Pattern Recognition 43 (2010) 3374–3390

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Cluster validity index for estimation of fuzzy clusters of different sizes and densities

Krista Rizman Žalik

University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, SI-2000 Maribor, Slovenia

ARTICLE INFO

Article history:
Received 4 May 2009
Received in revised form
14 March 2010
Accepted 29 April 2010

Keywords:
Unsupervised classification
Fuzzy clustering
Cluster validity
Fuzzy *c*-means

ABSTRACT

Cluster validity indices are used for estimating the quality of partitions produced by clustering algorithms and for determining the number of clusters in data. Cluster validation is difficult task, because for the same data set more partitions exists regarding the level of details that fit natural groupings of a given data set. Even though several cluster validity indices exist, they are inefficient when clusters widely differ in density or size. We propose a clustering validity index that addresses these issues. It is based on compactness and overlap measures. The overlap measure, which indicates the degree of overlap between fuzzy clusters, is obtained by calculating the overlap rate of all data objects that belong strongly enough to two or more clusters. The compactness measure, which indicates the degree of similarity of data objects in a cluster, is calculated from membership values of data objects that are strongly enough associated to one cluster. We propose ratio and summation type of index using the same compactness and overlap measures. The maximal value of index denotes the optimal fuzzy partition that is expected to have a high compactness and a low degree of overlap among clusters. Testing many well-known previously formulated and proposed indices on well-known data sets showed the superior reliability and effectiveness of the proposed index in comparison to other indices especially when evaluating partitions with clusters that widely differ in size or density.

1. Introduction

Clustering is an unsupervised process of discovering significant groups hidden in a data set. It is a process of the assignment of data objects into subsets (called *clusters*) so that data objects in the same cluster are similar in some sense [1,2].

An important problem in use of the most clustering algorithms is to determine the best number of clusters to appropriately represent natural partitions. Clustering algorithms usually require that the number of clusters is determined before clustering [3]. It is difficult to decide when a clustering result is the best and what the optimal number of clusters is. Assessment of each solution result of clustering algorithms is the most important. For the same data set more partitions exists regarding the level of details that fit natural groupings of a given data set. In many cases, especially with higher dimensional data and overlap clusters, it is not clear which clustering solution is the best. To overcome this problem, various algorithms have been proposed that determine the optional number of clusters automatically, such as hierarchical methods [4], compatible cluster merging [5] and evolutionary approach [6].

The optional number of clusters can be determined also by executing the clustering algorithm several times with different number of clusters and then selecting the number of clusters that provides the best result observing a predefined criterion function [3]. This is the simplest approach to determine the number of clusters. It requires an efficient criteria-validity index to measure the quality of the solution partitions [7,8]. Several cluster validity indices for clustering have been proposed [3,8–17]. They are usually based on a variance analysis and on comparison of intra and inter-cluster variability. A good partition should have a small intra-cluster variance and a large inter-cluster separation at the same time. Variance or compactness is used as a measure of closeness within clusters, while separation is a measure of the isolation of clusters. Most indices are based on some average values and cluster centres. They are inefficient in validation of partitions containing clusters that widely differ in density or size. Most methods ignore small clusters or clusters with low densities.

Fuzzy clustering indices evaluate fuzzy partitions. In fuzzy clustering each data object can belong to more clusters. Membership levels associated with each data object describe how strong data element belongs to each cluster. Two simple, but efficient indices based only on the fuzzy membership values of fuzzy partitions are partition coefficient (*PC*) [9] and partition entropy (*PE*) [10]. Both indices use only the fuzzy membership values and may lack of considering the geometrical structure of clusters. Therefore, researchers have suggested many cluster validity indices that include both fuzzy membership values and the information of structures. Most of them are based on ratio of compactness to separation measures, as for example index *XB* [11]. But also

E-mail address: krista.zalik@uni-mb.si

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.04.025

summation type validity indices have been proposed. They calculate index value as a sum or a difference of different cluster validity measures, usually compactness and separation measures. Kim and Ramakrishna [13] suggested six new indices with mitigating limitations of existing indices discovered by an analysis of their design and performances for both categories of validity indices summation and ratio. Some indices use exponential function that is highly useful in dealing with classical Shannon entropy as for example indices *PCAES* [12] and Zhang's index V_w [14]. Enhancement of fuzzy clustering method and some validation criteria have been proposed also for enhanced fuzzy clustering [15]. Many existing validity indices do not perform well when clusters overlap and some indices have been proposed that are efficient also when clusters become less and less separable [17]. A lot of different cluster validity indices proposed and tested over the years have been offering a conclusion that no universally best measure exists.

A wide variety of clustering algorithms have been proposed for different applications [2,20]. They can be divided into overlapping, partitional and hierarchical [18]. Hierarchical clustering is a sequence of nested partitional clusterings. In each step an input data set is partitioned into a different number of subsets. Most data sets cannot be adequately split into different number of non-overlapping subsets while partitions may overlap with each other to some degree. Fuzzy clustering methods can efficiently process such data sets, because they discover fuzzy clusters to which all the data objects belong to some degree [3,14], while in crisp clustering each data object belongs to only one cluster. Fuzzy clustering algorithms better fit natural partitions than crisp clustering. The fuzzy *c*-means (*FCM*) algorithm [24] is one of the most widely used algorithms for fuzzy clustering. There are extended types of *FCM* in the literature, such as [19–23], etc. The results of *FCM* algorithm are very sensitive to the initial number of clusters *c*. Validity indices mitigate this problem. While validity indices are independent on fuzzy clustering algorithms, we use the *FCM* algorithm.

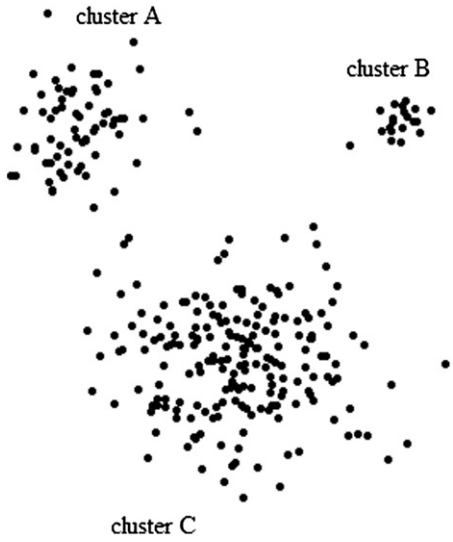


Fig. 1. Three clusters *A*,*B*,*C* with centres denoted by black rectangles. Cluster *B* is the most compact. Clusters *C* and *A* overlap more than clusters *C* and *B*, although the distances between centres of both pairs of clusters *A*,*C* and *B*,*C* are the same.

In this paper, we review and compare various clustering validity indices according to their efficiency when discovering clusters of different densities and sizes. We propose summation and ratio type validity index, in which the same compactness degree and the overlap degree are taken into comprehensive account. Experimental results on artificial and real-life data sets indicate that indices are stable and efficient also when validating partitions with clusters that widely differ in density or size.

In the next section, we review different cluster validity indices after introducing the fuzzy clustering. In Section 3, we propose two cluster validity indices based on compactness and overlap degree for evaluating fuzzy partitions. The comparison of the suggested indices against other popular validity indices for evaluating fuzzy clusters is given in Section 4. Conclusions are drawn in Section 5.

2. Fuzzy clustering and cluster validity

2.1. Fuzzy clustering and the fuzzy c-means algorithm

In fuzzy clustering, each data object belongs to more clusters with different membership degrees as in fuzzy logic rather than belonging to just one cluster. A data object near the centre of a cluster belongs to the cluster with a higher degree than an object on the edge of the cluster. For each data object x, a membership degree u_{ij} describes how strong the data object x_j belongs to the ith cluster.

Membership degree can have values between 0 and 1 (Eq. (1a)):

- The further away the object x_j from cluster centre of cluster i, the closer is the membership value u_{ij} to 0.
- Similarly, the closer the object x_j to the cluster centre of cluster i, the closer is the membership value u_{ij} to 1.
- The further away the object x_j is from all cluster centres, the closer the membership degree is to $1/c$, where c is the number of clusters.

The sum of membership coefficients expressing how strong one data object belongs to all clusters is 1 (Eq. (1b)). The sum of membership degrees of all n data objects to all c clusters is equal to n (Eq. (1c))

$$0 \le u_{ij} \le 1, \quad 1 \le j \le n, \quad 1 \le i \le c \tag{1a}$$

$$\sum_{i=1}^{c} u_{ij} = 1, \quad 1 \le j \le n \tag{1b}$$

$$\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} = n, \quad 1 \le j \le n \tag{1c}$$

Fuzzy *c*-means (*FCM*) [25] is a very often-used partition algorithm for fuzzy cluster analysis. This algorithm partitions the data set into *c* groups represented as fuzzy sets by maximizing

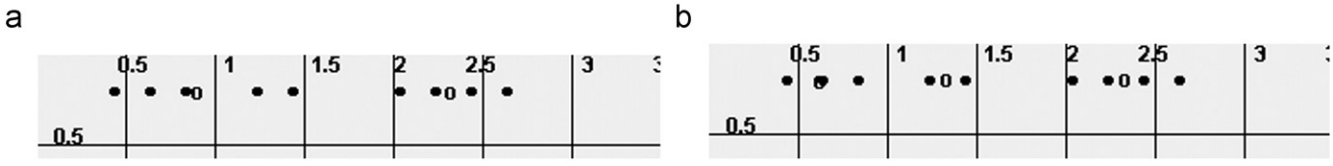


Fig. 2. Data set *A* with 9 data points partitioned into two (a) and three clusters (b). The cluster centres are marked with O.

the following cost function:

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n} {u_{ij}}^m \|x_j-v_i\|^2, \quad V=\{v_1,v_2,\ldots,v_c\}, \quad U=[u_{ij}],$$
$$1\le i\le n, 1\le j\le c \tag{2}$$

where c is the number of clusters, n is the number of data objects and m is fuzziness coefficient. $U=\lfloor u_{ij}\rfloor$ is c^*n fuzzy partition matrix and V is the set of cluster centres for all c clusters. Fuzzy partition is described by pairs (U, V). $\|x_j-v_i\|$ is Euclidean distance of data object x_j to cluster centre v_i. The *FCM* algorithm optimizes partitioning in iterations. It improves U and V in each iteration, and terminates when it reaches stable conditions. The *FCM* algorithm minimizes intra-cluster variance. It has the same problems as well-known crisp clustering algorithm *k*-means [20]: the obtained minimum can be a local minimum, and results depend on the initial choice of weights.

The fuzzy *c*-means algorithm is similar to *k*-means algorithm,

1. *Initialization*: Initialize membership coefficients—matrix $U=[u_{ij}]$, $U^{(0)}$, $k=0$.
2. $k=k+1$; Calculate the centre vectors $V^{(k)}=[v_i]$ with membership coefficients $U^{(k)}$. With fuzzy *c*-means, the centre of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$v_i=\frac{\sum_{j=1}^{n} {u_{ij}}^m x_j}{\sum_{j=1}^{n} {u_{ij}}^m}, \quad 1\le i\le c \tag{3}$$

3. For each point, compute its coefficients for being in the clusters, using the formula (4), and update $U^{(k)}$

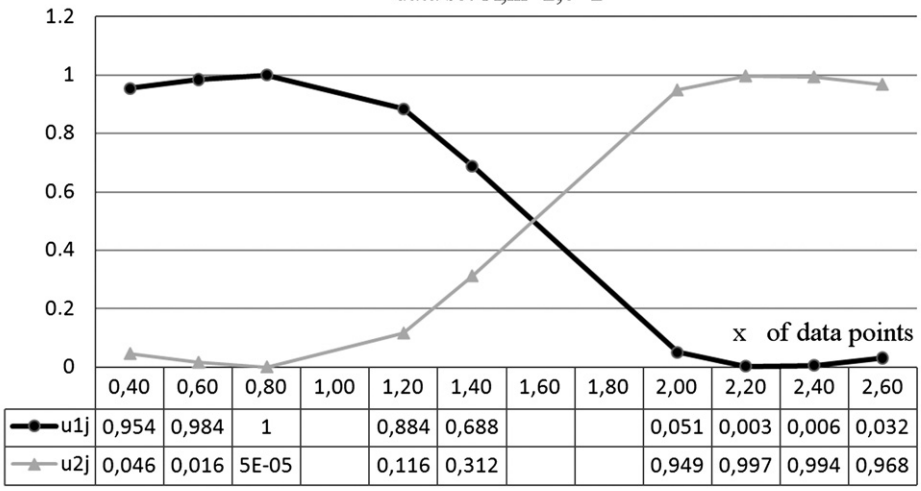
$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_j-v_i\|}{\|x_j-v_k\|}\right)^{2/(m-1)}}, \quad 1\le i\le c, 1\le j\le n \tag{4}$$

4. If $(U^{(k)}-U^{(k-1)})>\varepsilon$ go to step 2. Repeat steps 2 and 3 until the algorithm has converged.

2.2. Cluster validity indices for fuzzy clustering

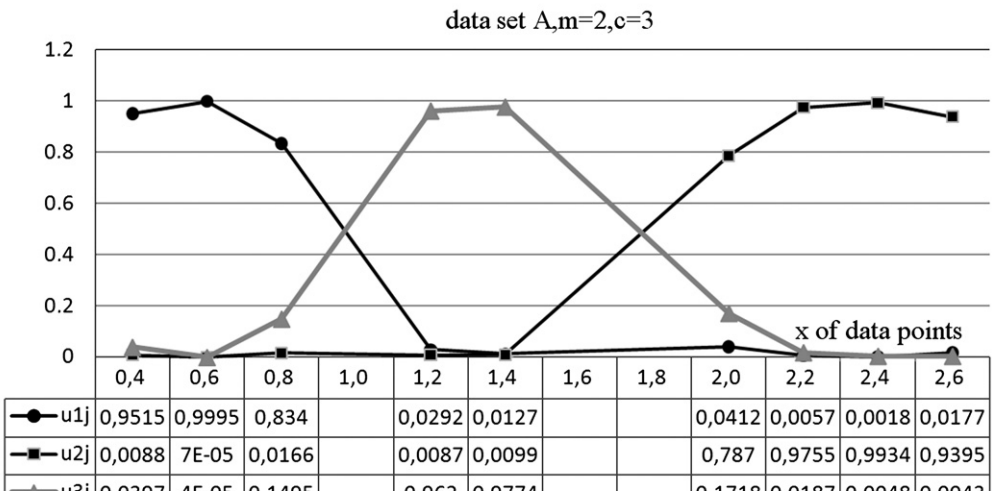
The first validity indices associated with *FCM* were partition coefficient (*PC*) [9] and partition entropy (*PE*) [10] defined

a



	0,40	0,60	0,80	1,00	1,20	1,40	1,60	1,80	2,00	2,20	2,40	2,60
u1j	0,954	0,984	1		0,884	0,688			0,051	0,003	0,006	0,032
u2j	0,046	0,016	5E-05		0,116	0,312			0,949	0,997	0,994	0,968

b



	0,4	0,6	0,8	1,0	1,2	1,4	1,6	1,8	2,0	2,2	2,4	2,6
u1j	0,9515	0,9995	0,834		0,0292	0,0127			0,0412	0,0057	0,0018	0,0177
u2j	0,0088	7E-05	0,0166		0,0087	0,0099			0,787	0,9755	0,9934	0,9395
u3j	0,0397	4E-05	0,1495		0,962	0,9774			0,1718	0,0187	0,0048	0,0043

c

no. of clusters	PC (max)	PE(min)	XB (min)	PCAES(max)	CO (max)	Cor(max)
2	**0,896**	**0,182**	**0,04**	**1,913**	0,866	1,499
3	0,886	0,232	0,056	1,47	**0,935**	**2**

Fig. 3. Curves u_{1i}, u_{2i} and u_{3i} describe the grade of membership of each data object of data set *A* to cluster 1, cluster 2 and cluster 3, respectively, in partitioning into two (a) and three clusters (b). Values of six observed indices for clustering of data set *A* into two and three clusters (c). Bold values show the recognized optional number of clusters (c).

as follows:

$$PC = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \tag{5}$$

$$PE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log(u_{ij}) \tag{6}$$

PC and *PE* are used to measure the fuzziness of the resulted clustering by the use of fuzzy partition matrix. The lower the fuzziness of a partition is, the larger is the *PC* value (or the smaller the *PE* value). The solution partition is obtained by maximizing *PC* or minimizing *PE* with respect to the number of fuzzy clusters.

Many validity indices have been proposed that combine membership degrees and the geometric structure of the data set. One of these is *XB* index proposed by Xie and Beni [11] that measures overall average compactness and separation. Smaller *XB* index means more compact and better-separated clusters. *XB* index is defined as follows:

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} {u_{ij}}^2 \|x_j - v_i\|^2}{n\left(\min_{i,j=1\ldots c,\ i\neq j}\|v_i - v_j\|^2\right)} \tag{7}$$

Wu and Yang [12] proposed *PCAES* (c) index, that uses exponential function:

$$PCEAS(c) = \sum_{i=1}^{c}\sum_{j=1}^{n}\frac{{u_{ij}}^2}{u_{Mj}} - \sum_{i=1}^{c} e\left(-\min_{k\neq i}\left(\frac{\|v_i - v_k\|^2}{\beta_T}\right)\right),$$

$$u_{Mj} = \min_{1\le i\le c}\sum_{j=1}^{n} u_{ij}^2, \quad \beta_T = \frac{1}{c}\sum_{i=1}^{c}\|v_i - \overline{v}\|^2, \quad \overline{v} = \sum_{j=1}^{n}\frac{x_j}{n}, \tag{8}$$

A large *PCAES* index means that all c clusters are compact and separated. The index consists of two terms. The first term is the normalized partition coefficient for measuring compactness. The second term is an exponential-type separation measure, which takes advantage of exponential function that measures the sum of distances between the closest pairs of cluster centres.

3. A new validity index using compactness and overlap measures

3.1. Motivation

Indices usually evaluate compactness by calculating the variance of all data objects belonging to a cluster. Greater variance means lower compactness. The common measure of variance of ith cluster is $\sum_{j=1}^{n} {u_{ij}}^2\|x_j - v_i\|^2$ and considers both distances of data objects from cluster centres and membership values describing fuzziness. These measure monotonic decreases to 0 with an increase in the number of clusters c: $\lim_{c\to n}\|x_j - v_i\| = 0$. Therefore, it is unable to validate partitions with large number of small clusters. Another shortcoming of the traditional measure of compactness is the wrong estimation of compactness for two clusters with the same number of elements and the same distribution but different size. For example, in Fig. 1

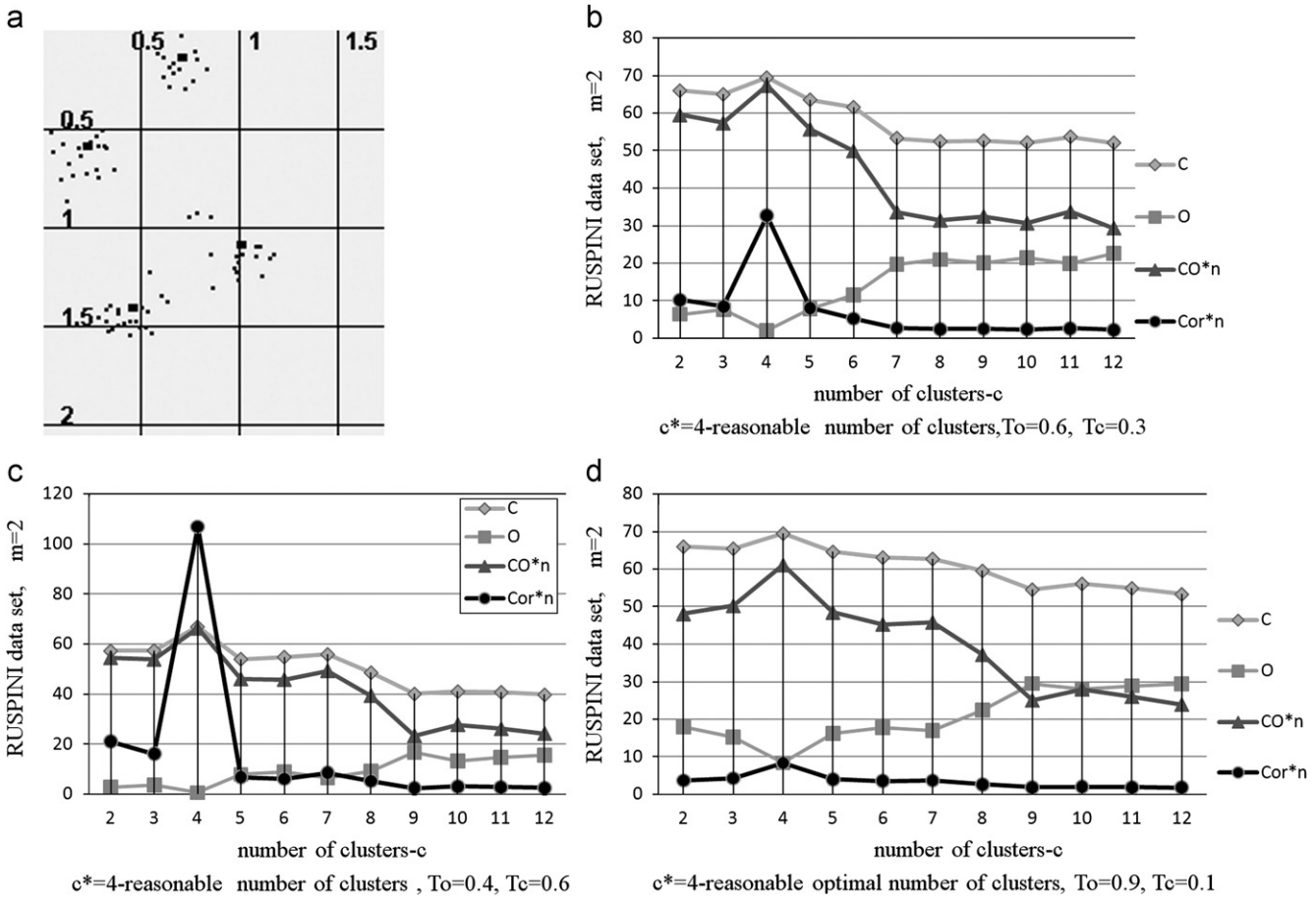
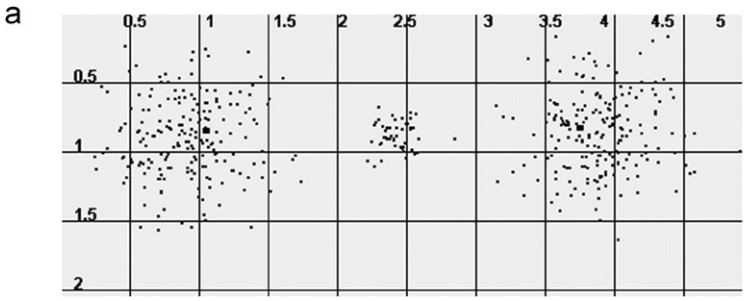


Fig. 4. Values of proposed compactness measure *C*, overlap measure *O*, summation type validity index *CO* and ratio type validity index CO_r for different values of T_o in T_c (b–d) for Ruspini data set [26] plotted in (a) for which four clusters is reasonable number of clusters.

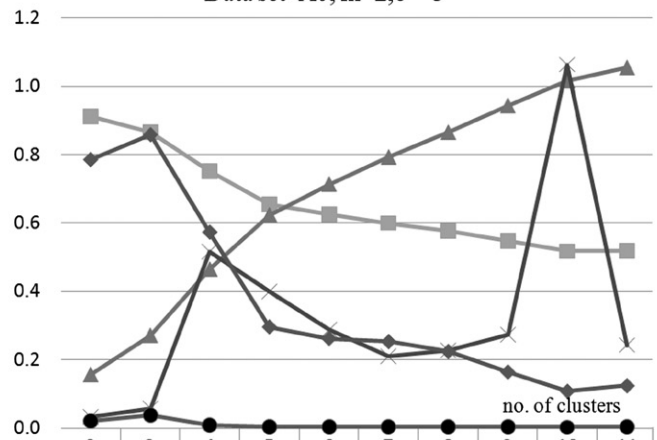
points in cluster A have larger distance to cluster centre that points in the cluster B. The traditional measure of compactness does not discover the cluster B as being more compact with small variance measure because the cluster B has smaller value $\|x_j - v_i\|^2$, but higher values for membership degrees u_{Bj}.

The separation measures have also shortcomings. Most often used separation measure between clusters is obtained by computing the distances between clusters' centres $\|v_i - v_j\|$. However, a separation measure based only on the distances between clusters' centres often does not give an accurate value for



b

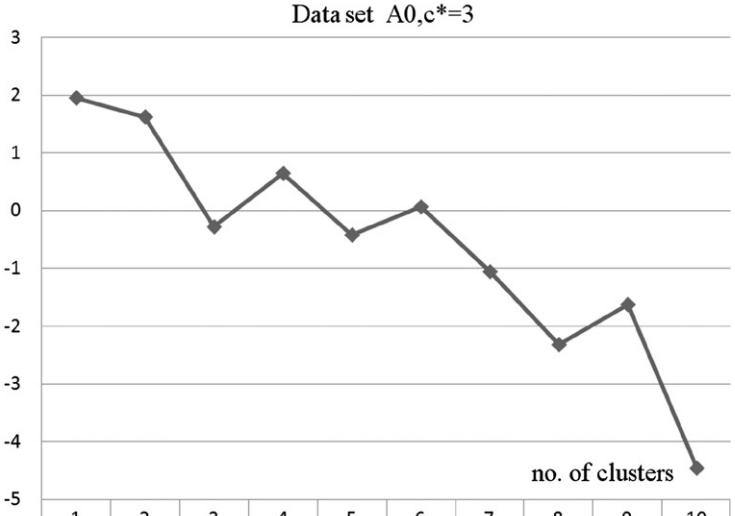
Data set A0, m=2,c*=3



	2	3	4	5	6	7	8	9	10	11
PC	0,911	0,865	0,752	0,655	0,625	0,600	0,577	0,548	0,517	0,519
PE	0,156	0,271	0,464	0,623	0,713	0,792	0,865	0,943	1,016	1,054
XB	0,033	0,057	0,515	0,399	0,288	0,211	0,226	0,274	1,063	0,243
CO	0,785	0,858	0,573	0,296	0,262	0,253	0,224	0,164	0,108	0,124
COr	0,020	0,039	0,008	0,004	0,004	0,004	0,004	0,003	0,003	0,003

c

Data set A0,c*=3



	1	2	3	4	5	6	7	8	9	10
PCAES m=2.0	1,950	1,621	-0,282	0,645	-0,423	0,065	-1,058	-2,323	-1,629	-4,462

Fig. 5. The data set $A0$ with three clusters with different sizes: (a) variation of the $PCAES$, PC, PE, XB, CO, CO_r with the number of clusters for data set $A0$ (b,c)

cluster separation. Fig. 1 shows that, for the same distance between two clusters, the separation of two clusters can be different. Intuitively, we see that clusters C and A are less separated than C and B. But with the use of conventional separation measure the separation of clusters C,A and C,B is equal, because $\|v_A - v_C\| = \|v_C - v_B\|$.

Some indices (PC, PE) use membership degrees as the only measure. But any increase in the number of clusters decreases the

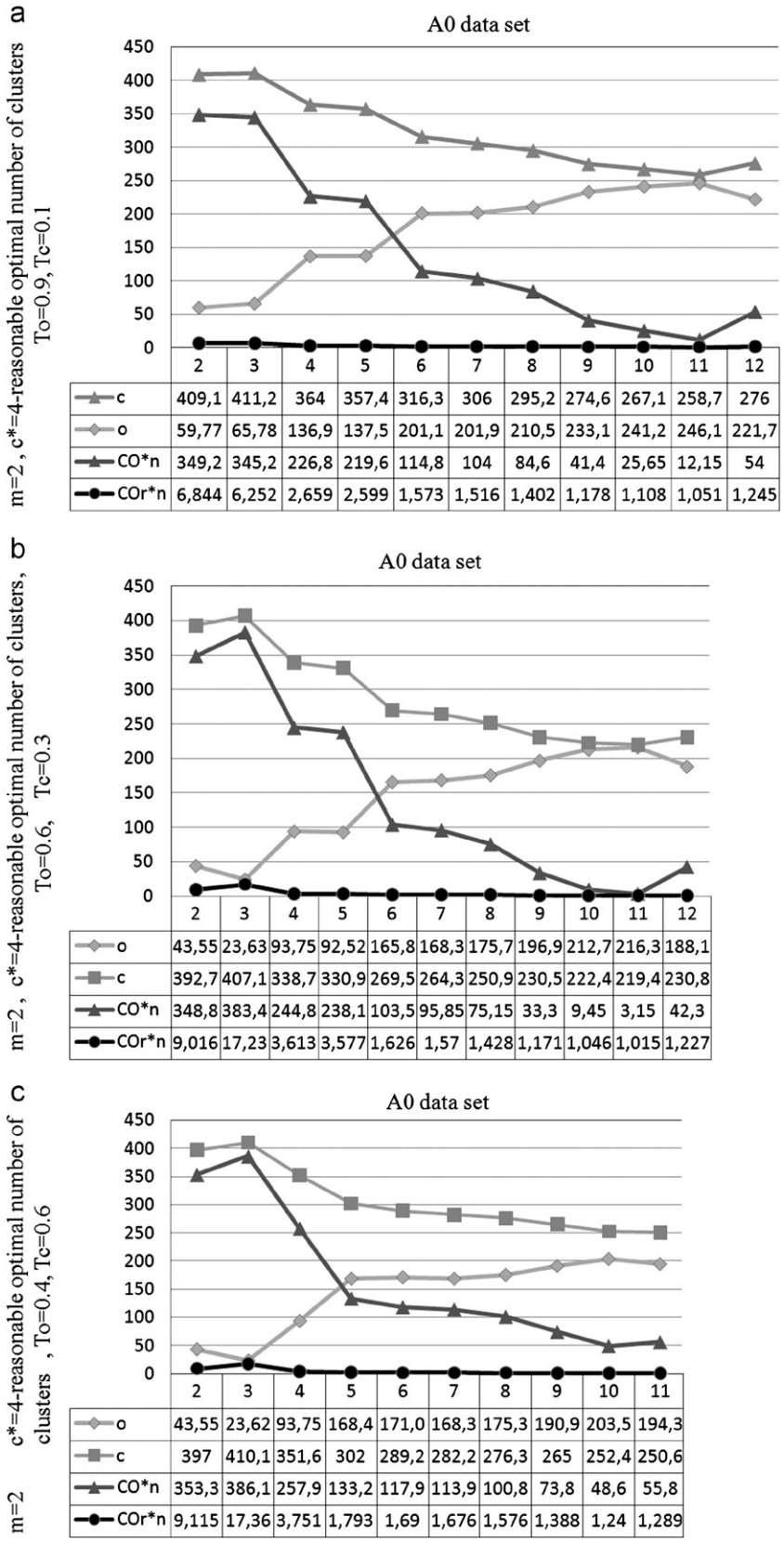
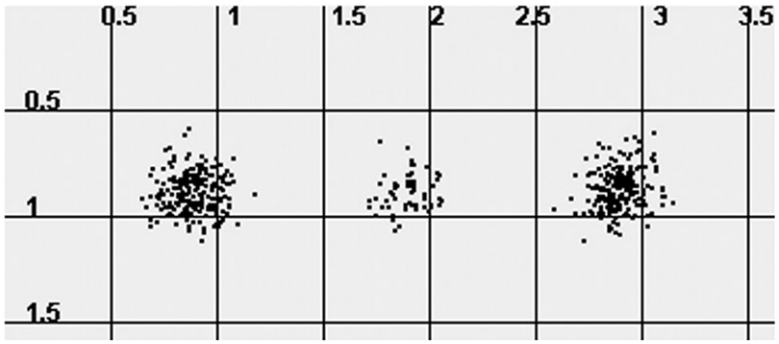
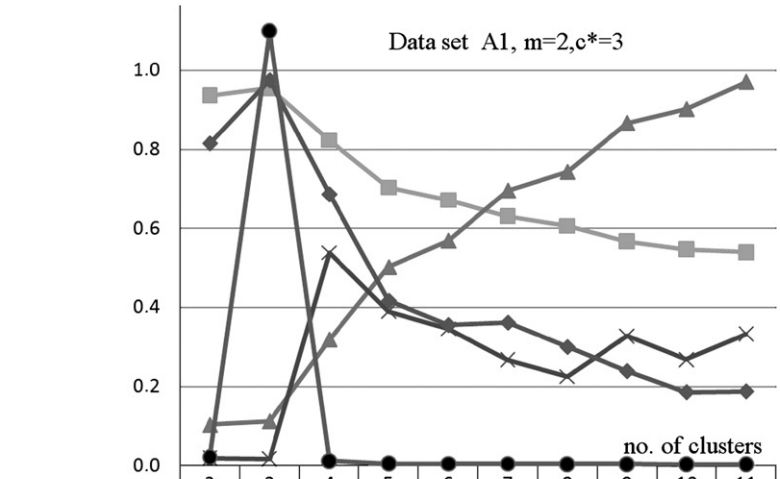


Fig. 6. Variation of the CO and CO_r with different values of T_c and T_o for data set $A0$ with three clusters with different sizes.

a

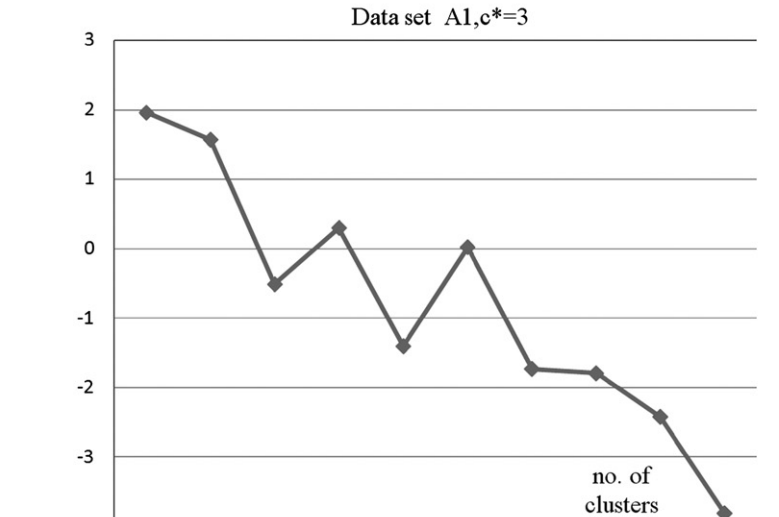


b



	2	3	4	5	6	7	8	9	10	11
PC	0,937	0,956	0,823	0,704	0,672	0,631	0,607	0,568	0,548	0,540
PE	0,103	0,112	0,319	0,502	0,569	0,696	0,743	0,867	0,902	0,971
XB	0,019	0,017	0,539	0,391	0,346	0,268	0,225	0,328	0,268	0,333
CO	0,816	0,976	0,687	0,417	0,356	0,362	0,300	0,239	0,186	0,188
COr	0,022	1,100	0,012	0,005	0,005	0,005	0,004	0,004	0,003	0,003

c



	1	2	3	4	5	6	7	8	9	10
PCAES m=2.0	1,961	1,571	-0,510	0,303	-1,403	0,025	-1,727	-1,791	-2,419	-3,806

Fig. 7. Data set *A*1 with three equal sized clusters with different densities (a). Variation of the indices *PC*, *PE*, *XB*, *CO*, CO_n, *PCAES* with the number of clusters for data set *A*1 (b,c).

membership degrees of all data objects that are not in a new cluster, as shown in the example in Fig. 3. We used a fuzzy c-means algorithm to group data set A with nine data objects into two (Fig. 2a) and three (Fig. 2b) clusters. In clustering into three clusters, the maximal membership values of all data objects in the new cluster 3 increased in comparison to membership values in clustering into two clusters. For two data points in cluster 3 with x coordinates 1.2 and 1.4, the maximal membership values increased from 0.88 and 0.68 (Fig. 3a) to 0.96 and 0.97 (Fig. 3b), respectively. The increase of membership values of data objects in cluster 3 indicates that these objects should be in a separate cluster. For most of the other data objects that remain in clusters 1 and 2 the maximal membership values decreased. For example, membership value of data object with x coordinate 0.8 decreased from $u_{1j}=0.99$ in the partition with two clusters to $u_{1j}=0.834$ in the partition with three clusters. Therefore, the validity index based on all membership values as an important part of the measure cannot be successfully used for validation index when validating clusters of different sizes and densities.

To overcome these shortcomings, we propose a new index that estimates two properties of fuzzy clusters: compactness and overlap. The compactness of clusters is used as a measure of variation or scattering of the data within the clusters. The smaller is scattering within the clusters, the higher is the compactness. The overlap measure indicates the degree of overlap between fuzzy clusters. When overlap is small, the separation is large, and each data object is clearly assigned to only one cluster. We assume that all clusters overlap, but at different degree. For example clusters C and A overlap more than clusters C and B on Fig. 1 although the distances between centres of both pairs of clusters A,C and B,C are the same. The basic goal of validation index is to find partition that maximizes compactness and minimizes overlapping.

3.2. The compactness measure

The degree of compactness for each partitioning can be quantified by summing the compactness of each cluster.

Definition 1. Let X be an n-data set $X=\{x_1,x_2,\ldots,x_n\}$ that is grouped into c clusters. Each data object x_j belongs to the ith cluster with a degree u_{ij}, satisfying constraint conditions for membership values u_{ij} described by Eqs. (1a), (1b) and (1c).

Definition 2. The fuzzy partition compactness degree $C(c,U)$ is defined as sum of compactness degrees of all cluster $C_i(c,U)$. The compactness degree of each cluster is calculated as a sum of compactness rates of all data objects $C_{ij}(c,U)$ as follows:

$$C(c,U)=\sum_{i=1}^{c}C_i(c,U),\quad C_i(c,U)=\frac{1}{n}\sum_{j=1}^{n}C_{ij}(c,U)$$
$$C_{ij}(c,U)=\begin{cases}u_{ij} & \text{if } (u_{ij}-u_{ik})\geq T_c,\quad k=1,\ldots,c,\quad k\neq j\\ 0 & \text{otherwise}\end{cases} \tag{9}$$

The jth data point increases the compactness of the ith cluster when it belongs to the ith cluster strong enough, that is when u_{ij} is greater than T_c from all other membership values describing the degrees of association of this data objects to other clusters. In all examples we took parameter T_c 0.6. Following Definition 2, the compactness degree $C_i(c,U)$ has the following properties:

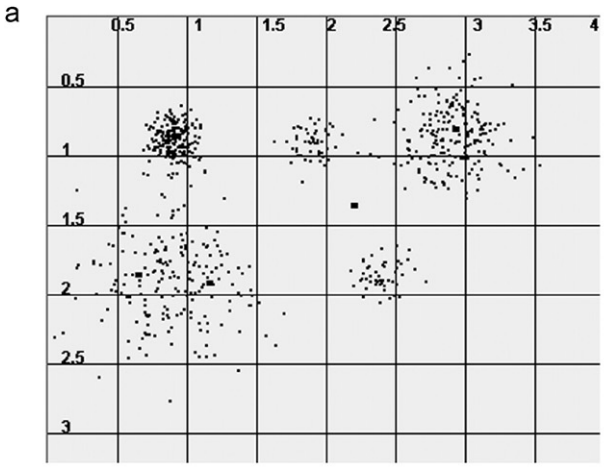
Property 1. *Compactness degree of each data object $C_{ij}(c,U)$ is bounded, $0\leq C_{ij}(c,U)\leq 1$.*

Property 2. *Compactness degree of a whole partition is also bounded, $0\leq C(c,U)\leq n$. The more compact the fuzzy partition, the higher the value of compactness degree. In contrast, the fuzzier the partition, the smaller is the value of compactness degree.*

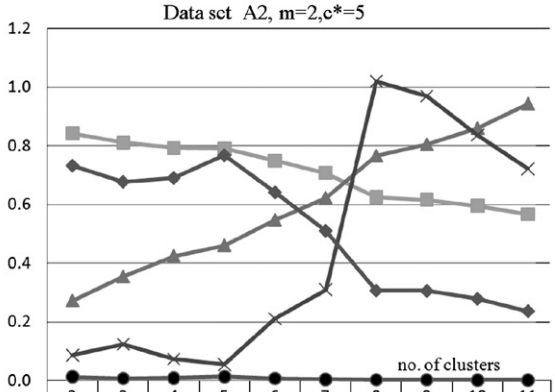
Property 3. *$C(c,U)=n$ in crisp clustering. Compactness degree of a whole partition reaches its maximum n in the crisp partition.*

Property 4. *$C(c,U)=0$ if $U=[1/c]$ and $T_c>0$. If all objects belong to all clusters uniformly, the compactness degree of fuzzy partition reaches its minimum 0. This is the fuzziest partition for any data set.*

The measure is based only on membership values that are calculated from distances from data object to all clusters' centres

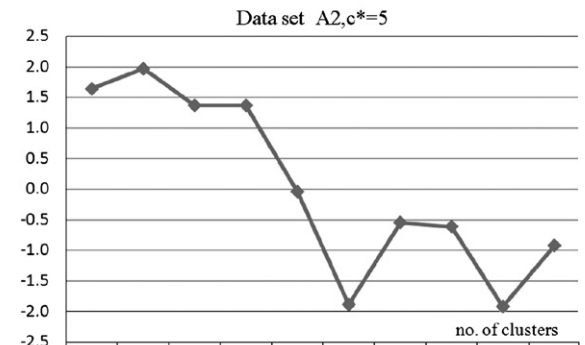


b



	2	3	4	5	6	7	8	9	10	11
PC	0,843	0,811	0,793	0,791	0,749	0,708	0,626	0,616	0,596	0,567
PE	0,273	0,354	0,425	0,461	0,547	0,622	0,767	0,805	0,860	0,943
XB	0,087	0,124	0,074	0,055	0,211	0,310	1,020	0,970	0,837	0,722
CO	0,733	0,678	0,691	0,768	0,642	0,511	0,307	0,306	0,279	0,237
COr	0,013	0,008	0,009	0,014	0,007	0,004	0,003	0,003	0,003	0,002

c



	1	2	3	4	5	6	7	8	9	10
PCAES m=2.0	1,645	1,974	1,376	1,375	-0,03	-1,88	-0,54	-0,60	-1,92	-0,92

Fig. 8. Data set $A2$ where five clusters is reasonably optimal number of clusters (a). Variation of the $PCAES$, PC, PE, XB, CO, CO_r with the number of clusters for data set $A2$. (b, c)

(Eq. (4)). The nearer is data object x_j is to the ith cluster centre v_i and the further away it is from other clusters' centres, the higher is the degree of membership u_{ij} and closer to 1.

3.3. The overlap measure

As mentioned earlier, in most cluster validity indices the separation measures are calculated based on distances among clusters' centres. However, this measure does not consider different cluster shapes. To overcome this shortcoming, the overlap degree is calculated from the membership values and used in the proposed index.

Definition 3. Overlap measure $O_{ab}(c,U: C_a, C_b)$ between two clusters C_a and C_b is computed from overlap degrees $O_{abj}(c,U: C_a, C_b)$ of each data objects x_j, that is associated strong enough to both fuzzy clusters C_a and C_b. A small value of overlap measure $O_{ab}(c,U: C_a, C_b)$ between two clusters C_a and C_b indicates that two clusters have small overlap and that clusters are well separated

$$O_{ab}(c,U)=\frac{1}{n}\sum_{j=1}^{n}O_{abj}(c,U),\quad a,b=1,\ldots,c,\quad a\neq b$$

$$O_{abj}(c,U)=\begin{cases}1-(u_{aj}-u_{bj}) & \text{if } (u_{aj}-u_{bj})\geq T_o \text{ and } a\neq b\\ 0 & \text{otherwise}\end{cases} \tag{10}$$

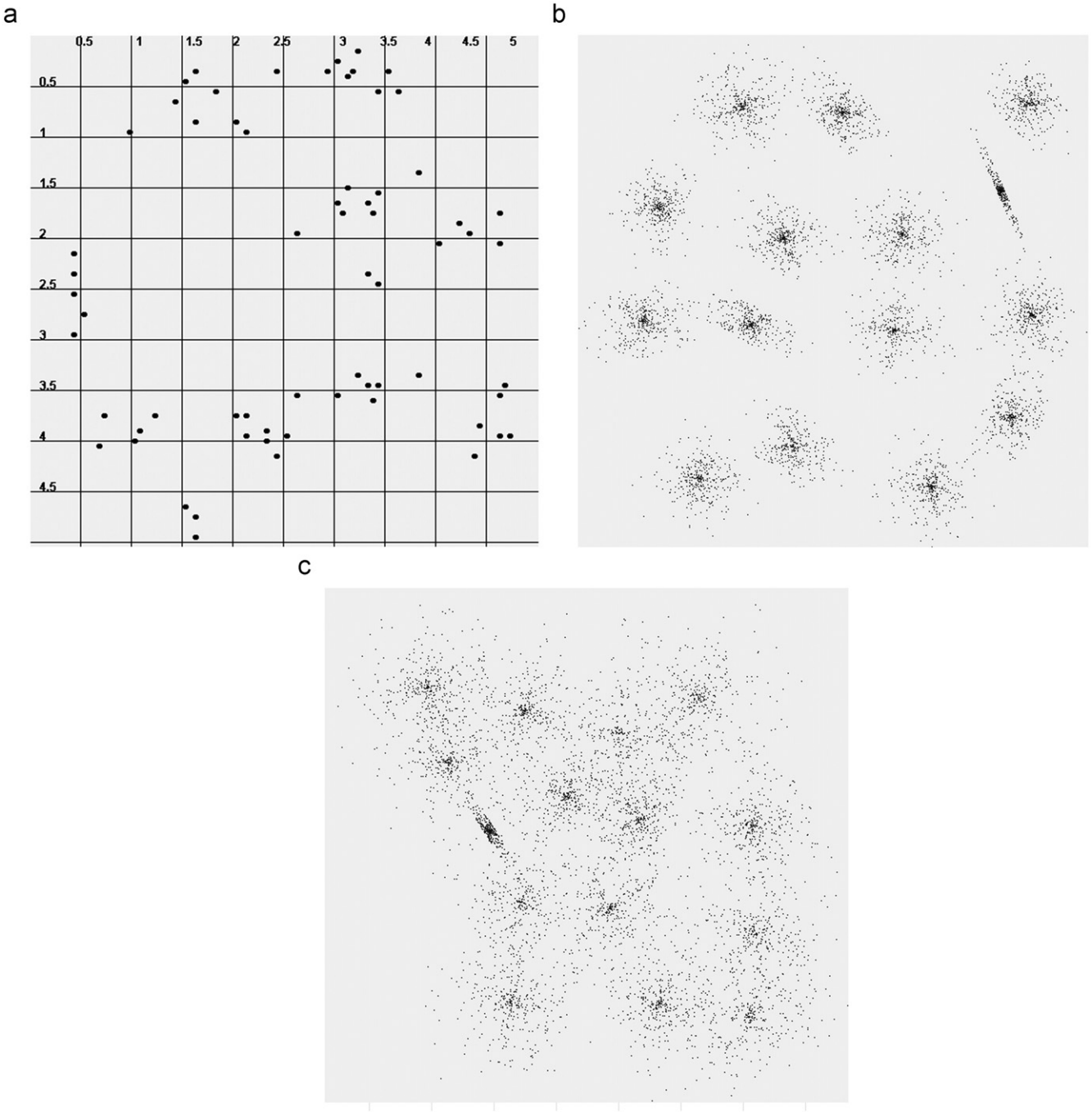


Fig. 9. STARFIELD data set (a) and two synthetic 2D data with 5000 vectors and 15 Gaussian clusters with different degree of cluster overlap $S1$ (b) and $S3$ (c).

The overlap measure satisfies the following properties:

Property 1. *The overlap degree is bounded*: $0 \le O_{abj}(c,U{:}C_a,C_b) \le 1-T_o$

Property 2. $O_{ab}(c,U{:}\ C_a, C_b)$—*overlap between two fuzzy clusters* C_a *and* C_b *is bounded* $0 \le O_{abj}(c,U{:}\ C_a, C_b) \le 1-T_o$.

Property 3. *The measure is commutative*: $O_{ab}(c,U{:}\ C_a, C_b)=O_{ba}(c,U{:}\ C_b, C_a)$.

Using the upper definition for each pair of clusters, we define the overlap measure for the whole partition.

Definition 4. The overlap measure of the whole fuzzy partition is defined as the sum of the overlap values for each pair of clusters:

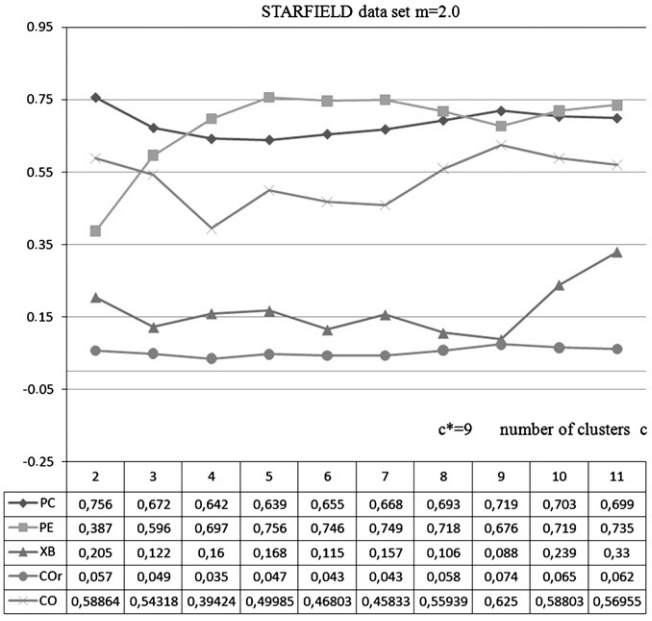
$$O(c,U)=\sum_{a=1}^{c-1}\sum_{b=a+1}^{c} O_{ab}(c,U) \tag{11}$$

Overlap gives the degree of overlap among all clusters in a fuzzy c-partition. A small overlap value means well-separated fuzzy partitions.

3.4. New validity indices

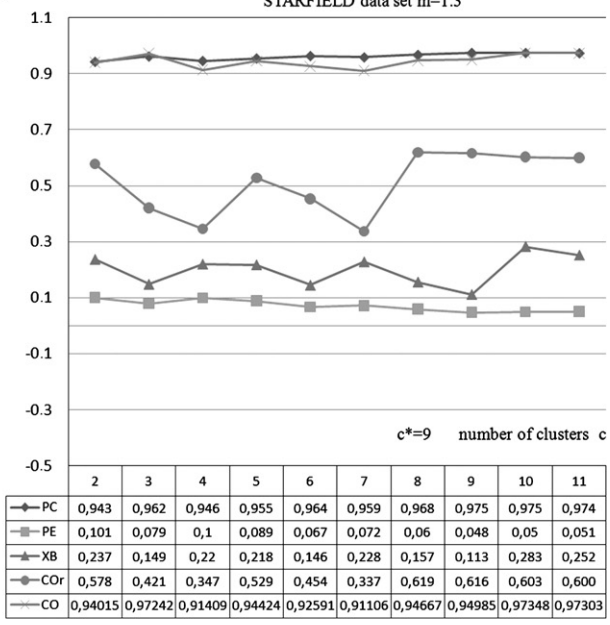
Cluster validity indices can be classified into two categories. Ratio type indices are characterized by the ratio of intra-cluster compactness distance to inter-cluster separation distance. Summation type indices are defined as sum of compactness and separation with appropriate weighting factor. However, the ratio type index cannot be used in initial state of clustering, which is mainly comprised with singleton clusters. Intra-cluster distance of a cluster with one data objects is 0 and value of index is 0 or ∞.

a



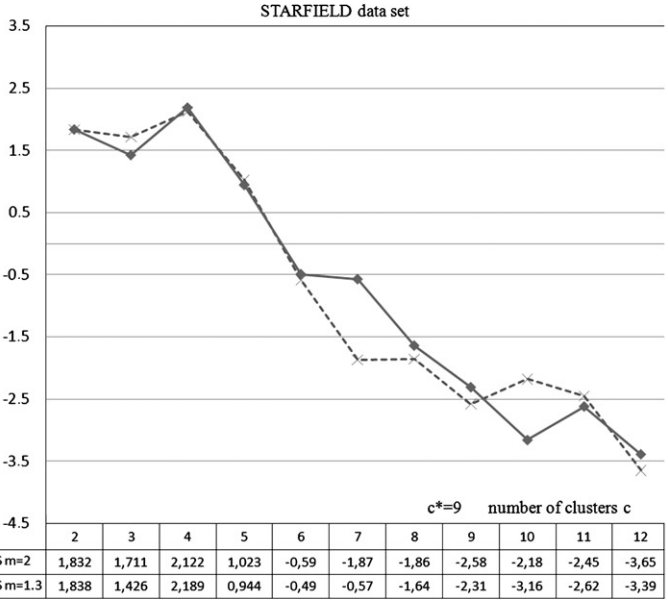
	2	3	4	5	6	7	8	9	10	11
PC	0,756	0,672	0,642	0,639	0,655	0,668	0,693	0,719	0,703	0,699
PE	0,387	0,596	0,697	0,756	0,746	0,749	0,718	0,676	0,719	0,735
XB	0,205	0,122	0,16	0,168	0,115	0,157	0,106	0,088	0,239	0,33
COr	0,057	0,049	0,035	0,047	0,043	0,043	0,058	0,074	0,065	0,062
CO	0,58864	0,54318	0,39424	0,49985	0,46803	0,45833	0,55939	0,625	0,58803	0,56955

b



	2	3	4	5	6	7	8	9	10	11
PC	0,943	0,962	0,946	0,955	0,964	0,959	0,968	0,975	0,975	0,974
PE	0,101	0,079	0,1	0,089	0,067	0,072	0,06	0,048	0,05	0,051
XB	0,237	0,149	0,22	0,218	0,146	0,228	0,157	0,113	0,283	0,252
COr	0,578	0,421	0,347	0,529	0,454	0,337	0,619	0,616	0,603	0,600
CO	0,94015	0,97242	0,91409	0,94424	0,92591	0,91106	0,94667	0,94985	0,97348	0,97303

c



	2	3	4	5	6	7	8	9	10	11	12
PCAES m=2	1,832	1,711	2,122	1,023	-0,59	-1,87	-1,86	-2,58	-2,18	-2,45	-3,65
PCAES m=1.3	1,838	1,426	2,189	0,944	-0,49	-0,57	-1,64	-2,31	-3,16	-2,62	-3,39

Fig. 10. Validity indices for STARFIELD data set, for $c=2,\ldots,12$ and for different degree of fuzziness $m=2.0$ (a, c) and $m=1.3$ (b, c).

We propose summation type of validity index CO using defined measures for compactness and overflow

$$CO(c,U) = C(c,U) - O(c,U) = \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{c} C_{ij}(c,U) - \sum_{a=1}^{c-1}\sum_{b=a+1}^{c} O_{abj}(c,U)\right) \quad (12)$$

We define also ratio type validity index CO_r using the same measures of compactness and overlap:

$$CO_r(c,U) = \frac{1}{n}\frac{C(c,U)}{O(c,U)} = \frac{1}{n}\sum_{j=1}^{n}\frac{\sum_{i=1}^{c} C_{ij}(c,U)}{\sum_{a=1}^{c-1}\sum_{b=a+1}^{c} O_{abj}(c,U)} \quad (13)$$

From Eqs. (12) and (13), we can see that compact and well-separated clusters correspond to the high value of validity index CO and CO_r, i.e. the high value of compactness C and the small value of overlap O. It tends to indicate a good cohesion within clusters and a small overlap between clusters.

The solution fuzzy c-partitioning and the number of clusters c^* of a given data set can be found by partitioning the data by fuzzy c-means algorithm into $2,3,\ldots,c_{max}$ $(c_{max} < \sqrt{n})$ clusters and finding the maximal value of validity index CO and CO_r. For fuzzy c-means we used the following parameters: $eps = 0.0001$; $m = 2$; iterations = 100. While fuzzy c-means is sensitive to initialization values, we repeat fuzzy c-means clustering for each number of cluster c ten times and calculate proposed validity indices CO and CO_r and then used average values.

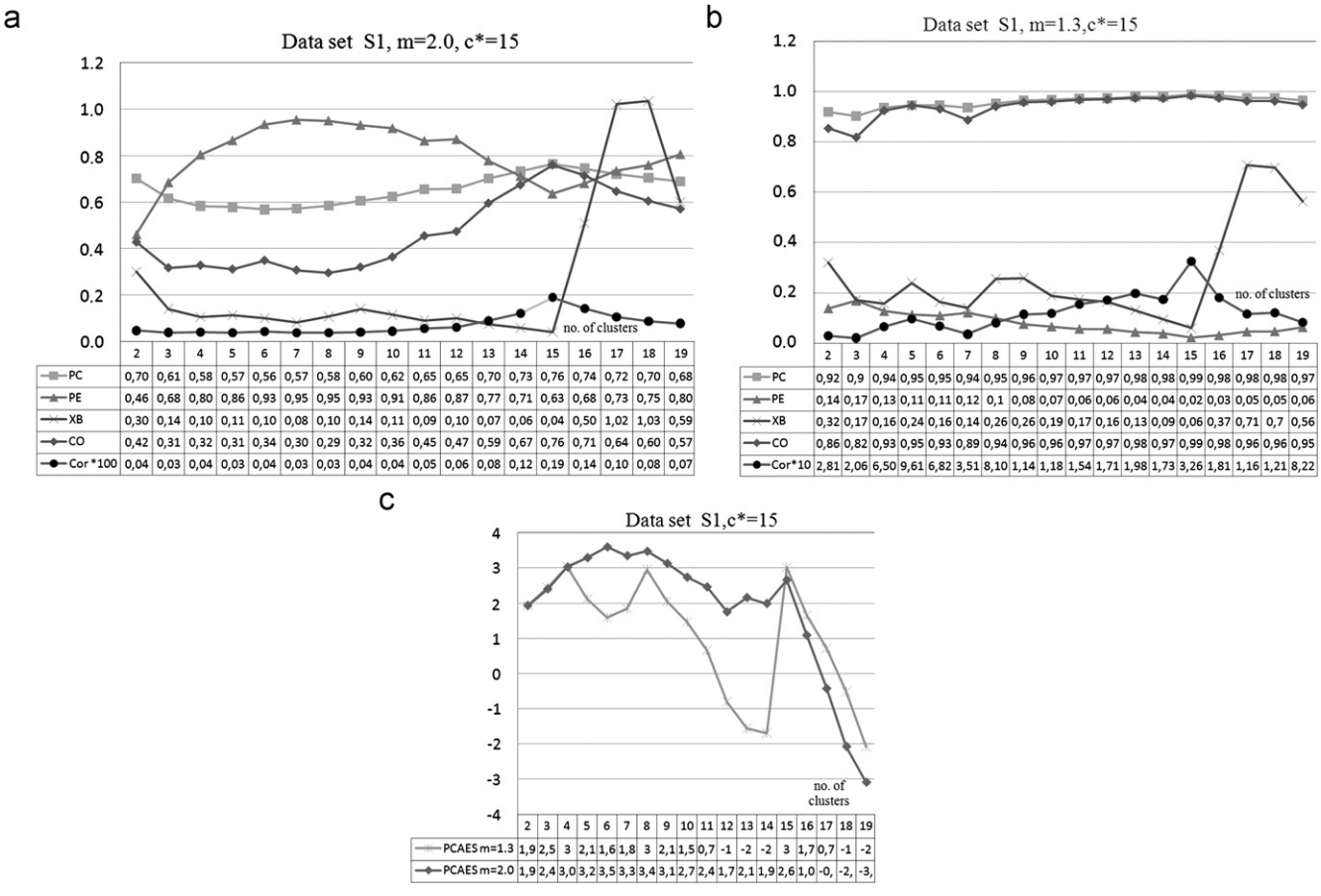
The variation of compactness C, overlap measures O and validity indices CO and CO_r for Ruspini data set [21], plotted in Fig. 4a, are shown in Fig. 4b–d for different values of parameters T_o and T_c. Parameters do not influence the results, since Ruspini data set consists of four clusters of equal size and density. From Fig. 4b–d we see that compactness decreases and overlap increases as c decreases or increases from optimal c^*. $T_o = 0.4$ and $T_c = 0.6$ provide the CO_r index to identify optional number of clusters with the greatest increase in value, while the changes of parameters do not influence the CO index. Small parameter T_o decreases the importance of overlap in the whole measure and can be used for data sets with overlapping clusters. $T_o = 1$ and $T_c = 0$ allow all membership values to be used for compactness and overflow measure, that cannot be successfully used for validation index when validating clusters of different sizes and densities.

4. Analysis and experimental results

To demonstrate the effectiveness with which the proposed indices CO and CO_r determine the optimal partition, we performed comparisons with other indices mentioned in Section 2.2 (*PC* [9], *PE* [10], *XB* [11], *PCAES* [12]) on nine synthetic and real data sets. We performed tests on synthetic data sets, known 2D data sets and real data sets. In the second experiment on known 2D data sets we tested the reliability of each index for different fuzziness coefficients 2.0 and 1.3.

4.1. Synthetic data sets

First we demonstrate the performance of proposed validity indices with testing it on three data sets with known structure. We created artificial 2D data sets by using different number of



a	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
PC	0,70	0,61	0,58	0,57	0,56	0,57	0,58	0,60	0,62	0,65	0,65	0,70	0,73	0,76	0,74	0,72	0,70	0,68
PE	0,46	0,68	0,80	0,86	0,93	0,95	0,95	0,93	0,91	0,86	0,87	0,77	0,71	0,63	0,68	0,73	0,75	0,80
XB	0,30	0,14	0,10	0,11	0,10	0,08	0,10	0,14	0,11	0,09	0,10	0,07	0,06	0,04	0,50	1,02	1,03	0,59
CO	0,42	0,31	0,32	0,31	0,34	0,30	0,29	0,32	0,36	0,45	0,47	0,59	0,67	0,76	0,71	0,64	0,60	0,57
Cor *100	0,04	0,03	0,04	0,03	0,04	0,03	0,03	0,04	0,04	0,05	0,06	0,08	0,12	0,19	0,14	0,10	0,08	0,07

b	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
PC	0,92	0,9	0,94	0,95	0,95	0,94	0,95	0,96	0,97	0,97	0,97	0,98	0,98	0,99	0,98	0,98	0,98	0,97
PE	0,14	0,17	0,13	0,11	0,11	0,12	0,1	0,08	0,07	0,06	0,06	0,04	0,04	0,02	0,03	0,05	0,05	0,06
XB	0,32	0,17	0,16	0,24	0,16	0,14	0,26	0,26	0,19	0,17	0,16	0,13	0,09	0,06	0,37	0,71	0,7	0,56
CO	0,86	0,82	0,93	0,95	0,93	0,89	0,94	0,96	0,96	0,97	0,97	0,98	0,97	0,99	0,98	0,96	0,96	0,95
Cor*10	2,81	2,06	6,50	9,61	6,82	3,51	8,10	1,14	1,18	1,54	1,71	1,98	1,73	3,26	1,81	1,16	1,21	8,22

c	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
PCAES m=1.3	1,9	2,5	3	2,1	1,6	1,8	3	2,1	1,5	0,7	-1	-2	-2	3	1,7	0,7	-1	-2
PCAES m=2.0	1,9	2,4	3,0	3,2	3,5	3,3	3,4	3,1	2,7	2,4	1,7	2,1	1,9	2,6	1,0	-0,	-2,	-3,

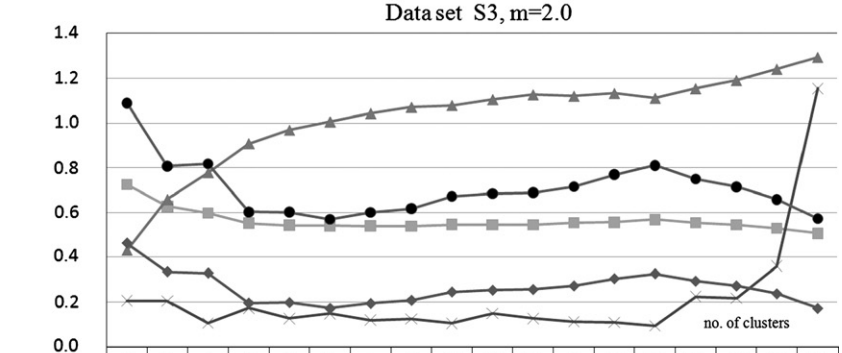
Fig. 11. Variation of validity indices for *S*1 data set, for $c = 2,\ldots,19$ and for different degree of fuzziness $m = 2.0$ (a, c) and $m = 1.3$ (b, c).

Gaussian functions having different means and variance. The scatter plots of all three data sets are shown in Figs. 5–7. We applied *FCM* algorithm to these three data sets separately using fuzziness $m=2.0$ and 10 different number of clusters, $c=2,\ldots,11$ and calculated values of CO and CO_r indices. For all results we used parameters $T_o=0.4$ and $T_c=0.6$.

4.1.1. Example 1

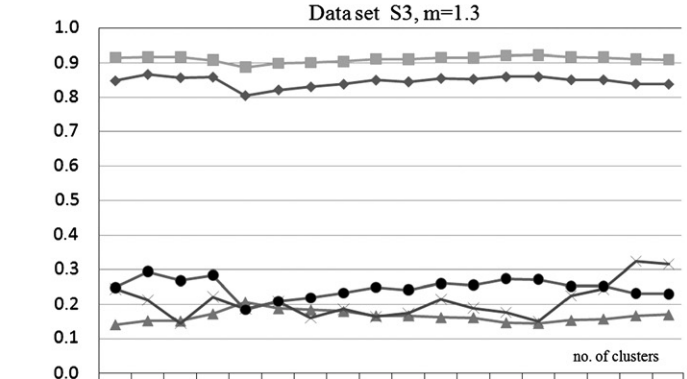
Firstly, we generated data set *A*0 from 450 data objects that form three clusters with different sizes, as shown in Fig. 5. One cluster contained 50 data objects and other two contained 200 data objects each. We used the fuzzines degree $m=2.0$. The reasonable number of clusters is three. Proposed validity indices CO and CO_r identified

a



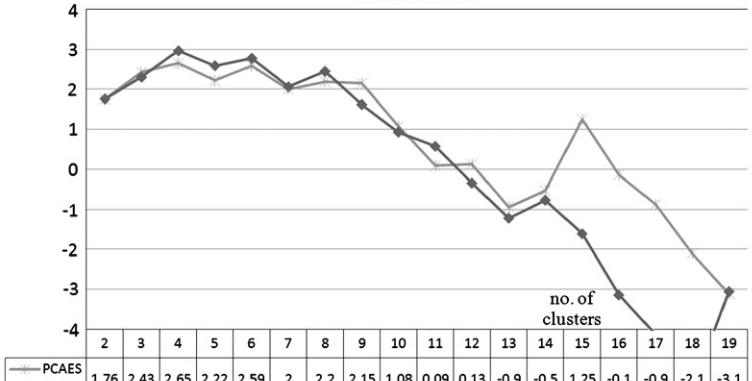
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
PC	0,725	0,627	0,598	0,552	0,544	0,541	0,54	0,54	0,546	0,545	0,545	0,554	0,556	0,569	0,555	0,546	0,53	0,507
PE	0,431	0,659	0,778	0,908	0,969	1,005	1,043	1,07	1,078	1,105	1,126	1,12	1,132	1,111	1,154	1,191	1,241	1,294
XB	0,206	0,205	0,105	0,173	0,126	0,148	0,12	0,124	0,105	0,15	0,127	0,112	0,11	0,093	0,224	0,217	0,361	1,154
CO	0,463	0,336	0,33	0,197	0,199	0,173	0,195	0,209	0,246	0,254	0,257	0,273	0,305	0,327	0,294	0,273	0,238	0,173
Cor * 100	1,088	0,807	0,816	0,602	0,601	0,570	0,600	0,617	0,671	0,684	0,688	0,715	0,768	0,810	0,749	0,714	0,658	0,572

b



	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
PC	0,91	0,92	0,92	0,91	0,89	0,9	0,9	0,9	0,91	0,91	0,91	0,91	0,92	0,92	0,92	0,92	0,91	0,91
PE	0,14	0,15	0,15	0,17	0,21	0,19	0,18	0,18	0,17	0,17	0,16	0,16	0,15	0,14	0,15	0,16	0,17	0,17
XB	0,24	0,21	0,15	0,22	0,19	0,21	0,16	0,19	0,16	0,17	0,22	0,19	0,18	0,15	0,22	0,24	0,32	0,32
CO	0,85	0,87	0,86	0,86	0,8	0,82	0,83	0,84	0,85	0,85	0,86	0,85	0,86	0,86	0,85	0,85	0,84	0,84
Cor*100	0,25	0,29	0,27	0,28	0,19	0,21	0,22	0,23	0,25	0,24	0,26	0,26	0,27	0,27	0,25	0,25	0,23	0,23

c



	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
PCAES m=1.3	1,76	2,43	2,65	2,22	2,59	2	2,2	2,15	1,08	0,09	0,13	-0,9	-0,5	1,25	-0,1	-0,9	-2,1	-3,1
PCAES m=2.0	1,76	2,3	2,96	2,59	2,77	2,06	2,45	1,62	0,93	0,58	-0,3	-1,2	-0,8	-1,6	-3,1	-4,1	-5,5	-3

Fig. 12. Variation of validity indices for *S*3 data set, for $c=2,\ldots,19$ and for different degree of fuzziness $m=2.0$ (a, c) and $m=1.3$ (b, c).

three clusters, but all other indices identified only two clusters (Fig. 5b and c). For most observed indices maximal value defines the optional number of clusters, except for *XB* and *PE*. Only proposed validity indices *CO* and CO_r (Fig. 5c) correctly identified three clusters in 1D example with one small and two bigger clusters in Fig. 1, but all other indices identified only two clusters.

For the data set *A*0 we changed parameters T_o and T_c: $T_o=0.4$ and $T_c=0.6$ (Fig. 6c), $T_o=0.6$ and $T_c=0.3$ (Fig. 6b) and $T_o=0.9$ and $T_c=0.1$ (Fig. 6a). Only for parameters $T_o=0.9$ and $T_c=0.1$ both indices *CO* and CO_r identified only two clusters as optimal number and small cluster is ignored.

4.1.2. Example 2

For the second artificial data set *A*1, 450 data objects were generated forming three clusters of equal size but different densities, shown in Fig. 7a. This data set was used for testing the efficiency of indices in discovering clusters with different densities. Two clusters contained 200 data objects and one cluster contained only 50 data objects. The performances for each validity index are given in Fig. 7b. Proposed indices *CO* and CO_r, *XB* and *PC* identified three clusters, while *PE* and *PCAES* identified two.

4.1.3. Example 3

Fig. 8a shows the third generated Gaussian-mixture data set *A*2. This data set consists of 590 data objects and forms five clusters that differ in size and density. Fig. 8b shows the results obtained by various validity indices for numbers of clusters $c=2,3,\ldots,10$. Only the proposed CO_r and *CO* indices and *XB* index identified all five clusters. *PCEAS* index identified three clusters and *PC* and *PE* index found two clusters as optimal number.

The results of all three different generated data sets (Figs. 5b, 7b, and 8b) indicate that both new indices can efficiently identify different models in data sets containing clusters of different sizes and densities.

In all of the three artificial data sets the new validity index has its maximum at the reasonably number of clusters c^*. The validity measure converges to 0 as $c > c^*$. In all data sets the proposed index showed asymptotical behaviour towards the larger number of clusters.

4.2. Known 2D data sets

In order to evaluate the performance of the proposed index we used three known 2D data sets: STARFIELD [11] and *S*1 and *S*3 data sets [28]. We applied *FCM* algorithm to these three data sets separately using two different degrees of fuzziness $m=1.3$ and 2.0 and calculated values of *CO* and CO_r index for each combination.

4.2.1. Example 4

Fig. 9a shows the STARFIELD data set, which contains 75 two-dimensional elements. Nine clusters is a reasonably optimal partition. Fig. 10 shows the results of validity index for $c=2,\ldots,11$ and for fuzziness $m=2.0$ and 1.3. Among the indices considered, only *CO* and CO_r correctly recognized the presence of nine clusters for both fuzziness coefficients. For fuzziness coefficient $m=2.0$, *PC* and *PE* considered two clusters to be a natural structure, *XB* recognized six clusters as the optimal partition and *PCEAS* four clusters.

4.2.2. Example 5

Fig. 9b and c show synthetic 2D data sets *S*1 and *S*3 with 5000 vectors and 15 Gaussian clusters with different degrees of cluster overlap. The degree of overlap in data set *S*3 is greater than in data set *S*1 (Fig. 9b and c). For the data set *S*1 and fuzziness $m=2.0$ all indices except *PE* recognized 15 clusters (Fig. 11). For the data set *S*3, (Fig. 12) all indices except *XB* recognized two clusters as the most reasonable cluster number and then 15 (the second or the third maximal or minimal value).

Very small validity index values denotes that the degree of overlap among clusters is high for all numbers of clusters 1,…,19. We decreased T_c from 0.8 to 0.6 and T_o from 0.2 to 0.1. The values of index was higher, but the result was the same: maximal values of index showed optimal number of clusters 2, 3, and then 15 (Fig. 13).

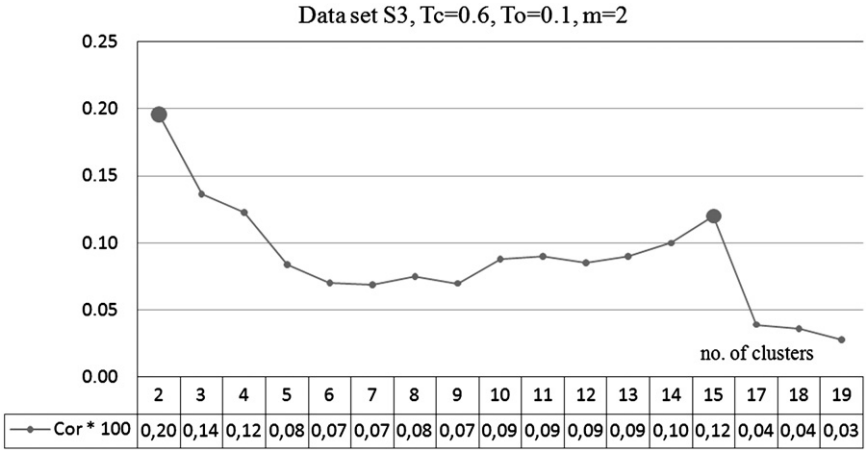
The new validity indices are not affected from changing values of fuzziness criteria $m=1.3$ and 2.

4.3. Real data sets

In addition, we used three real data sets from UCI repository [27] to demonstrate the performance of the proposed indices: Iris data set, Glass and Ionosphere data set.

4.3.1. Example 6: Iris data set

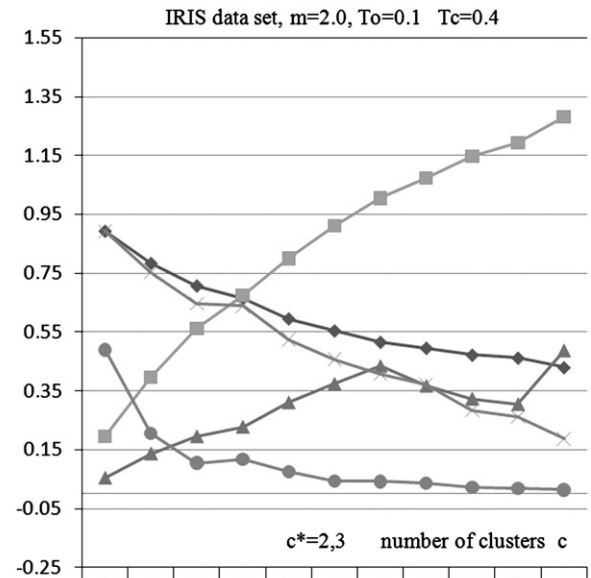
First used real data set is well-known IRIS data set [29] containing data from three types of iris (iris virginica, iris setose, and iris versicolor) of 150 data objects described by four dimensions (features). The four feature values represent the sepal length, sepal width, petal length and the petal width in centimetres. Two clusters are the optimal choice in view of the geometric structure of Iris data, where two clusters overlap



	2	3	4	5	6	7	8	9	10	11	12	13	14	15	17	18	19
Cor * 100	0,20	0,14	0,12	0,08	0,07	0,07	0,08	0,07	0,09	0,09	0,09	0,09	0,10	0,12	0,04	0,04	0,03

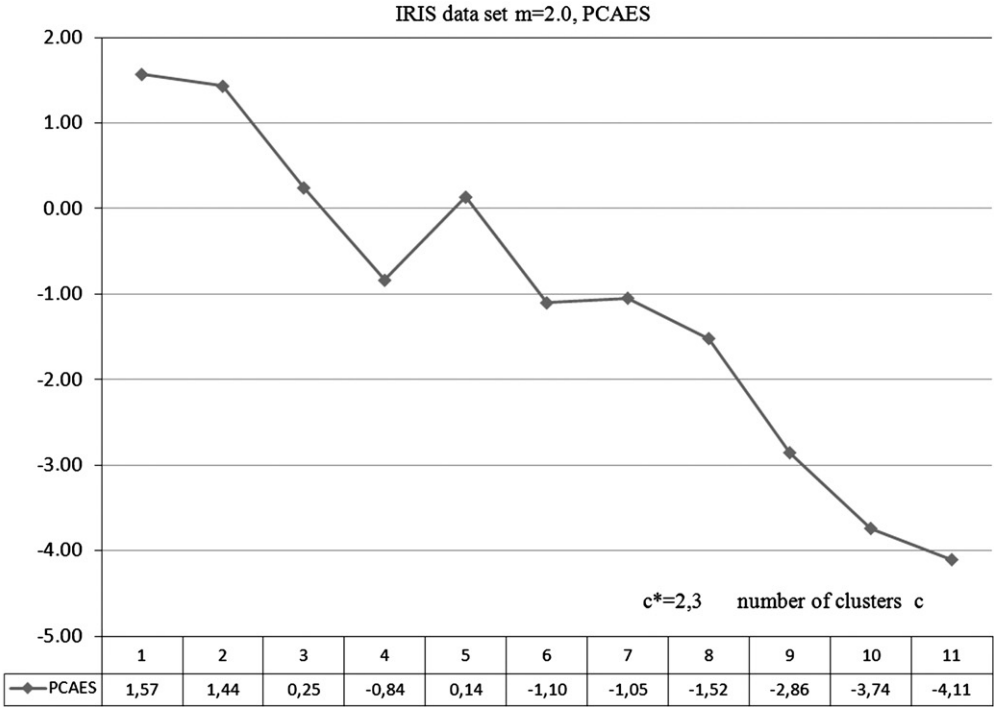
Fig. 13. Variation of validity index CO_r for *S*3 data set, for $c=2,3,\ldots,19$ and for fuzziness $m=2.0$.

a



	2	3	4	5	6	7	8	9	10	11	
PC	0,89	0,78	0,71	0,67	0,60	0,55	0,52	0,50	0,47	0,46	0,43
PE	0,20	0,40	0,56	0,68	0,80	0,91	1,01	1,07	1,15	1,19	1,28
XB	0,05	0,14	0,20	0,23	0,31	0,37	0,43	0,37	0,32	0,31	0,49
CO	0,89	0,75	0,65	0,64	0,53	0,46	0,41	0,37	0,29	0,26	0,19
COr	0,49	0,21	0,11	0,12	0,08	0,04	0,04	0,04	0,02	0,02	0,01

b



	1	2	3	4	5	6	7	8	9	10	11
PCAES	1,57	1,44	0,25	-0,84	0,14	-1,10	-1,05	-1,52	-2,86	-3,74	-4,11

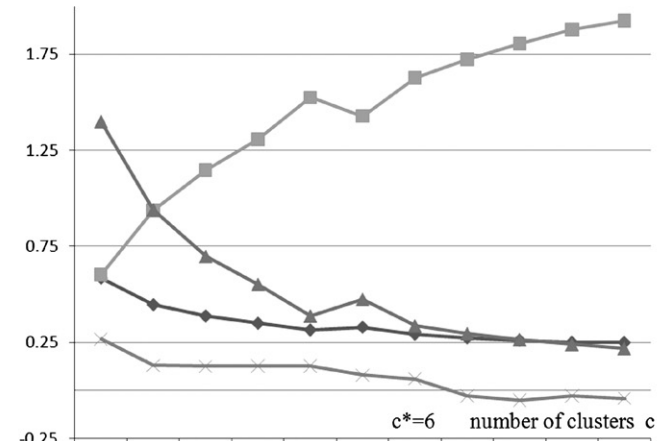
Fig. 14. Values of considered validity indices for different number of clusters (2,3,…,11) for Iris data set.

significantly, as mentioned by Pal and Bezdek [30]. Iris data set consists of three clusters. Two clusters have substantial overlapping. The proposed indices *CO* and *COr* estimated $c^*=2$ as an optimal cluster number, but they showed $c^*=3$ (Fig. 14) is second good cluster number estimate. Overall, most validity indices gave the optimal cluster number estimate $c^*=2$ or 3 for the Iris data set. There are small decreasing of CO_r validity index when c^* is greater than 3. Therefore, the CO_r index offers the information that $c^*=3$ is a good cluster number estimate for Iris data set.

4.3.2. Example 7: Glass data set

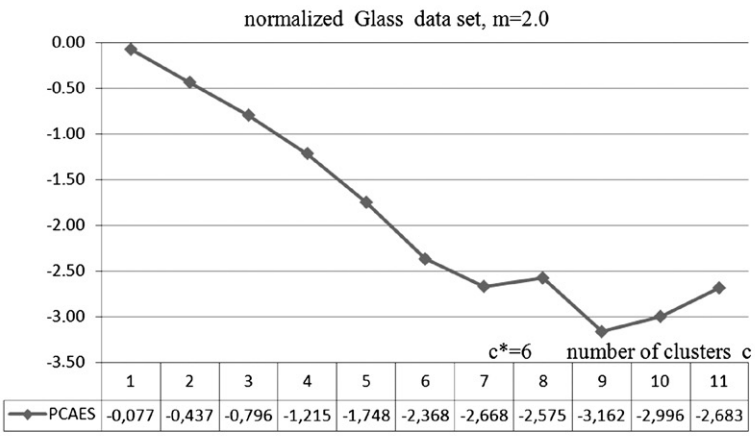
The Glass data set [27] has $n=214$ data objects in 9 dimensional space. It consists of six clusters. We normalized the

a



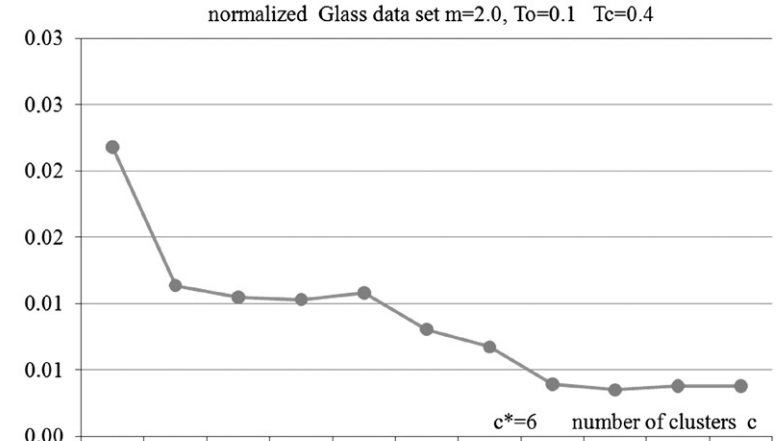
	2	3	4	5	6	7	8	9	10	11	12
PC	0,585	0,446	0,387	0,350	0,314	0,329	0,293	0,274	0,261	0,250	0,249
PE	0,602	0,936	1,146	1,307	1,524	1,428	1,628	1,722	1,807	1,879	1,924
XB	1,400	0,940	0,698	0,552	0,388	0,475	0,336	0,296	0,265	0,239	0,218
CO	0,267	0,130	0,125	0,126	0,126	0,079	0,060	-0,03	-0,05	-0,03	-0,04

b



	1	2	3	4	5	6	7	8	9	10	11
PCAES	-0,077	-0,437	-0,796	-1,215	-1,748	-2,368	-2,668	-2,575	-3,162	-2,996	-2,683

c



	2	3	4	5	6	7	8	9	10	11	12
COr	0,022	0,011	0,010	0,010	0,011	0,008	0,007	0,004	0,004	0,004	0,004

Fig. 15. Values of considered validity indices for different number of clusters (2,3,…,11) for Glass data set.

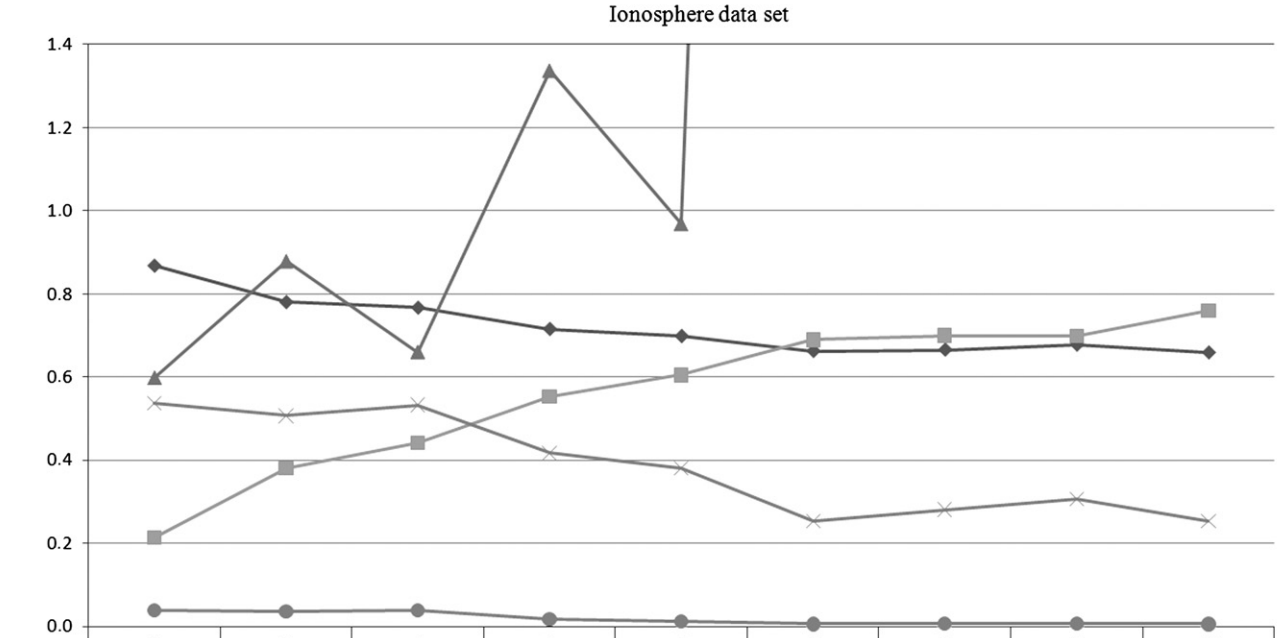
glass data set as in [12] so that all attributes have equal range. Because the clusters are heavily overlapped, it is difficult to perform good cluster number estimates. The optimal cluster number estimate for this data set obtained by *PC*, *PE* and *PCAES* is 2. Others present the monotonic tendency of the cluster number c. There is a large decreasing in CO_r and *CO* index when $c^*=6$ and therefore the index offers the information that $c^*=6$ is a good cluster number estimate for the Glass data set (Fig. 15).

4.3.3. Example 8: Ionosphere data set

In addition, we used Ionosphere data set from UCI repository [27] to demonstrate the performance of new indices. This is binary classification data set of 351 instances collected by radar with 34 attributes. Both indices CO_r and CO identified four clusters as optimum number of clusters and were able to identify two small clusters, but all other considered indices identified two clusters (Fig. 16).

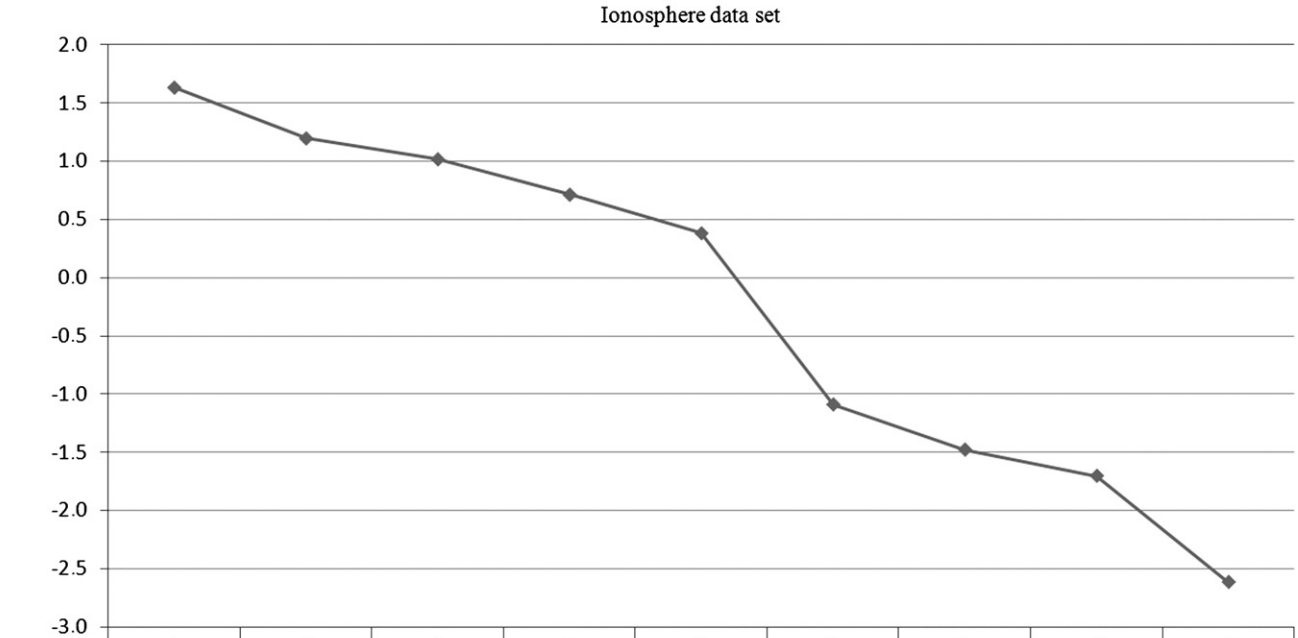
In this experiment, we demonstrated that the proposed indices CO and CO_r could actually identify the number of clusters of real data sets. In all nine tested data sets, the validity indices CO and CO_r demonstrate the stability to provide the correct number of clusters.

a



	2	3	4	5	6	7	8	9	10
PC	0,869	0,781	0,767	0,715	0,698	0,662	0,665	0,678	0,658
PE	0,215	0,381	0,442	0,553	0,605	0,690	0,700	0,699	0,760
XB	0,598	0,879	0,660	1,338	0,969	8,985	6,248	5,504	19,872
CO	0,537	0,507	0,533	0,417	0,381	0,253	0,281	0,306	0,254
COr	0,039	0,037	0,039	0,018	0,012	0,006	0,007	0,007	0,006

b



	1	2	3	4	5	6	7	8	9
PCAES	1,632	1,195	1,016	0,712	0,381	-1,090	-1,479	-1,704	-2,616

Fig. 16. Values of considered validity indices for different number of clusters (2,…,11) for Ionosphere data set.

The proposed indices identify the correct number of clusters over a wide range of parameters T_o and T_c. Both indices give similar results for T_c within the wide range 0.8—0.2 and for T_o from 0.4 to 0.1. The values of proposed validity index CO changes little with the change of parameters T_o and T_c, but the CO_r validity index usually identifies optional cluster number with bigger maximum.

5. Conclusions

The cluster validity index estimates the quality of clustering and tells us which partition is the best. It can be used when searching for an optimal number of clusters or when the number of clusters is unknown before clustering. Most of the existing cluster validity indices depend too much on average values and clusters' centres. This causes incorrect evaluation of partitions containing clusters that widely differ in size and density.

In this paper, two new cluster validity criterions are introduced that does not suffer from the drawbacks of traditional measures for the compactness and separation used in the most cluster validity indices. It measures two properties: compactness and overlap. The smaller the degree of overlap, the more separated are clusters. Both measures are calculated from membership degrees. A good fuzzy partition is expected to have small overlap and great compactness. The overlap measure describes the degree of overlap among all fuzzy clusters. It is calculated from membership values of all data objects that are strong enough belong to two or more fuzzy clusters. Compactness is obtained from membership values of data objects that strong enough belong to only one fuzzy cluster. We propose ratio and summation type of validity index based on the same compactness and overlap measures. Thus optimal fuzzy partition is obtained by maximizing the CO and CO_r index with respect to the number of clusters. The performance of the proposed indices was tested on various data sets demonstrating its validity and efficiency. The results indicate that for different fuzziness values of the experimental tests and for artificial and real data sets the new validity indices were more efficient than the rest of considered validity indices especially when identifying clusters that widely differ in density or size.

References

[1] L. Kaufman, P.J. Rousseeuw, in: Finding Groups in DataWiley, New York, 1990.
[2] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.
[3] F. Hoppner, F. Klawon, R. Kruse, T. Runkler, Fuzzy Cluster Analysis: Methods for Classifications, Data Analysis and Image RecognitionWiley, New York, 1999.
[4] A. Devillez, P. Billaudel, G.V. Lecolier, A fuzzy hybrid hierarchical clustering method with a new criterion able to find the optimal partition, Fuzzy Sets Syst. 128 (3) (2002) 323–338.
[5] H. Frigui, R. Krishnapuram, A robust algorithm for automatic extraction of an unknown number of clusters from noisy data, Pattern Recognition Lett. 17 (1996) 1223–1232.
[6] E.R. Hruschka, R.J.G.B. Campello, L.N. de Castro, Evolutionary search for optimal fuzzy c-means clustering, in: Proceedings of the 13th IEEE International Conference on Fuzzy Systems, Budapest, Hungary, 2004, pp. 685–690.
[7] R. Kothari, D. Pitts, On finding the number of clusters, Pattern Recognition Lett. 20 (1999) 405–416.
[8] G.V. Milligan, M.C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (2) (1985) 159–179.
[9] J.C. Bezdek, Numerical taxonomy with fuzzy sets, J. Math. Biol. 1 (1974) 57–71.
[10] J.C. Bezdek, Cluster validity with fuzzy sets, J. Cybern. 3 (1974) 58–78.
[11] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern. Anal. Mach. Intell. 13 (8) (1991) 841–847.
[12] K.-L. Wu, M.-S. Yang, A cluster validity index for fuzzy clustering, Pattern Recognition Lett. 26 (2005) 1275–1291.
[13] M. Kim, R.S. Ramakrishna, New indices for cluster validity assessment, Pattern Recognition Lett. 26 (15) (2005) 2353–2363.
[14] Y. Zhang, W. Wang, X. Zhang, Li Yi, A cluster validity index for fuzzy clustering, Inform. Sci. 178 (4) (2008) 1205–1218.
[15] A. Celikyilmaz, I.B. Türkşen, Validation criteria for enhanced fuzzy clustering, Pattern Recognition Lett. 29 (2) (2008) 97–108.
[16] W. Wang, Y. Zhang, On fuzzy cluster validity indices, Fuzzy Sets Syst. 158 (19) (2007) 2095–2117.
[17] M. Bouguessa, S. Wang, H. Sun, An objective approach to cluster validation, Pattern Recognition Lett. 27 (13) (2006) 1419–1430.
[18] A.K. Jain, R.C. Dubes, in: Algorithms for Clustering DataPrentice-Hall, Englewood Cliffs, New York, 1988.
[19] J.C. Bezdek, in: Pattern Recognition with Fuzzy Objective Function AlgorithmPlenum Press, New York, 1981.
[20] J. Hartigan, in: Clustering AlgorithmsWiley, New York, 1975.
[21] D.E. Gustafson, W. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: Proceedings of the IEEE Conference on Decision Control, San Diego, CA, 1979, pp. 761–766.
[22] R. Krishnapuram, J. Kim, A note on the Gustafson–Kessel and adaptive fuzzy clustering algorithms, IEEE Trans. Fuzzy Syst. 7 (1999) 453–461.
[23] K.L. Wu, M.S. Yang, Alternative c-means clustering algorithms, Pattern Recognition 35 (2002) 2262–2278.
[24] J.C. Bezdek, in: Pattern Recognition with Fuzzy Objective Function AlgorithmPlenum Press, New York, 1981.
[25] J. MacQueen, Some methods for classification and analysis multivariate observations, in: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1967, 1-281-297.
[26] E.H. Ruspini, Numerical methods for fuzzy clustering, Inform. Sci. 2 (1970) 319–350.
[27] C.L. Blake, C.J. Merz, 1998. UCI repository of machine learning databases, a huge collection of artificial and realworld data sets. ⟨http://www.ics.uci.edu/~mlearn/MLRepository.html⟩.
[28] P. Fränti, O. Virmajoki, Iterative shrinking method for clustering problems, Pattern Recognition 39 (5) (2006) 761–765.
[29] J.C. Bezdek, Numerical taxonomy with fuzzy sets, J. Math. Biol. (1974) 57–71.
[30] N.R. Pal, J.C. Bezdek, On cluster validity for the fuzzy c-means model, IEEE Trans. Fuzzy Syst. 3 (1995) 370–379.

Krista Rizman Žalik received the M.Sc. and Ph.D. degrees from University of Maribor in 1990 and 1993, respectively, all in computer science. She is currently an Assistant Professor at University of Maribor. Her primary research interests include pattern recognition, data mining information retrieval and databases.