OPTIMAL NUBER OF CLUSTERS PROVIDED BY k-MEANS AND E-M ALGORITHM

Vedran Novoselac* and Zlatko Pavić

Mechanical Engineering Faculty in Slavonski Brod, J. J. Strossmayer University of Osijek, Croatia
**Corresponding author e-mail: vnovosel@sfsb.hr*

Abstract

The paper considers the problem of determining the optimal number of clusters in data set by grouping index. The problem of clustering are provided with k-means and E-M (Expectation Maximization) algorithm. In addition to well-known indexes that are frequently used, two new indexes are presented. New indexes are based on the orthogonal distances from data to the line which represent corresponding cluster in the partition obtained with mentioned algorithms.

Keywords: k-means, E-M, grouping index

1. Introduction

The paper considers the problem of partitioned a set of data $\mathcal{A} = \{a_i \in \mathcal{R}^n : i = 1,\ldots,m\} \subset \mathcal{R}^n$ into k nonempty disjoint subsets $\pi_1,\ldots,\pi_k$, $1 \le k \le m$, such that $\bigcup_{i=1}^{k} \pi_i = \mathcal{A}$. The partition will be denoted by $\Pi(\mathcal{A}) = \{\pi_1,\ldots,\pi_k\}$. The elements of partition Π are called clusters and the set of all such partitions are denoted by $\mathcal{P}(\mathcal{A},k)$. For this purpose k-means and E-M algorithms are described [1, 2, 3, 6, 8]. Calculation of the various indexes on final partition indicates the quality of separateness and compactness of clusters. Some of the most popular indexes that are frequently used are Davies-Bouldin, Dunn, Calinski-Haradsz, and Simplify Silhouette Width Criterion [2, 6, 7, 9]. In addition to these indexes two new indexes Orthogonal Distances Criterion (ODC) and Width Ortogonal Distances Criterion (WODC) are presented. Indexes ODC and WODC are based on line which is determined by eigenvector of corresponding largest eigenvalue of covariance matrix and mean of observed cluster. The optimal number of clusters of data set $\mathcal{A}$ are provided with observation on mentioned indexes.

2. Grouping algorithms

This section presents the standard k-means and E-M algorithm. Mentioned algorithms have very broad application and they are often closely modified to the related issue [4,5].

2.1. k-means algorithm

Let $d(x,y) = \|x-y\|_p$, $p \ge 1$, be a metric (in paper we use Euclidian norm, i.e. $p = 2$). In the sense of the given metric center c_i of corresponding cluster π_i is defined as

$$c_i = \arg\min_{c \in \mathcal{R}^n} \sum_{a \in \pi_i} d(a,c). \qquad (1)$$

Data $a \in \mathcal{A}$ is attached to cluster π_i if is closest to the center c_i in comparison with the distances from the centers of other clusters. In that sense objective function $F : \mathcal{P}(\mathcal{A},k) \to \mathcal{R}$ is defined as

$$F(\Pi) = \sum_{i=1}^{k} \sum_{a \in \pi_i} d(a,c_i), \quad \Pi \in \mathcal{P}(\mathcal{A},k). \qquad (2)$$

Thus defined stopping criterion of k-means algorithm due to the so-called threshold $\varepsilon > 0$. Stoppage criterion is reached if absolute value of the difference between the objective function of iteration step does not exceed defined threshold ε. Thus, the k-means algorithm can be written as follows:

ALGORITHM 1. (k-means)

STEP 0.
Input number of clusters k, data set $\mathcal{A}$, stoppage criterion $\varepsilon > 0$, and initial centers $c_1^0,\ldots,c_k^0$. Set step of the algorithm $s = 0$;

STEP 1.
Apply the principle of minimum distance to determine the initial partition $\Pi^s = \{\pi_1^s,\ldots,\pi_k^s\}$, with an initial clusters

$$\pi_i^s = \{a \in \mathcal{A} : d(a,c_i^s) \le d(a,c_j^s),\ j = 1,\ldots,k\},$$

for every $i = 1,\ldots,k$;

STEP 2.

Form a new centroids c_i^{s+1}, $i=1,\ldots,k$, which are obtained by solving minimization problems

$$c_i^{s+1} = \arg\min_{c\in\mathcal{R}^n} \sum_{a\in\pi_i^s} d(a,c);$$

Create a new partition $\Pi^{s+1} = \{\pi_1^{s+1},\ldots,\pi_k^{s+1}\}$ with clusters

$$\pi_i^{s+1} = \{a \in \mathcal{A} : d(a,c_i^{s+1}) \le d(a,c_j^{s+1}),\ j=1,\ldots,k\},$$

for every $i=1,\ldots,k$;

STEP 3.

If $|F(\Pi^s) - F(\Pi^{s+1})| \le \varepsilon$ then **STOP**, else $s = s+1$ and go to **STEP 2.**

2.2. E-M algorithm

E-M algorithm is based on the principle of soft grouping, where the boundaries between clusters are not solid. Specifically, it is a probabilistic grouping that each element of the reference data set determines the probability of belonging to each cluster. E-M algorithm is generally based on the Gaussian mixture model. Gaussian mixture model approximates the data as a linear combination of k density

$$p(x) = \sum_{i=1}^{k} w_i\, f_i(x\,|\,\theta_i), \qquad (3)$$

where x is n-dimensional vector, and weights w_i, $i = 1,2,\ldots,k$ respectively represent the percentage of data belonging to a cluster π_i, $i = 1,2,\ldots,k$, what imply $\sum_{i=1}^{k} w_i = 1$. Parameter $\theta_i = (\mu_i, \Sigma_i)$ of density function $f_i(x|\theta_i)$ in Gaussian mixture model is presented with expectation μ_i and covariance matrix Σ_i determining the density function for the normal (Gaussian) distribution, i.e.

$$f_i(x\,|\,\theta_i) = \frac{1}{\sqrt{(2\pi)^l\,|\Sigma_i|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right), \qquad (4)$$

Quality of Gaussian mixture model presented with parameters $\Phi = \{(w_i, \mu_i, \Sigma_i) : i = 1,\ldots,k\}$ is measured with log-likelihood

$$L(\Phi) = \sum_{x\in S} \log\left(\sum_{i=1}^{k} w_i\, f_i(x\,|\,\mu_i,\Sigma_i)\right). \qquad (5)$$

The process is repeated until the log-likelihood of the mixture model at the previous iteration is sufficiently close to the log-likelihood of the current model. The algorithm proceeds as follows for Gaussian mixture model:

ALGORITHM 2. (E-M)

STEP 0.

Initialization of parameters $\Phi^0 = \{(w_i^0, \mu_i^0, \Sigma_i^0) : i = 1,\ldots,k\}$ (zero partition), $s = 0$, and stopage criterion $\varepsilon > 0$ (set by user or at random).

STEP 1. (E step)

For every $a \in \mathcal{A}$ calculate π_i cluster probability as

$$w_i^s(a) = \frac{w_i^s\, f_i(a\,|\,\mu_i^s,\Sigma_i^s)}{\sum_{n=1}^{k} w_n^s\, f_n(a\,|\,\mu_n^s,\Sigma_n^s)}, \quad i=1,\ldots,k.$$

STEP 2. (M step)

Calculation of new parameters for Gaussian mixture model for every $i = 1,2,\ldots,k$:

$$w_i^{s+1} = \sum_{a\in\mathcal{A}} w_i^s(a),$$

$$\mu_i^{s+1} = \frac{\sum_{a\in\mathcal{A}} w_i^s(a)\,a}{\sum_{a\in\mathcal{A}} w_i^s(a)},$$

$$\Sigma_i^{s+1} = \frac{\sum_{a\in\mathcal{A}} w_i^s(a)(a-\mu_i^{s+1})(a-\mu_i^{s+1})^T}{\sum_{a\in\mathcal{A}} w_i^s(a)}.$$

STEP 3.

If $|L(\Phi^s) - L(\Phi^{s+1})| \le \varepsilon$ then **STOP**, else $s = s+1$ and go to **STEP 1.**

3. Grouping indexes

In order to measure the compactness and separateness of k optimal partitions one of the most common indexes are used: DB (Davies-Bouldin), D (Dunn), CH (Calinski-Harabasz), SSC (Simplify Silhouette Width Criterion). In addition to these , we construct two new indexes ODC (Orthogonal Distances Criterion) and WODC (Width Orthogonal Distances Criterion). They are based on the sum of the orthogonal distance to the line which is passing through the centroid and is determined by the eigenvector of the largest eigenvalue of covariance matrix of observed cluster. Next figure

present such a lines of the case when set $\mathcal{A}$ are contained of the two clusters.

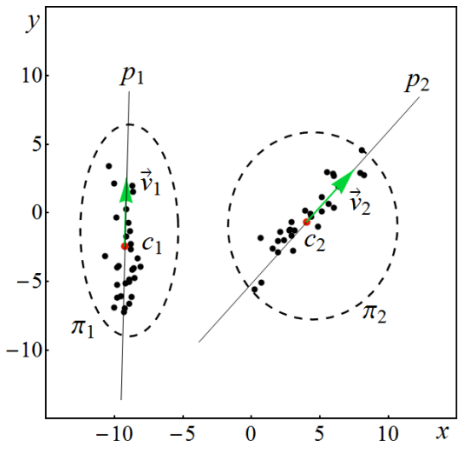


Figure 1. Index construction

Below we present a detailed definition of the aforementioned indexes:

a) Davies-Bouldin indeks

$$DB = \frac{1}{k}\sum_{i=1}^{k} R_i, \tag{6}$$

where

$$R_i = \max_{i \neq j} R_{ij}, \quad R_{ij} = \frac{r_i + r_j}{D_{ij}}, \quad D_{ij} = d(c_i, c_j),$$

$$r_i = \frac{1}{m_i}\sum_{a \in \pi_i} d(a, c_i).$$

More compact and better separated clusters will result smaller DB index.

b) Dunn indeks

$$D = \min_{1 \le i < j \le k} \left(\frac{D(\pi_i, \pi_j)}{\max_{1 \le s \le k} diam\, \pi_s} \right), \tag{7}$$

where

$$D(\pi_i, \pi_j) = \min_{a \in \pi_i, b \in \pi_j} d(a, b),$$

$$diam\, \pi_s = \max_{a,b \in \pi_s} d(a, b).$$

More compact and better separated clusters will result smaller D index.

c) Chalinski-Harabasz indeks

$$CH = \frac{(m-k)\mathcal{G}(\Pi)}{(k-1)\mathcal{F}(\Pi)}, \tag{8}$$

where $\mathcal{F}, \mathcal{G} : \mathcal{P}(\mathcal{A}, k) \to \mathcal{R}$ are functions defined as:

$$\mathcal{F}(\Pi) = \sum_{i=1}^{k}\sum_{x \in \pi_i} \|c_i - x\|_2^2, \quad \mathcal{G}(\Pi) = \sum_{i=1}^{k} m_i \|c_i - c\|_2^2,$$

$$c_i = \frac{1}{m_i}\sum_{a \in \pi_i} a, \quad c = \frac{1}{m}\sum_{a \in \mathcal{A}} a.$$

More compact and better separated clusters will result larger DB index.

d) Simplify Silhouette Width Criterion

$$SSC = \frac{1}{m}\sum_{a \in \mathcal{A}} \frac{\beta_{ai} - \alpha_{ai}}{\max\{\alpha_{ai}, \beta_{ai}\}}, \tag{9}$$

where for all $a \in \pi_i \cap \mathcal{A}$ follows that $\alpha_{ai} = d(a, c_i), \quad \beta_{ai} = \min_{j \neq i} d(a, c_j).$

e) Orthogonal Distances Criterion

$$ODC = \sum_{i=1}^{k} D_i, \tag{10}$$

where

$$D_i = \sum_{a \in \pi_i} \boldsymbol{d}(a, p_i).$$

$\boldsymbol{d}(a, p_i)$ present orthogonal distance from a to the line p_i which is determined by centroid c_i and eigenvector of corresponding largest eigenvalue. More compact and better separated clusters will result smaller ODC index.

f) Width Orthogonal Distances Criterion

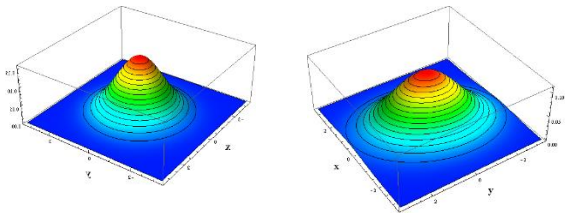
$$WODC = \sum_{i=1}^{k} W_i, \tag{11}$$

where

$$W_i = \frac{1}{d_i}\sum_{a \in \pi_i} \boldsymbol{d}(a, p_i), \quad d_i = \min_{\substack{j=1,..,k \\ j \neq i}} d(c_i, c_j).$$ More compact and better separated clusters will result smaller WODC index.

4. Experimental results1

In this chapter we examine data sets $\mathcal{A} \subset \mathcal{R}^2$ on implementing problem of determining k optimal partition obtained by k-means and E-M algorithm. The problem of finding an optimal partition of the set $\mathcal{A}$ can be reduced to the global optimization problem of objective function of k-means and E-M. In our case we run observed grouping algorithms for many different initial parameters and choose the solution with the best quality. The experimental data were generated using Gaussian random variable, i.e. $X \sim \mathcal{N}(\mu, \Sigma)$. The figure below shows illustrative examples of such functions density. In Figure 2(a) is present case with expectation $\mu = (0,0)$, and identity covariance matrix $\Sigma = I$. Figure 2(b) present case with $\mu = (0,0)$ and covariance matrix $\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$.



(a) (b)

Figure2 . PDF of Gaussian distribution

In the paper we also generated data with the extended model of the Gaussian distribution defined as

$$X \sim \mathcal{N}(\rho A + (1-\rho)B; \Sigma), \qquad (12)$$

where $A, B \in \mathcal{R}^n$ and ρ is uniformly distributed random variable within the interval $[0,1]$, i.e. $\rho \sim \mathcal{U}([0,1])$. Such data are distributed in a way that the expectation of the Gaussian distribution is uniformly distributed along the $\overline{AB}$. The density function of such defined random variable is given as

$$f(x \mid A, B, \Sigma) = \int_0^1 f(x \mid \rho A + (1-\rho)B, \Sigma)\mathrm{d}\rho. \qquad (13)$$

The Figure 3 below shows illustrative examples of such extended Gaussian PDF, where $A = (0,0)$, $B = (5,5)$, and identity covariance matrix $\Sigma = I$ are taken into the calculation of (13).

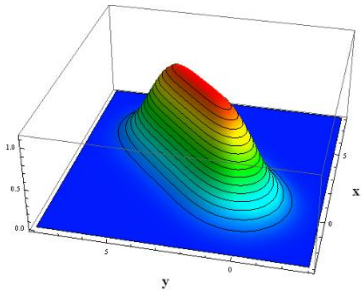


Figure 3. PDF of extended Gaussian distribution

The following examples illustrate the calculation of optimal number of clusters of data set $\mathcal{A} = \bigcup_{i=1}^{N} \mathcal{A}_i$, where every set $\mathcal{A}_i$ is generated with Gaussian random variable or extended Gaussian random variable defined by (12). To determinate the k optimal partition we have apply 100 randomly initializations of k-means and E-M algorithm respectively. Among the all randomly generated algorithms we presente those one with the best solution, i.e. best observed indexs value.

Example 1. ***2*** *The set of data $\mathcal{A}$ is generated by the Gaussian random variable which is presented in Figure 4. The Figure 5 shows the movement of the indexes of the k optimal partition.*

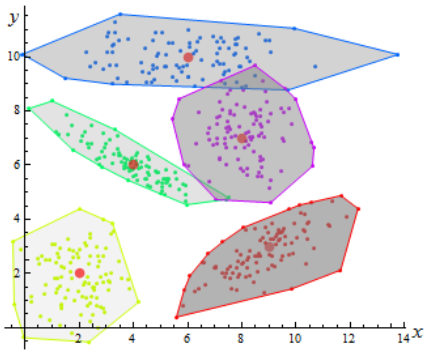
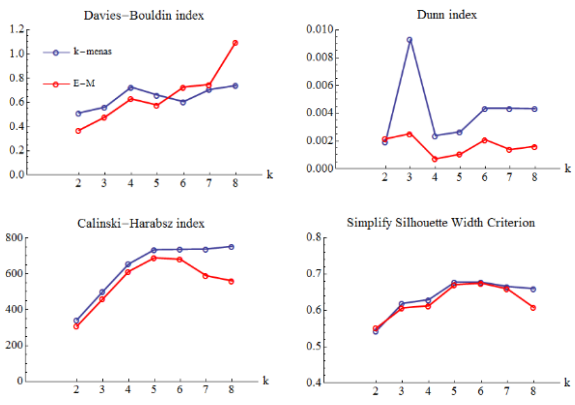


Figure 4. Data set $\mathcal{A}$



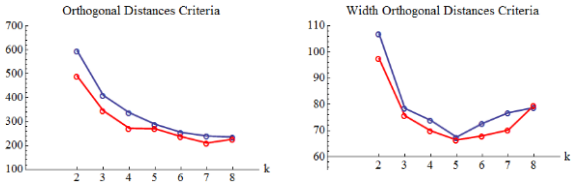


Figure 5. Indexes results

Example 2 *The set of data $\mathcal{A}$ is generated by the extended Gaussian random variable which is presented in Figure 6. The Figure 7 shows the movement of the indexes of the k optimal partition.*

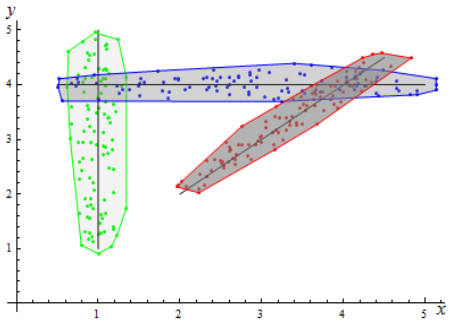


Figure 6. Data set $\mathcal{A}$

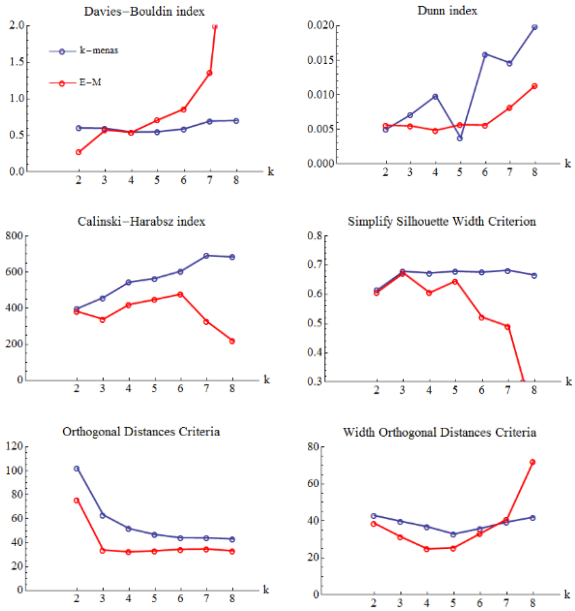


Figure 7. Indexes results

Example 3. ***3*** *The next set of data are generated by combination of Gaussian random variable and extended Gaussian random variable which is presented in Figure 6. The Figure 7 shows the movement of the indexes of the k optimal partition.*

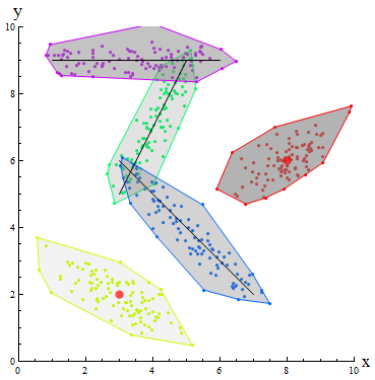


Figure 8. Data set $\mathcal{A}$

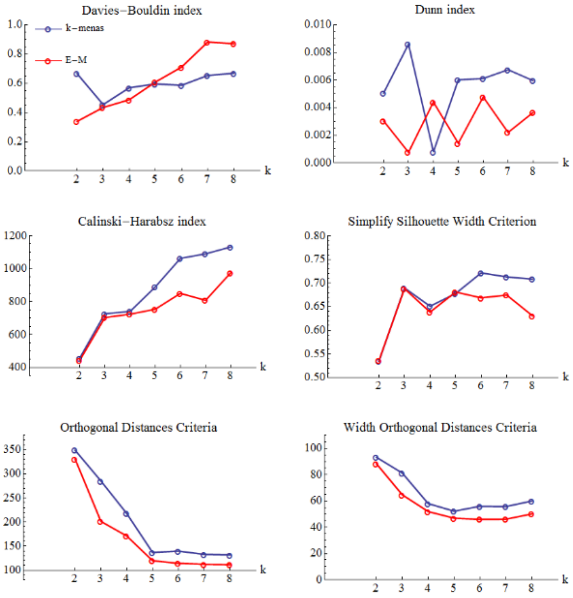


Figure 9. Indexes results

5. Conclusion

Problem of determination of the optimal numbers of clusters in observed data present a problem which we have solved by the investigation of the gruping indexes. Among the common used indexes we have construct two new indexes which shows good properties of finding the optimal number of clusters. Experimental results shows that the mentioned problem depends on many facts and indexes do not clearly shows unique optimal number. This shows that the problem must be precisely studied in order to finde optimal number of cluster what is present in the papper.

6. References

[1] P. S. Bradley, U. M. Fayyad, C. A. Reina, *Scaling EM (Expectation-Maximization) Clustering to Large Databases*, Microsoft Research, 1999.

[2] G. Gan, C. Ma, J. Wu, *Data Clustering, Theory, Algorithms and Applications*, SIAM, 2007.

[3] A. Kak, *Expectation-Maximization Algorithm for Clustering Multidimensional Numerical Data*, An RVL Tutorial Presentation, Summer 2012, Purdue University, 2013..

[4] T. Marošević, R. Scitovski, *Multiple ellipse fitting by center-based clustering*, Croatian Operational Research Review. 6(2015), 1; 43-53.

[5] V. Novoselac and Z. Pavić, *Outlier detection in experimental dana using a modified expectation maximization algorithm*, Proceedings of 6th International Scientific and Expert Conference of the International TEAM Society, Kecskemet, 2014, 112-115.

[6] K. Sabo, R. Scitovski, I. Vazler, *Grupiranje podataka: klasteri*, Osječki matematički list 10(2010), 149-176.

[7] R. Scitovski, S. Scitovski, *A fast partitioning algorithm and its application to earthquake investigation*, Computers and Geosciences, 59, 2013, 124-131.

[8] S. Theodoridis, K. Koutroumbas, *Pattern Recognition, Fourth Edition*, Elsevier, 2009.

[9] Lucas Vendramin, Ricardo J. G. B. Campello, Eduardo R. Hruschka, *On the Comparison of Relative Clustering Validity Criteria* , Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA. SIAM 2009, 733-744.