are equally distant from each other the *VRC* takes the value of one. (2) When all points are uniformly distributed in the space the relationship between *k* and *VRC* is monotonic and smooth. In this case CH suggest that each point should be considered its own cluster. (3) When the clear structure is present there should be a noticeable jump in the value of *VRC* when going from *k-1* to *k*.

3.2 Krzanowski-Lai criterion (1985)

Krzanowski and Lai (KL) in 1985 introduced a criterion for determining the optimal number of groups in a data set. KL assume that the criterion will be used with clustering algorithms based on minimizing the within-group sum of squares.

KL showed that under the assumption that the data set consists of *v* independent variables distributed uniformly with equal variances the expression $k^{\frac{2}{v}} \cdot \frac{WGSS}{TSS}$ equals approximately one for any value of *k*. *TSS* denotes the total sum of squares. Unfortunately KL simulations showed that this result holds poorly in small samples typically encountered in practical data analysis. KL also point out that the criterion often gives multiple local optima. However they found out that despite this facts the criterion is still useful for determining the optimal number of groups.

3.3 Rousseeuw criterion (1987)

Rousseeuw in 1987 paper introduced a silhouette graph. This graph was meant to be a new useful tool aiding in the interpretation of clustering results. It was designed to differentiate the clear group structure from merely a data set partition to non-overlapping sets of points. As

a side effect a criterion for determining the optimal number of groups in a data set was introduced.

For the *i*-th data point the average distance between the point and other points belonging to the same cluster is calculated and denoted $A(i)$. Next for the *i*-th data point the average distance between the point and other points belonging to some other cluster is calculated and repeated for all other clusters. The minimum of those averages is denoted $B(i)$. And finally for the *i*-th data point one gets the $S(i)$:

$$S(i) = \frac{B(i) - A(i)}{max\{A(i), B(i)\}}. \tag{6}$$

$S(i)$ takes values in range from -1 to +1. The higher the positive value the more certain it is that the point belongs to the correct cluster, since it is close to the rest of the cluster members and is far from the members of all other clusters. $S(i)$ close to zero shows uncertain cluster membership and the negative value suggests that the point is probably in the wrong cluster.

The global silhouette index (S) is calculated as an average value of all $S(i)$:

$$\mathrm{S} = \frac{1}{n}\sum_{i} S(i). \tag{7}$$

The *k* that reaches the maximum S is considered optimal. Rousseeuw points out that the advantage of the S index is that it was developed without any particular clustering algorithm in mind.

4 Simulation studies

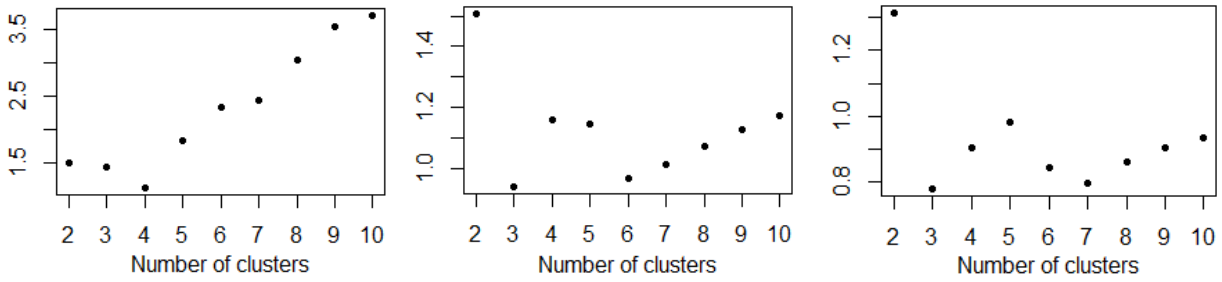
The performance of the proposed separability index AS has been studied through simulations carried under some theoretical models. The accuracy of the number of groups specification, has been compared to three other indexes described in the previous section. Simulations were carried out in R using *clusterSim* package for Caliński-Hrabasz, Krzanowski-Lai and Rousseeuw criteria. On each run of the simulation data was clustered by *k*-medoids method (R package *cluster*).

Experiment 1

Equal samples for two groups have been generate from two two-dimensional normal distributions: $N(-d,0,1,1,0)$ and $N(d,0,1,1,0)$, taking $d = 1$ and $d = 3$. The sample sizes were $n = 20, 100, 200$.

For $d = 3$ all criteria correctly identified the number of clusters, as two groups (See Fig. 1 for AS index). For small sample ($n = 20$) and $N(-1,0,1,1,0)$ and $N(1,0,1,1,0)$ model, all criteria

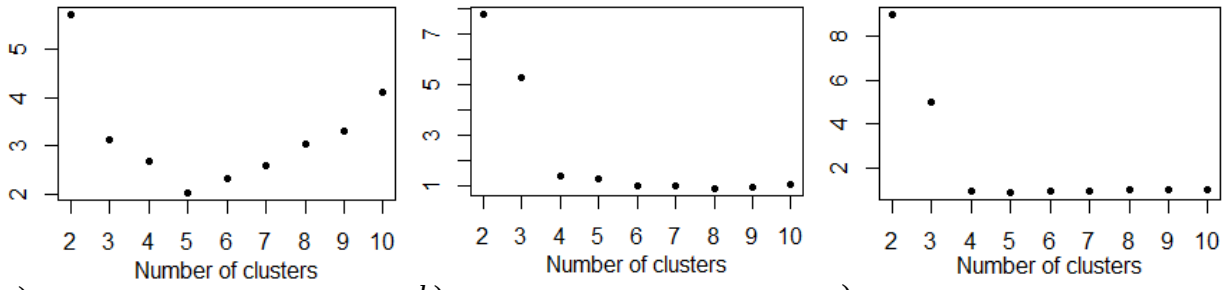
got lost showing number of clusters bigger than 2. For large samples and d = 1 three criteria, all except Krzanowski-Lai identified two clusters. On Fig 1, values of AS index are shown against number of clusters.



a) *b*) *c*)

Fig. 1 a, b, c. AS indexes for different number of clusters with a) n=20, b) n=100, c) n=200 samples generated from N(-1,0,1,1,0) and N(1,0,1,1,0).

Source: own calculations.



a) *b*) *c*)

Fig. 2 a, b, c. AS indexes for different number of clusters with a) n=20, b) n=100, c) n=200 samples generated from N(-3,0,1,1,0) and N(3,0,1,1,0).

Source: own calculations.

With two groups of points with expected values lying in 6 standard deviations distance, the AS index definitely suggests the correct number of clusters.

Experiment 2

In this experiment samples were generated from four two-dimensional normal distribution, moving along coordinate axes. Thus expected values were equal to (d,0), (0,d), (-d,0), (0,-d). A unit variance-covariance matrix has been used. Distance d from the center of coordinate system has been taken as d = 1, 2, …, 100. We generated four samples of equal size n = 20 and n=100 objects. Partitions for 2, 3, …, 10 groups were found by k-medoids method.

All four indexes performed very well. With small samples of $n = 20$, Caliński-Harabasz and Silhuette identified 4 groups in 99% runs, Krzanowski-Lai in 97%, and AS in 96%. With large sample $n = 100$, all criteria (except Krzanowski-Lai – 98%) got the score of 99%.

Experiment 3

In this experiment we have been trying to estimate the probability of correct identification of the number of groups in case when they are rather distant to each other. Four groups have been generated with the following model: $N(3,0,1,1,0)$, $N(0,3,1,1,0)$, $N(-3,0,1,1,0)$, $N(0,-3,1,1,0)$. The probability has been estimated by 1000 simulation runs. The data has been also clustered for some "incorrect" number of clusters 2, 3, …, 10, with *k*-medoids method.

Four samples of equal size *n*	**AS**	**CH**	**KL**	**S**
30	0.907	0.995	0.664	1.000
50	0.984	1.000	0.870	1.000
100	1.000	1.000	0.992	1.000

Table 1. Probability of correct identification of the number of groups in Experiment 3.
Source: own calculations.

The probability of correct identification increases with the sample size. Silhuette index looks ideal even for small samples.

Experiment 4

Four samples were generated from four two-dimensional normal distributions with unit variance-covariance matrix and expected values equal to $(d,0)$, $(0,d)$, $(-d,0)$, $(0,-d)$ with $d = 1, 2, \ldots, 100$. Then random noise was added, as extra 5% or 10% of previously generated points. The noise came from uniform distribution based on rectangular area defined by $((x_{min},y_{min}),(x_{min},y_{max}),(x_{max},y_{min}),(x_{max},y_{max}))$. On Fig. 3 we can see one example of data generated in above described way.

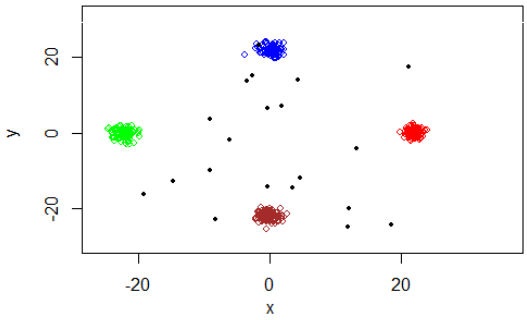


Fig. 3. Four clusters, each consisting of 100 points, d=22, with 5%.

Source: own calculations.

Random Noise	AS	CH	KL	S
5%	98%	16%	21%	77%
10%	99%	14%	11%	56%

Table 2. Percentage of correct identifications of number of clusters for d=1,…100.

Source: own calculations.

Simulation experiment with noise shows (See Table 2) the great advantage of our proposal over three other criteria.

Conclusions

Initial simulation studies reveals that the proposed separability index identifies the number of groups generally similarly to other three measures, expect the presence of random noise. In that case AS clearly outperforms the other three indexes.

Acknowledgement

The authors acknowledge support from research funds granted to the Faculty of Management at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

References

Arabie, P., Hubert, L., & De Soete, G. (1978). Clustering and Classification, *Journal of Classification*, 15 (2), 286-286

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1-27.

Everitt, B.S. (1995). Classification and Cluster-Analysis – Commentary, *British Medical Journal*, 311(7004), 536-536.

Everitt, B. S., Landau, S., Leese M., & Stahl, D. (2011). Cluster analysis. 5th Edition, John Wiley & Sons Ltd, Chichester.

Krzanowski, W. J., & Lai, Y. T. (1985). A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, *44*, 23-34.

Migdał-Najman, K. (2011). Ocena jakości wyników grupowania – przegląd bibliografii. *Przegląd Statystyczny*, R. LVIII, Zeszyt 3-4, 281-299.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

R packages used

cluster: M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, M. Studer, P. Roudier, 2015, *"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et. al.*, R Package version 2.0.3, https://cran.r-project.org/package=cluster

clusterSim: M. Walesiak, A. Dudek, *Searching for Optimal Clustering Procedure for a Data Set*, R Package version 0.44-2, cran.r-project.org/web/packages/clusterSim/index.html