Pattern Anal Applic (2017) 20:687–700
DOI 10.1007/s10044-015-0525-8

THEORETICAL ADVANCES

A new validity index for crisp clusters

Artur Starczewski[1]

Received: 21 November 2014 / Accepted: 14 November 2015 / Published online: 28 November 2015

Abstract In this paper, a new cluster validity index which can be considered as a measure of the accuracy of the partitioning of data sets is proposed. The new index, called the STR index, is defined as the product of two components which determine changes of compactness and separability of clusters during a clustering process. The maximum value of this index identifies the best clustering scheme. Three popular algorithms have been applied as underlying clustering techniques, namely complete-linkage, expectation maximization and *K*-means algorithms. The performance of the new index is demonstrated for several artificial and real-life data sets. Moreover, this new index has been compared with other well-known indices, i.e., Dunn, Davies-Bouldin, PBM and Silhouette indices, taking into account the number of clusters in a data set as the comparison criterion. The results prove superiority of the new index as compared to the above-mentioned indices.

Keywords Clustering · Validity index · Unsupervised classification

1 Introduction

Clustering is named as unsupervised learning or unsupervised classification which uses unlabelled patterns and where structural information about data is not available. In this process data is partitioned into homogeneous subsets (called clusters), inside which elements are similar to each other while being different from items in other groups. In many clustering methods clusters are represented by their centers.

Nowadays, a large number of clustering algorithms exist having found use in various fields such as data mining, bioinformatics, exploration data, etc. In general, these algorithms can be classified into two basic categories, i.e., partitional and hierarchical methods [8]. The first-group methods provide one-level partitioning data, and the well-known algorithms of this type are, e.g., *K*-means and its variations [5, 21] or expectation maximization (EM) [15]. The second category of methods comprises multi-level partitioning data, and the representative examples of such algorithms are hierarchical agglomerative approaches such as single-linkage, complete-linkage or average-linkage [12, 16, 22]. However, these algorithms are seldom used for large sets since their computational complexity is high [29]. It should be noted that the results of partitioning of the same data may be different when input parameters of the clustering algorithm vary within a certain range. The significant input parameter of many clustering algorithms is a number of clusters, which is often selected in advance. Thus, the key issue is how to properly evaluate results of data clustering. There are three techniques which can be used to evaluate partitioning of data sets, namely, external, internal or relative approaches [12, 25]. The first two techniques are based on statistical testing, and their computational demands are high. On the other hand, the relative methods perform the comparison of partitioning schemes obtained by a clustering algorithm using different values of input parameters multiple times. Then, cluster validity indices are used to find the best partitioning of data. A great number of such indices have been introduced, e.g., [3, 9, 10, 13, 17, 26, 28, 30, 31]. In many validity indices two properties of clusters are taken into account,

✉ Artur Starczewski
starcz@kik.pcz.czest.pl

[1] Institute of Computational Intelligence, Częstochowa University of Technology, Al. Armii Krajowej 36, 42-200 Czestochowa, Poland

i.e., compactness and separability [11]. The first property is associated with the within-cluster spread, and the second with the inter-cluster separation. Validity indices are most often a ratio of a measure of cluster compactness to cluster separation or vice versa. They can also be the sum or the product of these measures. Then, according to the type of the validity indices, the right partitioning of a data set is associated with the maximum or minimum value of the validity index. In the literature well-known cluster validity indices such as, e.g., Dunn [7], Davies-Bouldin (DB) [6], PBM [18] or Silhouette (SIL) indices [23] are frequently used when comparing results of different clustering techniques. The Dunn index is the ratio of the minimum inter-cluster distance to the maximum cluster diameter. In turn, the Davies-Bouldin (DB) index is the ratio of the sum of the within-cluster scatter to the inter-cluster separation. On the other hand, the PBM index is a composition of three factors, namely, the number of clusters, the measure of cluster compactness and the measure of cluster separation. It is proposed to be used to form a small number of compact clusters. The silhouette (SIL) index is the mean of the means of so-called silhouettes through all the clusters. Recently, numerous new interesting solutions have been proposed for cluster evaluation. For example, paper [14] presents a new validity index for crisp clustering, which emphasizes the cluster shape by using a high order characterization of its probability. In turn, to represent the separation among clusters a new measure called dual center is proposed in [27]. A new measure of connectivity is presented in [24]. This measure is based on the concept of the relative neighborhood graph. Proposed new indices are able to automatically detect clusters of any shape and size. In turn, the stability index based on the variation on some information measures over the partitions generated by a clustering model is proposed in [20]. Moreover, in paper [32] the authors note that the knee point detection is often required because most indices show monotonicity with an increasing number of clusters. Thus, indices with a clear minimum or maximum value are preferred. They present an index called the *WB* index. However, it should be noted that existing validity indices have limitations and lack generalization in evaluation of clustering results [1].

In this paper, a new cluster validity index called the STR index is proposed and its maximum value indicates the best partitioning of the data set for non-overlapping clusters. Unlike most indices, this proposed approach uses the knee point detection, and so a maximum value of the index is very clear. It consists of the product of two components, which determine changes of compactness and separability of clusters in partitioning schemes [see Eq. (18)]. It should be noted that values of these changes have different ranges, but do not need to be normalized because they are multiplied. In order to present effectiveness of the new validity index several experiments were performed for different data sets. This paper is organized as follows: Sect. 2 presents an overview of several well-known validity indices. Section 3 describes the new validity index and the basic dependencies referring to cluster properties. Section 4 illustrates experimental results on artificial and real-life data sets. Finally, Sect. 5 presents conclusions.

2 Chosen popular validity indices

Nowadays, in the clustering literature there is a large number of various validity indices. Some of them are very well known and are often used for comparing with other indices. Among them are those mentioned above, i.e., Dunn, Davies-Bouldin (DB), PBM and Silhouette (SIL) indices. Below, their detailed description is presented.

Dunn index This index is expressed as:

$$D = \min_{1 \le i \le K} \left(\min_{1 \le j \le K, i \ne j} \left(\frac{d(C_i, C_j)}{\max_{1 \le k \le K} (\delta(C_k))} \right) \right) \tag{1}$$

where K is a number of clusters in a data set, $\delta(C_k)$ is the diameter of cluster C_k, i.e., the largest distance between two points within the cluster, and $d(C_i, C_j)$ is the minimum distance between two clusters C_i and C_j, which is calculated as a distance between the closest points from these two clusters. For well-separable clusters, distances between clusters are large and their diameter is small. Thus, the maximum value of the index indicates the right partitioning of data.

Davies–Bouldin (DB) index This index is defined as the ratio of the sum of the within-cluster scatter to the inter-cluster separation and can be expressed as follows:

$$\mathrm{DB} = \frac{1}{K} \sum_{i=1}^{K} R_i \tag{2}$$

where the factor R_i can be written as:

$$R_i = \max_{j \ne i} \frac{S_i + S_j}{d_{ij}} \tag{3}$$

S_i and S_j denote the within-cluster scatter for i_{th} and j_{th} clusters, respectively, and, e.g., S_i can be expressed as follows:

$$S_i = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{v}_i\| \tag{4}$$

where n_i is a number of $\mathbf{x}$ in the cluster C_i, and $\mathbf{v_i}$ is the center of this cluster. Moreover, the d_{ij} is the distance between the cluster centers, i.e., $d_{ij} = \|\mathbf{v}_i - \mathbf{v}_j\|$. The minimum of the *DB* index indicates the appropriate partitioning of a data set.

PBM index This index is defined as follows:

$$\mathrm{PBM} = \left(\frac{1}{K} \times \frac{E_o}{E} \times D\right)^2 \tag{5}$$

where E identifies the total within-cluster scatter, such that

$$E = \sum_{k=1}^{K} \sum_{j=1}^{n} \mu_{kj} \left\| \mathbf{x}_j - \mathbf{v}_k \right\| \tag{6}$$

and n is a number of elements in the data set, $\mathbf{U} = [\mu_{kj}]$ is a partition matrix of the data, and $\mathbf{v}_k$ is the center of the cluster C_k. On the other hand, the factor E_0 represents total scatter of all patterns belonging to one cluster in the given data set. It is expressed as follows:

$$E_0 = \sum_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{v}\| \tag{7}$$

where $\mathbf{v}$ is the center of patterns $\mathbf{x} \in X$. The next factor—D—is a measure of cluster separation. It is defined as a maximum distance between cluster centers:

$$D = \max_{i,j=1}^{K} \left\| \mathbf{v}_i - \mathbf{v}_j \right\| \tag{8}$$

The maximum value of the index corresponds to the best partitioning of a given data set.

Silhouette (SIL) index This index can be defined as:

$$\mathrm{SIL} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{SIL}(C_k) \tag{9}$$

where $\mathrm{SIL}(C_k)$ is the Silhouette width for the given cluster C_k and can be expressed as follows:

$$\mathrm{SIL}(C_k) = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \mathrm{SIL}(\mathbf{x}) \tag{10}$$

where n_k is a number of patterns in C_k, and $\mathrm{SIL}(\mathbf{x})$ is the Silhouette width for the pattern $\mathbf{x}$ and can be written as:

$$\mathrm{SIL}(\mathbf{x}) = \frac{b(\mathbf{x}) - a(\mathbf{x})}{max(a(\mathbf{x}), b(\mathbf{x}))} \tag{11}$$

$a(\mathbf{x})$ is the within-cluster mean distance and it is defined as the average distance between $\mathbf{x}$ and the rest of the patterns belonging to the same cluster, $b(\mathbf{x})$ is the smallest of the mean distances of $\mathbf{x}$ to the patterns belonging to the other clusters. The maximum of the SIL index provides the best partitioning of a data set. It needs to be noted that unlike the above-mentioned indices, it can be used for clusters of arbitrary shapes.

3 The new validity index

First, the definition of the index is presented, and next the role of its components and interactions between them are explained in detail.

Let us denote a data set by $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where n is a number of patterns. Moreover, let C_k indicate kth cluster, where $k = 1, \ldots, K$. Notice that in the given data set the number of clusters K is limited by the number of patterns n. Measure of cluster compactness can be expressed as the ratio of the total scatter of all patterns to the total scatter of the within clusters. Thus, for the K partition scheme, it is defined as follows:

$$E(K) = \frac{E_0}{E_K} \tag{12}$$

Here, E_0 denotes the total scatter of all patterns of X and is expressed as:

$$E_0 = \sum_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{v}\| \tag{13}$$

where $\mathbf{v}$ is the center of the data set X. Whereas, E_K is the total scatter of the within clusters, such that

$$E_K = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} \|\mathbf{x} - \mathbf{v}_k\| \tag{14}$$

and $\mathbf{v}_k$ is the center of the kth cluster. In turn, the measure of cluster separation can be defined as the ratio of the maximum to the minimum distance between cluster centers and can be written as:

$$D(K) = \frac{D_{K\max}}{D_{K\min}} \tag{15}$$

and

$$D_{K\max} = \max_{i,k=1}^{K} \|\mathbf{v}_i - \mathbf{v}_k\| \tag{16}$$

$$D_{K\min} = \min_{i,k=1}^{K} \|\mathbf{v}_i - \mathbf{v}_k\| \tag{17}$$

where $\mathbf{v}_i$ and $\mathbf{v}_k$ are the centers of the ith and kth clusters.

Based on these measures of cluster properties (Eqs. 12 and 15), the new validity index, called the STR index, is defined as:

$$\mathrm{STR} = [E(K) - E(K-1)] \cdot [D(K+1) - D(K)] \tag{18}$$

where $E(K-1)$ is the measure of compactness of clusters calculated for the $K-1$ partition scheme, that is:

$$E(K-1) = \frac{E_0}{E_{K-1}} \tag{19}$$

and

$$E_{K-1} = \sum_{k=1}^{K-1} \sum_{\mathbf{x} \in C_k} \|\mathbf{x} - \mathbf{v}_k\| \tag{20}$$

while $D(K+1)$ is the measure of cluster separation calculated for the $K+1$ partition configuration and is expressed as follows:

$$D(K+1) = \frac{D_{(K+1)\max}}{D_{(K+1)\min}} \tag{21}$$

and

$$D_{(K+1)\max} = \max_{i,k=1}^{K+1} \|\mathbf{v}_i - \mathbf{v}_k\| \tag{22}$$

$$D_{(K+1)\min} = \min_{i,k=1}^{K+1} \|\mathbf{v}_i - \mathbf{v}_k\| \tag{23}$$

To determine the proper partitioning of a data set, the maximum value of the STR index is found (see Eq. 18).

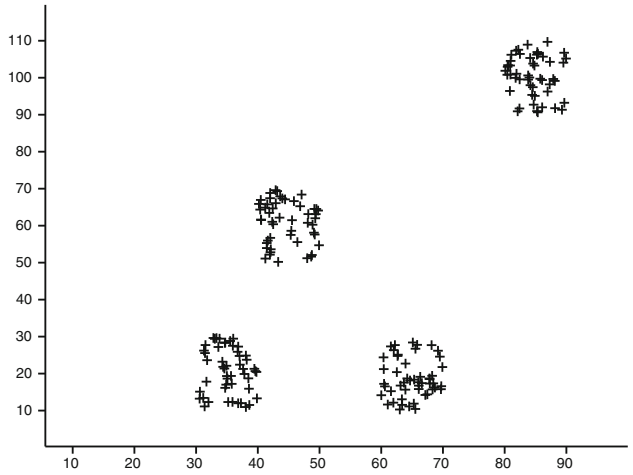


Fig. 1 An example of a data set consisting of four clusters

3.1 Detailed explanation

Let us denote by c^* the actual number of clusters present in a data set X. For instance, Fig. 1 shows an example of a data set consisting of four clusters $c^* = 4$, which contain 50 instances per class. In order to demonstrate changes of compactness and separability of clusters, for these data a partitioning process was carried out using the complete-linkage clustering algorithm. The number of clusters K varied from 12 to 1, and the variation of the STR index factors is presented in Fig. 2.

It should be noted that when $K > c^*$, the compact clusters are subdivided into smaller ones. Thus, the scatter of the patterns in individual clusters becomes small and

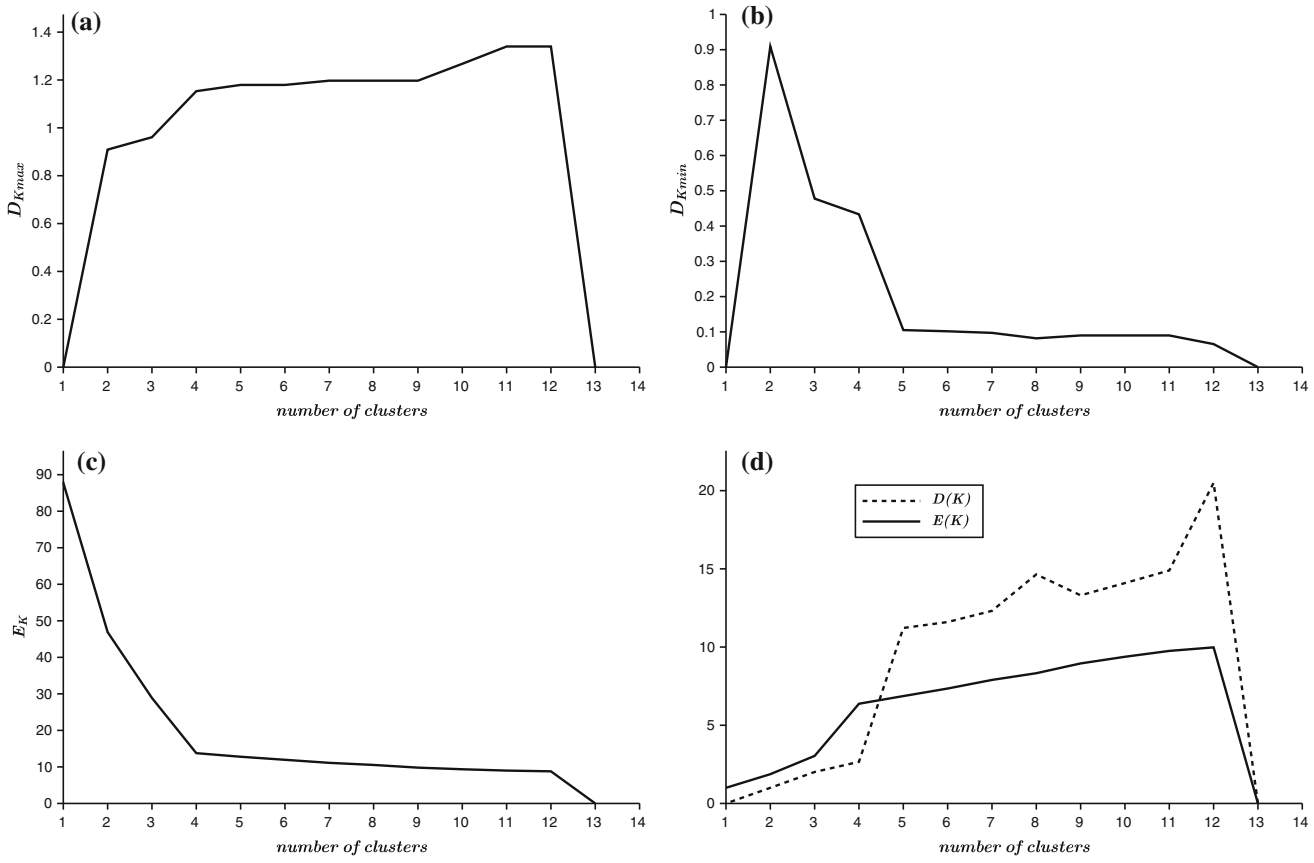


Fig. 2 Variation of the STR index factors with respect to the number of clusters for the example data set: **a** $D(K)_{\max}$, **b** $D(K)_{\min}$, **c** $E(K)$, **d** D_K and E_K

then the compactness of these new clusters does not change so much. Similarly, if $K = c^*$, the within-cluster scatter is small because the data consist of compact clusters with a small spread of patterns. On the other hand, when $K < c^*$, individual clusters are merged into larger ones and the total scatter of the within clusters increases significantly. This means that abrupt changes of compactness occur when the number of clusters varies from $K = c^*$ to $K = c^* - 1$.

In the proposed new index the measure of compactness of clusters is expressed by $E(K)$ (see Eq. 12). For example, Fig. 2c shows variation of E_K (denominator of the $E(K)$) with respect to the number of clusters. As it can be observed, E_K increases abruptly when the number K varies from c^* to $c^* - 1$. This phenomenon forms the knee point at the number of clusters $c^* = 4$. Notice that the factor $E(K)$ will also have the knee point because it is the ratio of E_0 to E_K, where E_0 is constant. However, the behavior of this factor around c^* is inverse, that is, it is large for $K = c^*$, and it is small for $K = c^* - 1$ (see Fig. 2d). Of course, when K equals 1, the value of $E(K)$ is 1. Thus, changes of compactness of clusters shown by $E(K)$ are greatest between $K = c^*$ and $K = c^* - 1$ partition schemes. Therefore, the difference between $E(K)$ and $E(K-1)$ is used by the new index (see Eq. 18) to determine these changes of compactness.

The second property of clusters is their separability. Measure of this property can be defined in different ways, for example, as a minimum distance between clusters. But the key issue is to find the knee point when the number K varies from $K_{\max}$ to $K_{\min}$. It should be noted that when the number of clusters $K > c^*$, the minimum distance between clusters does not change so much because clusters are still small and exist in their natural groups. But when $K = c^*$, this measure of separability increases abruptly, because there are well-separable clusters in the given data set and the distances between them are large. Similarly, if $K < c^*$, then clusters are merged by a clustering algorithm and are far away from each other.

In this proposed index, a separability measure called *D(K)* (see Eq. 15) is defined as the ratio of two inter-cluster distances. The first one, $D_{K\max}$, is the maximum distance between cluster centers, and the other one, $D_{K\min}$, is the minimum distance between them (see Eqs. 16 and 17). Of course, $D_{K\max}$ is limited by the maximum separation between two patterns in the given data set. Notice that when the number of clusters $K > c^*$, the distance $D_{K\max}$ is large, because clusters are still small and also the maximum distance does not change so much when these clusters are merged. Similarly, for $K = c^*$, the distance $D_{K\max}$ is also large because the data consist of well-separable clusters. Whereas, for $K < c^*$ this factor decreases, because clusters are merged into large ones and the maximum distance between the centers is smaller (see Fig. 2a). On the other hand, the $D_{K\min}$ is small when $K > c^*$, because clusters are subdivided into smaller ones and hence their centers are close to each other. But when $K = c^*$, the factor $D_{K\min}$ abruptly increases since distances between the centers will be proportionally larger. This applies also to $K < c^*$, and then the value of $D_{K\min}$ increases further until the number of clusters $K = 2$. Thus, for $D_{K\min}$ the knee point occurs when the number of clusters is equal to $c^* + 1$ (see Fig. 2b). It can be seen that the change of the $D(K)$ between $K = c^* + 1$ and $K = c^*$ partition schemes is the biggest (see Fig. 2d). Therefore, the difference between $D(K+1)$ and $D(K)$ can be used to determine significant changes of separability. Notice that if $K = 1$, it is assumed that $D(K)$ equals 0.

In order to understand the details of the STR index better, an example of calculating of the index will be presented. Let us denote this index as $\mathrm{STR} = A \cdot B$, where the component A denotes $E(K) - E(K-1)$, and B is $D(K+1) - D(K)$. As mentioned above, the example data were partitioned by the complete linkage clustering algorithm, and the number of clusters K was varied from 12 to 1. Since the proposed index is based on $E(K)$ and $D(K)$, so they must be computed for each K partition scheme of the data. In Table 1 are presented values of the STR index and of its components with respect to number of clusters. It should be noted that if the number of cluster equals 12 or 11, the index cannot be computed, because, e.g., if $K = 12$, the factor $D(K+1)$ is not calculated. Consequently, the index calculation starts when the number of clusters is equal to $K - 1 = 10$, and then $K = 11$ and $K + 1 = 12$. In this case, components of the index are as follows: $A = E(11) - E(10) = 9.75 - 9.38 = 0.37$ and $B = D(12) - D(11) = 20.50 - 14.90 = 5.6$. Finally, the STR index is equal to $0.37 \cdot 5.6 = 2.07$ (see Table 1). It should be observed that if the number of clusters equals 7, the component B is negative. This is so because $D_{K\min}$ can

Table 1 Values of the STR index and of its components with respect to the number of clusters K for the example data

K	$E(K)$	$D(K)$	A	B	STR index
12	9.98	20.50	–	–	–
11	9.75	14.90	–	–	–
10	9.38	14.09	0.37	5.60	2.07
9	8.95	13.31	0.43	0.81	0.35
8	8.33	14.65	0.62	0.78	0.48
7	7.90	12.32	0.43	−1.34	0
6	7.35	11.60	0.55	2.33	1.28
5	6.87	11.22	0.48	0.72	0.35
4	6.38	2.66	0.49	0.38	0.19
3	3.04	2.01	3.34	8.56	28.59
2	1.87	1	1.17	0.65	0.76
1	1	0	0.87	1.01	0.88

decrease when the number of clusters varies from $K_{\max}$ to $K_{\min}$, and then the factor $D(K)$ achieves large values (see Fig. 2b, d). In these cases, the value of the index is assumed to equal 0. Notice that the factor $E(K)$ is positive because sizes of clusters are always increased during a clustering process. In Fig. 3 a variation of the STR index with respect to the number of clusters for the example data is presented.

Thus, it seems reasonable that the definition of the new index includes the product of these two components calculated as the differences of the cluster compactness and separability between K and $K-1$, and also $K+1$ and K partition schemes (Eq. 18). Unlike most other indices, this new index uses the knee point detection when the number K varies within a certain range and is an input parameter of an underlying clustering algorithm. Although these two components of the index may have different scales, they need not be normalized because they are multiplied. Furthermore, the maximum value of the index occurs when a number of clusters $K = c^* - 1$, because the measure of cluster compactness changes abruptly between $K = c^*$ and $K = c^* - 1$ partition schemes. Consequently, the right number of clusters equals $c^* = K + 1$.

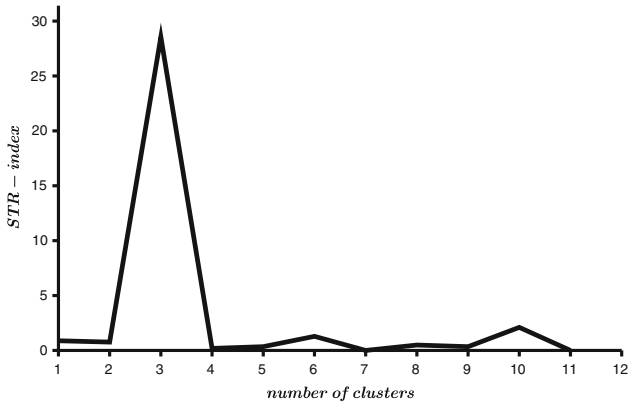


Fig. 3 Variation of the STR index with respect to the number of clusters for the example data

4 Experimental results

Several experiments were carried out to verify effectiveness of the new index. The first ones relate to determining the number of clusters for artificial and real-life data sets when the complete-linkage algorithm is applied as the underlying clustering method. The subsequent experiments are to show how effectively this new index works in comparison to the other popular validity indices such as Dann, DB, PBM and SIL indices. Here, three well-known algorithms were selected for clustering of data sets, namely, complete-linkage, K-means and EM methods.

4.1 Artificial data

Randomly generated six artificial data sets with a various number of clusters were used in the experiments. The first three of them called Data 1, Data 2 and Data 3 are 2-dimensional with 3, 4 and 15 clusters, respectively. The next three sets called Data 4, Data 5 and Data 6 are 3-dimensional with 4, 7 and 9 clusters, respectively. Table 2 presents a detailed description of these data taking also into account the number of elements per class.

As it can be observed in Fig. 4 clusters are mostly circular and located in various distances from each other; some of them are quite close. For example, in Fig. 4c clusters are small and most of them are very near each other. On the other hand, Fig. 4d–f presents various large clusters of 3-dimensional data sets. Here, clusters are more scattered, and the distances between them are also very different.

Table 2 Detailed description of the artificial data sets

Data sets	No. of elements	Features	Classes	No. of elements per class
Data 1	134	2	3	39, 48, 47
Data 2	400	2	4	50, 50, 150, 150
Data 3	429	2	15	31, 39, 38, 18, 29 30, 32, 27, 10, 39 22, 27, 39, 20, 28
Data 4	550	3	4	100, 100, 150, 200
Data 5	820	3	7	80, 90, 100, 100 100, 150, 200
Data 6	391	3	9	68, 62, 22, 22, 32 52, 36, 47, 50

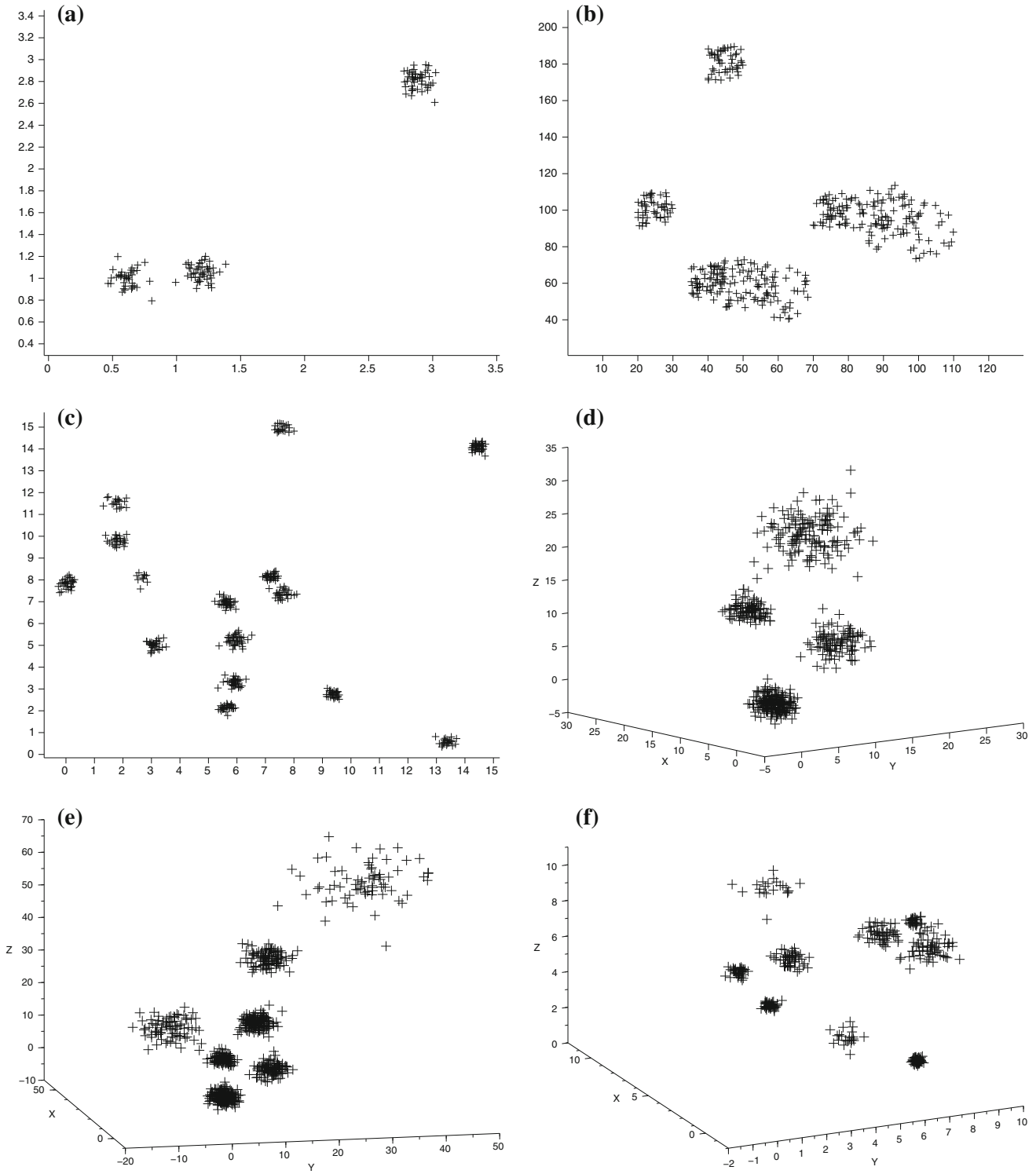


Fig. 4 Artificial data sets: **a** Data 1, **b** Data 2, **c** Data 3, **d** Data 4, **e** Data 5, **f** Data 6

4.1.1 Determination of the cluster number for the artificial data sets

Several tests were performed with the artificial data. The complete-linkage method as the underlying clustering algorithm was used for partitioning of these data, and the cluster number K varied from $K_{\max} = \sqrt{n}$ to $K_{\min} = 1$. Note that the maximum number of clusters should not be greater than $\sqrt{n}$, where n is the number of elements in a given data set. This value is an accepted rule in the

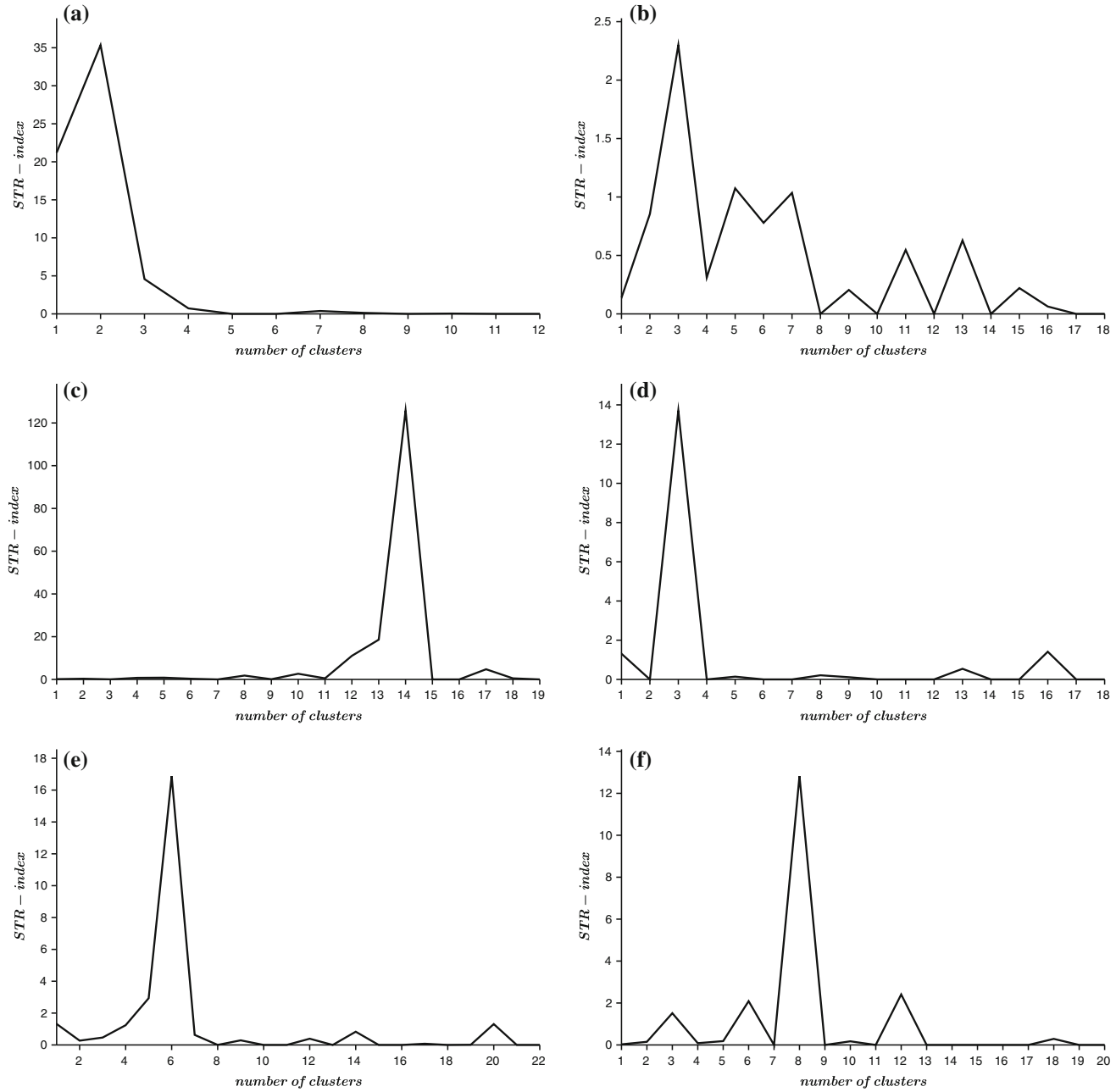


Fig. 5 Variation of the STR index with respect to the number of clusters for: **a** Data 1, **b** Data 2, **c** Data 3, **d** Data 4, **e** Data 5, **f** Data 6

clustering literature [19]. To demonstrate the behavior of this index, in Fig. 5 the variation of the STR index with respect to the number of clusters is presented.

As it can be noticed, in all these cases the maximum values of this index indicate $K = c^* - 1$ partition scheme of data. To calculate the correct number of clusters we need to increase K by 1—this issue is explained in detail in Sect. 3.1. It can be seen that for most of the well-separable data, the index peaks are high and explicit. However, Fig. 5b shows several high distinct peaks for the set Data 2, which consists of large ellipsoidal clusters. Notice that for $K > c^*$ these clusters are subdivided into smaller ones. But when they are merged into larger ones by the clustering algorithm, their compactness and separability change significantly. Therefore, this index provides several large values. On the other hand, in Fig. 5d–f are presented values of the index for the 3-dimensional data. Despite the fact that the patterns in some clusters are much more scattered, the STR index generates clear peaks which are related to the correct partitioning of the data.

Moreover, the number of the peaks and their height can provide interesting information about data structure, e.g., it

can indicate how much these clusters are separated; however, this subject requires further study.

4.1.2 Determination of suitable clustering for the artificial data sets

As it was mentioned above, in all those experiments the STR index proved reliable and made it possible to specify the correct number of clusters in the artificial data sets. However, a very important issue is also the appropriate partitioning of data, which means that all patterns belong to suitable clusters. This can be demonstrated graphically but only for 2-dimensional and 3-dimensional data sets. As in Sect. 4.1.1, the complete-linkage method in conjunction with the STR index was used for partitioning of the above-mentioned artificial data, and Fig. 6 presents clustered data, where each cluster is denoted by a successive number. It can be seen that despite various size and number of clusters, all the patterns are assigned to correct groups. Notice that the right number of clusters given as an input parameter of a clustering algorithm does not guarantee that all patterns are associated with appropriate clusters. It mainly depends on properties and additional input parameters of clustering algorithms. Certainly, an incorrect number of K results in poor partitioning of data by these algorithms.

4.2 Real-life data

The complete-linkage method was also used for the partitioning of the real-life data sets, where the cluster number K also varied from $\sqrt{n}$ to 1. The appropriate number of clusters in the data was found for the following sets: Breast cancer, Breast tissue, Glass, Haberman, Iris, Parkinsons, Vertebral column and Wine, which were drawn from the UCI repository [2]. The description of these data is presented in Table 3.

The first set called Breast cancer is the Wisconsin Breast Cancer data. It consists of 683 patterns belonging to two classes: Benign (444 instances) and Malignant (239 instances). Each pattern is characterized by nine features. The next set called Breast tissue includes measurements of electrical impedance of tissue samples excised from breasts. This set includes 106 elements, which are located in 6 classes, and each sample is described by 9 features. Next, the Glass data set contains information about 6 types of glass, which are defined in terms of their oxide content. In more detail, the set has 214 instances and each of them is described by 9 attributes. The Haberman data set consists of the cases from a study on the survival of patients who had undergone surgery for breast cancer. The set has 306 cases belonging to two classes, and the number of features equals 3. The Iris data are very well known and extensively used in many comparisons of classifiers. This set has three classes Setosa, Virginica and Versicolor, which contain 50 instances per class. Moreover, each pattern is represented by four features, and two classes Virginia and Versicolor are overlapping each other. On the other hand, the third class Setosa is well separated from the others. The next set is the Parkinsons data set and it consists of 195 cases, which are described by 22 attributes. These data are composed of biomedical voice measurements from people and are used to discriminate healthy people from those with Parkinson's disease (2 classes). The following set, Vertebral column, contains values of six biomechanical features, which are used to classify orthopedic patients into 3 classes. In this set, the total number of cases equals 310. Finally, the Wine data set shows the results of a chemical analysis of wines. It comprises three classes of wines, which consist of 59, 71 and 48 samples per class, respectively. Altogether, the data set contains 178 patterns represented by 13 features.

Figure 7 shows values of the STR index with respect to the number of clusters when the complete-linkage algorithm was used for partitioning of the data. As it can be observed, despite the multidimensional data and a various number of the 'natural' clusters in these data sets, the STR index provides the right number K in most cases. For example, Fig. 7a presents the maximum of the STR index for $K = 1$, and so, the appropriate number of clusters equals 2 for the Breast cancer data set. For the other data, when the actual number of clusters equals 2 or 3, this new index also provides correct indications (see Fig. 7d–h). On the other hand, there are two cases where this number of clusters is incorrect, and it concerns two sets, i.e., Breast tissue and Glass. It should be noted that these data possess several clusters, but with a small number of elements. Thus, in this case the appropriate partitioning data are very hard, and most validity indices give wrong results. In addition, the properties of clustering algorithms greatly affect the shape and the size of created clusters. For example, the complete-linkage method favors creation of compact clusters and imposes spherical-shape clusters on data. However, it should be emphasized that the proposed index accurately indicates the appropriate number of clusters for the other data. Thus, these experiments confirm the effectiveness of this approach in the partitioning of these data sets.

4.3 Comparison of several validity indices

In order to demonstrate the effectiveness of the proposed index, several experiments have been performed for the above-mentioned data sets. For comparison four indices have been used, i.e., Dunn, DB, PBM and SIL indices. Detailed information regarding these indices is presented in

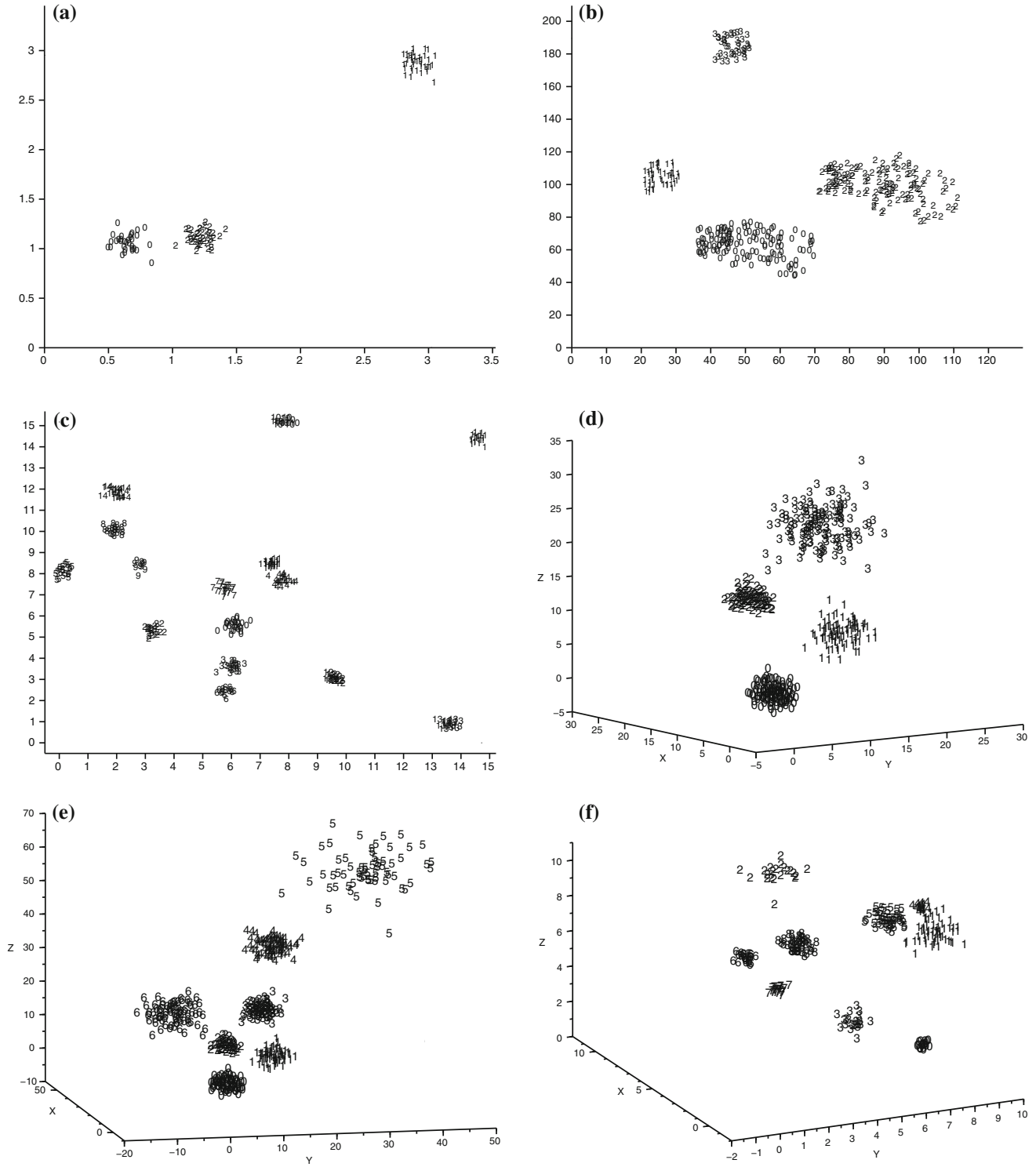


Fig. 6 Clustered data by the complete-linkage algorithm, corresponding to the maximal value of *STR* index and indicated by numbers, for: **a** Data 1, **b** Data 2, **c** Data 3, **d** Data 4, **e** Data 5, **f** Data 6

Table 3 Description of real-life data sets

Data set	No. of elements	Features	Classes
Breast cancer	683	9	2
Breast tissue	106	9	6
Glass	214	9	6
Haberman	306	3	2
Iris	150	4	3
Parkinsons	195	22	2
Vertebral column	310	6	3
Wine	178	13	3

Sect. 2. Three methods were used as underlying clustering techniques, namely, complete-linkage, EM and *K*-means. Each of these has different properties and approach to data set partitioning. Moreover, for each of these algorithms, the value of the input parameter K varied from $K_{\max} = \sqrt{n}$ to $K_{\min} = 1$. These experiments concern the determining of the proper number of clusters present in the given data sets, and, as mentioned, it is the key parameter for clustering algorithms.

Additionally, the accuracy rate is defined to determine the accuracy of a validity index in detecting the number of clusters. Here, the rate equals $A/$total number of data sets. A is the sum of the ratios of the difference $|p - o|$ to p, where the factor p denotes the actual number of clusters present in a given data set, and the other factor o is the number of clusters provided by the validity index. Of course, if the *rate* used for the index is close to 0, this means that this index is perfect.

Table 4 presents the comparison of the five indices while taking into account the number of clusters. As mentioned above, the complete-linkage algorithm creates compact clusters of approximately equal diameters and it is sensitive to outliers. It is so due to the fact that the similarity measure of clusters is the maximum distance between two patterns. As it can be seen from the table, the STR index provides the right cluster number for all data sets, apart from Glass and Breast Tissue. These two data sets possess 6 clusters while the maximum value of this new index indicates 5 clusters. But these results are good when compared to the other indices. For Glass, the four indices, i.e., Dunn, DB, PBM and SIL fail to detect the appropriate number of clusters and show 4, 10 ,2 and 8 clusters, respectively. Similarly, these indices provide an incorrect number of clusters for the Breast tissue data. Thus, the results confirm very good effectiveness of the STR index.

Table 5 provides the comparison of these five indices for the *EM* method, which looks for Gaussian-shape clusters. Notice that the choice of initial parameters for this clustering method is of great importance for obtaining correct results. Here, the STR index was able to provide the right number of clusters for the eight data sets, i.e., Data 1, Data 2 and from Data 4 to Data 6 and three real-life data sets. In turn, the Dunn index indicated the appropriate number of clusters only for three data sets, the *DB* index for eight sets, the *PBM* and the *SIL* for six sets. It can be seen that when compared to the other indices, the results obtained by the STR index are very good.

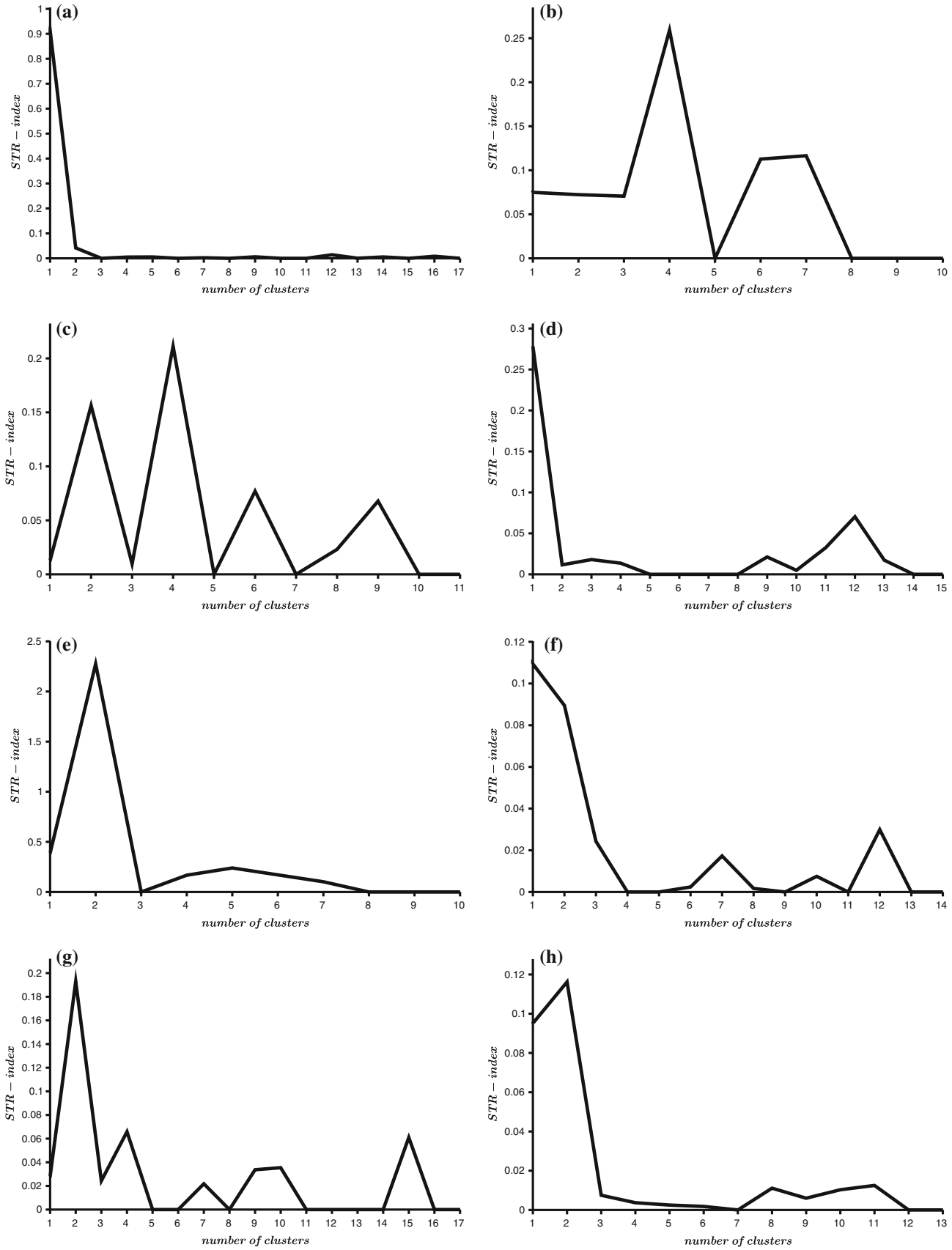
In Table 6 the results for *K*-means method are shown. It is generally known that this algorithm often gets stuck at suboptimal configurations. In order to overcome this problem, several re-initializations are used for different initial cluster centers. This algorithm looks for compact clusters around a mean. From this table it can be seen that the STR index indicates the proper number of clusters in almost all the cases except for Data 3, Iris and Vertebral column data sets. In comparison to the others, it is the best result.

To summarize, regardless which one of the three underlying clustering algorithms was used, the STR index provides very impressive results. The proposed index consistently outperforms the other indices in terms of the correct indication of the cluster number. The values of the accuracy rate also prove the superiority of this new index.

5 Conclusion

There is a large number of cluster validity indices in the clustering literature. Generally, these indices can be used to assess crisp and/or fuzzy clustering of data. The above-mentioned validity indices, i.e., the Dunn index, the DB index, the PBM index or the SIL index are popular and widely used by different clustering algorithms. However, there is no validity index which works well with all the clustering algorithms for a wide range of data sets. Hence, there is a constant need to develop efficient indices which can be used with different algorithms for various data sets.

In this paper, a new cluster validity index was proposed and the detailed analysis of its work was also done. Similar to the other reported studies of indices, this index was mainly used to identify the right number of clusters, and was also the measure of the correctness of various partitioning of data. The proposed index is defined as the product of two components, and its maximum value indicates the appropriate partition scheme. The first component measures changes of cluster compactness, and the second one measures changes of cluster separability. Here, unlike most of the other indices, this approach makes it possible to


(a)
STR − index
number of clusters
(b)
STR − index
number of clusters
(c)
STR − index
number of clusters
(d)
STR − index
number of clusters
(e)
STR − index
number of clusters
(f)
STR − index
number of clusters
(g)
STR − index
number of clusters
(h)
STR − index
number of clusters

◀**Fig. 7** Variation of the STR index with respect to the number of clusters for: **a** Breast cancer, **b** Breast tissue, **c** Glass, **d** Haberman,**e** Iris, **f** Parkinsons, **g** Vertebral column, **h** Wine

Table 4 Comparison of the number of clusters obtained by means of the complete-linkage algorithm in conjunction with the Dunn index, the DB index, the PBM index, the SIL index and the STR index

Data set	*N*	Number of clusters obtained				
		Dunn	DB	PBM	SIL	STR
Data 1	3	2	2	3	2	**3**
Data 2	4	2	4	4	4	**4**
Data 3	15	13	14	15	14	**15**
Data 4	4	4	4	4	4	**4**
Data 5	7	2	7	7	7	**7**
Data 6	9	7	9	9	7	**9**
Cancer	2	15	2	2	2	**2**
Tissue	6	2	2	3	2	**5**
Glass	6	4	10	2	8	**5**
Haberman	2	6	12	3	12	**2**
Iris	3	12	3	3	3	**3**
Parkinsons	2	2	2	2	2	**2**
Vertebral column	3	2	2	2	2	**3**
Wine	3	12	12	3	2	**3**
Accuracy rate		1.26	0.72	0.14	0.52	0.02

The values of the STR index are in bold

N denotes the actual number of clusters in the data sets. The Accuracy rate determines the accuracy of the validity index in detecting the proper number of clusters (Sect. 4.3)

detect the knee point occurring in the measuring of cluster compactness and separability. Thus, the maximum value of this index is very clear. Moreover, although the two components have different scales, they do not need to be normalized. It can be seen that measures of cluster properties are also appropriately chosen so that the value of this index is very large when a number of clusters equals the actual number of clusters present in a data set. To investigate the behavior of the proposed validity index, as the underlying clustering algorithms, three well-known methods characterizing different approaches to partitioning of data sets were selected. They are, the complete-linkage, the *K*-means and the EM algorithms.

The performed tests have proven the advantages of the proposed index compared to the above-mentioned indices, i.e., Dunn, DB, PBM and SIL indices. In these experiments, several artificial and real-life data sets were used, where artificial data were two or three dimensional, and the number of clusters varied from three to fifteen. The dimensionality of the real-life data was from three to twenty two. All the presented results confirm high

Table 5 Comparison of the number of clusters obtained by means of the EM algorithm in conjunction with the Dunn index, the DB index, the PBM index, the SIL index and the STR index

Data set	*N*	Number of clusters obtained				
		Dunn	DB	PBM	SIL	STR
Data 1	3	2	2	3	2	**3**
Data 2	4	4	4	7	4	**4**
Data 3	15	12	14	17	13	**14**
Data 4	4	4	4	4	4	**4**
Data 5	7	2	7	6	7	**7**
Data 6	9	7	9	9	7	**9**
Cancer	2	2	2	2	2	**2**
Tissue	6	3	2	5	2	**5**
Glass	6	7	6	2	2	**5**
Haberman	2	4	11	3	3	**3**
Iris	3	2	2	3	2	**2**
Parkinsons	2	7	2	3	2	**2**
Vertebral column	3	14	2	4	2	**2**
Wine	3	6	3	3	3	**3**
Accuracy rate		0.62	0.44	0.26	0.23	0.11

The values of the STR index are in bold

N denotes the actual number of clusters in the data sets. The accuracy rate determines the accuracy of the validity index in detecting the proper number of clusters (Sect. 4.3)

Table 6 Comparison of the number of clusters obtained by means of the *K*-means algorithm in conjunction with the Dunn index, the DB index, the PBM index, the SIL index and the STR index

Data set	*N*	Number of clusters obtained				
		Dunn	DB	PBM	SIL	STR
Data 1	3	2	2	3	2	**3**
Data 2	4	4	4	4	4	**4**
Data 3	15	11	14	13	13	**14**
Data 4	4	4	4	4	4	**4**
Data 5	7	2	7	6	7	**7**
Data 6	9	7	9	9	9	**9**
Cancer	2	2	2	2	2	**2**
Tissue	6	4	2	4	2	**6**
Glass	6	4	6	6	6	**6**
Haberman	2	2	6	2	2	**2**
Iris	3	2	2	3	2	**2**
Parkinsons	2	4	2	3	2	**2**
Vertebral column	3	2	2	3	2	**2**
Wine	3	4	3	3	3	**3**
Accuracy rate		0.27	0.31	0.079	0.18	0.075

The values of the STR index are in bold

N denotes the actual number of clusters in the data sets. The accuracy rate determines the accuracy of the validity index in detecting the proper number of clusters (Sect. 4.3)

efficiency of the STR index where this index in most cases outperforms the other indices in the conducted experiments. Further work will include application of the new index for the fuzzy clustering of various data sets.

References

1. Arbelaitz O, Gurrutxaga I, Muguerza J, Prez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recogn 46(1):243–256
2. Bache K, Lichman M (2013) UCI machine learning repository. University of California, School of Information and Computer Science, Irvine. http://archive.ics.uci.edu/ml
3. Baskir MB, Türksen IB (2013) Enhanced fuzzy clustering algorithm and cluster validity index for human perception. Expert Syst Appl 40:929–937
4. Bezdek JC, Pal NR (1998) Some new indexes of cluster validity. IEEE Trans Syst Man Cybern 28:301–315
5. Bradley P, Fayyad U (1998) Refining initial points for k-means clustering. In: Proceedings of the fifteenth international conference on knowledge discovery and data mining. AAAI Press, New York, pp 9–15
6. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1(4):224–227
7. Dunn JC (1974) Well separated clusters and optimal fuzzy partitions. J Cybern 4:95–104
8. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York
9. Rezaei Fränti PM, Zhao Q (2014) Centroid index: cluster level similarity measure. Pattern Recogn 47(9):3034–3045
10. Fred LN, Leitao MN (2003) A new cluster isolation criterion based on dissimilarity increments. IEEE Trans Pattern Anal Mach Intell 25(8):944–958
11. Halkidi M, Batistakis Y, Vazirgiannis M (2002) Clustering validity checking methods: Part II. ACM SIGMOD Record 31(3):19–27
12. Jain A, Dubes R (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs
13. Kim M, Ramakrishna RS (2005) New indices for cluster validity assessment. Pattern Recogn Lett 26:2353–2363
14. Lago-Fernández LF, Corbacho F (2010) Normality-based validation for crisp clustering. Pattern Recogn 43(3):782–795
15. Meng X, van Dyk D (1997) The EM algorithm: an old folk-song sung to a fast new tune. J Roy Stat Soc Ser B (Methodol) 59(3):511–567
16. Murtagh F (1983) A survey of recent advances in hierarchical clustering algorithms. Comput J 26(4):354–359
17. Ozkan I, Türksen IB (2012) MiniMax ε-stable cluster validity index for Type-2 fuzziness. Inform Sci 184:64–74
18. Pakhira MK, Bandyopadhyay S, Maulik U (2004) Validity index for crisp and fuzzy clusters. Pattern Recogn 37(3):487–501
19. Pal NR, Bezdek JC (1995) On cluster validity for the fuzzy c-means model. IEEE Trans Fuzzy Syst 3(3):370–379
20. Pascual D, Pla F, Sánchez JS (2010) Cluster validation using information stability measures. Pattern Recogn Lett 31(6):454–461
21. Pelleg D, Moore A (2000) X-means: extending k-means with efficient estimation of the number of clusters. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining. ACM Press, San Diego, pp 277–281
22. Rohlf F (1982) Single link clustering algorithms. In: Krishnaiah P, Kanal L (eds) Handbook of statistics, vol 2. North-Holland, Amsterdam, pp 267–284
23. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
24. Saha S, Bandyopadhyay S (2012) Some connectivity based cluster validity indices. Appl Soft Comput 12(5):1555–1565
25. Sameh AS, Asoke KN (2009) Development of assessment criteria for clustering algorithms. Pattern Anal Appl 12:79–98
26. Shieh H-L (2014) Robust validity index for a modified subtractive clustering algorithm. Appl Soft Comput 22:47–59
27. Shihong Y, Jianpei W, Jeenshing W, Xiujuan B (2015) A new validity index for evaluating the clustering results by partitional clustering algorithms. Soft Comput. doi:10.1007/s00500-014-1577-1
28. Wu KL, Yang MS, Hsieh JN (2009) Robust cluster validity indexes. Pattern Recogn 42:2541–2550
29. Zait M, Messatfa H (1997) A comparative study of clustering methods. Future Gener Comput Syst 13(2–3):149–159
30. Zalik KR (2010) Cluster validity index for estimation of fuzzy clusters of different sizes and densities. Pattern Recogn 43:3374–3390
31. Zhang D, Ji M, Yang J, Zhang Y, Xie F (2014) A novel cluster validity index for fuzzy clustering based on bipartite modularity. Fuzzy Sets Syst 253:122–137
32. Zhao Q, Fränti P (2014) WB-index: a sum-of-squares based index for cluster validity. Data Knowl Eng 92:77–89