Improvement of the Simplified Silhouette Validity Index

Artur Starczewski[1(✉)] and Krzysztof Przybyszewski[2,3]

[1] Institute of Computational Intelligence, Częstochowa University of Technology, Al. Armii Krajowej 36, 42-200 Częstochowa, Poland
artur.starczewski@iisi.pcz.pl
[2] Information Technology Institute, University of Social Sciences, 90-113 Łódź, Poland
[3] Clark University, Worcester, MA 01610, USA

Abstract. The fundamental issue of data clustering is an evaluation of results of clustering algorithms. Lots of methods have been proposed for cluster validation. The most popular approach is based on internal cluster validity indices. Among this kind of indices, the *Silhouette* index and its computationally simpled version, i.e. the *Simplified Silhouette*, are frequently used. In this paper modification of the *Simplified Silhouette* index is proposed. The suggested approach is based on using an additional component, which improves clusters validity assessment. The performance of the new cluster validity indices has been demonstrated for artificial and real datasets, where the *PAM* clustering algorithm has been applied as the underlying clustering technique.

Keywords: Clustering · Cluster validity index
PAM clustering technique

1 Introduction

Data clustering aims to discover natural existing structures in a dataset. For this purpose, data are partitioned into groups (clusters) of objects. Objects within a cluster are similar, whereas they are dissimilar in different clusters. Since there is a large variety of datasets different clustering algorithms and their configurations are still created, e.g. [9,11,12,31]. Note that among clustering methods two major categories are distinguished: partitioning and hierarchical clustering. For example, the well-known partitioning algorithms are, e.g. *K-means*, *Partitioning Around Medoids* (*PAM*) [5,24] and *Expectation Maximization* (*EM*) [21]. Whereas the agglomerative hierarchical clustering includes such methods as, e.g. the *Single-linkage*, *Complete-linkage* or *Average-linkage* [16,22,25]. Data clustering is applied in many areas, such as biology, spatial data analysis, business and so on. It can be noted that there is no a clustering algorithm, which creates the right data partition for all datasets. Moreover, the

L. Rutkowski et al. (Eds.): ICAISC 2018, LNAI 10842, pp. 433–444, 2018.
https://doi.org/10.1007/978-3-319-91262-2_39

same algorithm can also give different results depending on the input parameters. Therefore, cluster validation should be used to assess the results of data clustering. Generally, it is a very difficult task and is the most frequently realized by validity indices. Techniques of the cluster validation are usually classified into three groups, i.e. external, internal and relative validation [16,30]. The external validation is based on a comparison of partitions of a dataset obtained by a clustering algorithm with the correct partition of this data. In turn, the internal approach uses only the intrinsic properties of the dataset. On the other hand, the relative validation method compares the data partitions obtained by chaining input parameters of a clustering algorithm. It should be noted that the number of clusters is the key parameter for many clustering algorithms. So far, a number of authors have proposed different validity indices or modifications of existing indices, e.g., [1,10,18,29,32,33,36,38]. Among internal cluster validity indices, the *Silhouette (SIL)* [26] and *Simplified Silhouette (SimSIL)* [15] indices are frequently used to evaluate the efficacy of the clustering algorithms in detecting the right data partitioning. It is important to note that clustering methods in conjunction with cluster validity indices can be used during a process of designing various neural networks [2–4,6,17], neuro-fuzzy structures [7,8,20,27,28] and creating some algorithms for identification of classes [13,14].

In this paper, new cluster validity indices called the $SimSILA$ and the $SimSILAv1$, are presented. These new indices modify the *Simplified Silhouette (SimSIL)* index. The proposed approach is based on an additional component and it is a detailed explanation in Sect. 3. In order to present effectiveness of the validity indices, several experiments were performed for various datasets. This paper is organized as follows: Sect. 2 presents a detailed description of the *Silhouette*, $SILA$ and $SILAv1$ indices. In Sect. 3 the *Simplified Silhouette*, $SimSILA$ and $SimSILAv1$ indices are outlined. Section 4 illustrates experimental results on datasets. Finally, Sect. 5 presents conclusions.

2 Modification of the Silhouette Index

In this section modification of the *Silhouette (SIL)* index is described. This approach was proposed and discussed in papers [34,35]. Let us denote K-partition scheme of a dataset X by $C = \{C_1, C, ..., C_K\}$, where C_k indicates k_{th} cluster, $k = 1, .., K$. The original SIL index is presented as follows:

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

where $SIL(C_k)$ is the *Silhouette width* for the given cluster C_k and is defined as:

$$SIL(C_k) = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \frac{b(\mathbf{x}) - a(\mathbf{x})}{max\left(a(\mathbf{x}), b(\mathbf{x})\right)} \tag{2}$$

n_k is a number of elements in C_k, and $a(\mathbf{x})$ is the within-cluster mean distance, i.e. it is 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 elements belonging to the other clusters. The values of the index are from the range -1 to 1 and a maximum value (close to 1) provides the best partitioning of the dataset.

Now let us turn to the modification of this index [34]. This approach is based on using an additional component, which improves a performance of the index. The new index is called *SILA* index and it is defined as follows:

$$SILA = \frac{1}{n}\left(\sum_{\mathbf{x}\in X}\left(\frac{b(\mathbf{x}) - a(\mathbf{x})}{\max\left(a(\mathbf{x}), b(\mathbf{x})\right)} \cdot \frac{1}{(1 + a(x))^q}\right)\right) \tag{3}$$

where the exponent q is equal 1 and n is the number of elements in a dataset. A maximum value of the new index indicates the right partition scheme. Noted that the choice of the value of the q is very important and $q = 1$ can be too small for the very large differences of distances between data points. Hence, the new concept was proposed in paper [35]. This new index, called $SILAv1$, can be presented by Eq. (3), where q is defined as below:

$$q = 2 + \frac{K^2}{n} \tag{4}$$

Generally, the $SILA$ and $SILAv1$ indices ensure a better performance compared to the original *Silhouette* index. In the next section, a detailed explanation of modification of the *Simplified Silhouette* index is presented.

3 Modification of the Simplified Silhouette Index

It can be noted that the *Silhouette* index depends on of the computation of all the distances between data elements and it can lead to a computational cost $O(mn^2)$ [37], where m is the number of features. On the other hand, the *Simplified Silhouette* index is much less computationally expensive, and the overall complexity of the computation of the index is estimated as $O(kmn)$ [37]. Although the *Simplified Silhouette* index is similar to the *Silhouette* index, there are very significant differences. First, the distance of $\mathbf{x}$ to the cluster is not the average distance between $\mathbf{x}$ and the rest of the elements belonging to the same cluster. It is calculated as the distance between $\mathbf{x}$ and the centroid of the cluster and can be written as follows:

$$\hat{a}(\mathbf{x}) = d\left(\mathbf{x}, \bar{C}_k\right) \tag{5}$$

where $\bar{C}_k$ is the centroid of the cluster C_k and $d\left(\mathbf{x}, \bar{C}_k\right)$ is a function of the distance between $\mathbf{x}$ and $\bar{C}_k$. Next, the distance of $\mathbf{x}$ to the other cluster is defined as follows:

$$\hat{b}(\mathbf{x}) = \min_{\substack{l=1 \\ l\neq k}}^{K} d(\mathbf{x}, \bar{C}_l) \tag{6}$$

where $\bar{C}_l$ is the centroid of the cluster C_l and $l \neq k$. Finally, the *Simplified Silhouette* ($SimSIL$) index is defined as:

$$SimSIL = \frac{1}{\mathrm{n}} \sum_{\mathbf{x} \in X} \frac{\hat{b}(\mathbf{x}) - \hat{a}(\mathbf{x})}{\max(\hat{a}(\mathbf{x}), \hat{b}(\mathbf{x}))} \tag{7}$$

where n is the number of elements in the dataset X. The value of the index is also from the range -1 to 1 and a maximum value indicates the right partition scheme.

As in the previous index, the modification of the $Simplified\ Silhouette$ index is based on using the additional component, which is expressed as:

$$\hat{A}(\mathbf{x}) = \frac{1}{(1 + \hat{a}(\mathbf{x}))^q} \tag{8}$$

For the exponent $q = 1$, the newly proposed index is called $SimSILA$ and can be written as:

$$SimSILA = \frac{1}{n} \left(\sum_{\mathbf{x} \in X} \left(\frac{\hat{b}(\mathbf{x}) - \hat{a}(\mathbf{x})}{\max\left(\hat{a}(\mathbf{x}), \hat{b}(\mathbf{x})\right)} \cdot \frac{1}{(1 + \hat{a}(x))^q} \right) \right) \tag{9}$$

It can be noted that the additional component $\hat{A}(\mathbf{x})$ corrects the value of the index. When a clustering algorithm greatly increases sizes of clusters, the ratio of $1/(1+\hat{a}(x))^q$ decreases significantly and the value of the index is also decreased. However, the value $q = 1$ can be too small to appropriately correct the $SimSILA$ index. Hence, the issue of the choice of the exponent q for $\hat{A}(\mathbf{x})$ is a very significant problem. As with the previous index, the new index called $SimSILAv1$ is proposed and contains a formula of the change of the exponent q depending on the number of clusters. This formula is expressed by (4). Thus, the $SimSILAv1$ index can be presented by Eq. (9), where q is calculated by (4). It should be noted that the new indices can take values between $1/(1+\hat{a}(x))^q$ and $-1/(1+\hat{a}(x))^q$. A maximum value of the index selects the right data partitioning for a dataset. In the next section, the results of the experimental studies are presented to confirm the effectiveness of these new indices.

4 Experimental Results

In this section, several experiments have been conducted on artificial and real datasets using the *Partitioning Around Medoids* (PAM) clustering algorithm. This algorithm is a realisation of *K-medoid* clustering, which is a more robust version of *K-means* method. Both *K-medoids* and the *K-means* algorithm are partitional, but the first method searches K representative data elements (medoids) among all elements of a dataset. After finding K medoids, K clusters are created by assigning each data point to the nearest medoid. In contrast to the *K-means*, the *K-medoids* algorithm chooses data elements as centers (medoids). Moreover,

the Manhattan Norm is used to define distances between elements of the dataset. These make that the PAM algorithm is robust to noise and outliers. As mentioned in Sect. 1, the different parameter configurations of clustering algorithms can lead to different results. Thus, the choice of these input parameters is a key issue. Furthermore, one of the essential configuration parameters is a number of clusters. This parameter should be set before the start of the algorithm, but it is not usually known in advance. The common way to resolve this problem is to run the clustering algorithm multiple times with a different number of clusters and select the best result. For the clustering analyze, the range of the different number of clusters should be varied from $K_{min} = 2$ to $K_{max} = \sqrt{n}$, [23]. Whereas, the evaluation of results is usually realized by cluster validity indices. In experiments conducted on artificial and real datasets, the six indices, i.e. the *Silhouette (SIL)*, $SILA$, $SILAv1$, *Simplified Silhouette* ($SimSil$), $SimSILA$ and $SimSILAv1$ are used to determine the right number of clusters. To show the efficacy of the new validity indices, the results are also presented on the plots. It is assumed that the value of the validity indices equals 0 for $K = 1$. Furthermore, the *min-max* normalization of data has been applied to all the datasets used in the experiments. In order to better compare of the new indices, the maximum value of all the indices is modified and it is equal to 1.

4.1 Datasets

In the conducted experiments four artificial and six real datasets are used. The artificial data was called *Data* 1, *Data* 2, *Data* 3 and *Data* 4 and they were 2-dimensional with 3, 5, 8 and 11 clusters, respectively. Note that they consist of various cluster structure and densities. The scatter plot of these data is presented in Fig. 1. As it can be observed on the plot the distances between clusters are very different and some clusters are quite close. Generally, clusters are located in groups and some of the clusters are very close and others quite far. Moreover, the sizes of the clusters are different and they contain the various number of elements. Hence, many clusters validity indices can provide incorrect partitioning schemes. The real datasets are numeric data from the UCE Irvine Machine Learning Repository [19]: $Diabetes$, $Ecoli$, $Glass$, $Iris$, $Spectf$, $Wine$. The *Diabetes* dataset includes results of studies relating to the signs of diabetes in patients. This set includes 768 instances belonging to 2 classes and each item is described by 8 features. The second set is *Ecoli* dataset consisting of 336 instances, and the number of attributes equals 7. It has 8 classes, which represent the protein localization sites. Next comes the *Glass* dataset, which contains information about 6 types of glass defined in terms of their oxide content. The set has 214 instances and each of them is described by 9 attributes. The well-known *Iris* data are extensively used in many comparisons of classifiers. This set has three classes, which contain 50 instances per class. Moreover, each item is represented by four features. The *Spectf* dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography images. This set includes 267 instances and each of them is described by 44 features. It has 2 classes. Finally, the $Wine$ dataset shows the results of a chemical analysis of wines. It comprises

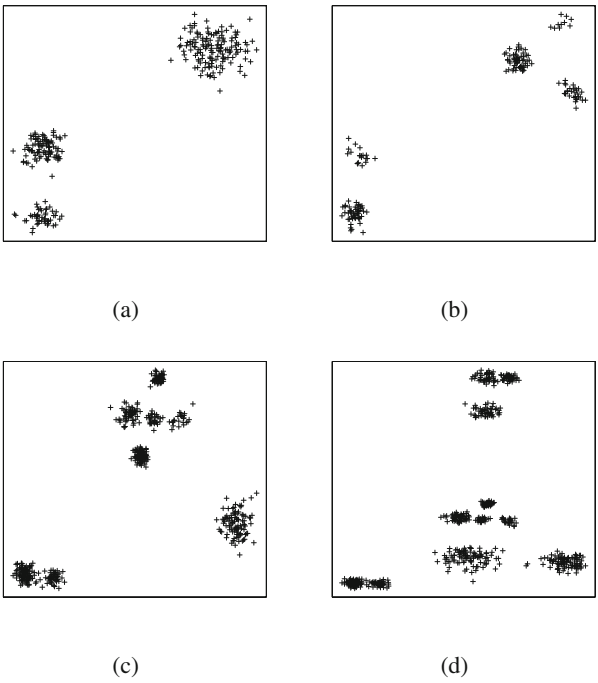

Fig. 1. 2-dimensional artificial datasets: (a) *Data* 1, (b) *Data* 2, (c) *Data* 3, and (d) *Data* 4

Table 1. A detailed description of the artificial datasets

Datasets	No. of elements	Features	Classes
Data 1	300	2	3
Data 2	170	2	5
Data 3	495	2	8
Data 4	665	2	11
Diabetes	768	8	2
Ecoli	336	7	8
Glass	214	9	6
Iris	150	4	3
Spectf	267	44	2
Wine	178	13	3

three classes of wines. Altogether, the dataset contains 178 patterns, where each of them is described by 13 features.

Additionally, Table 1 shows a detailed description of these datasets used in experiments.

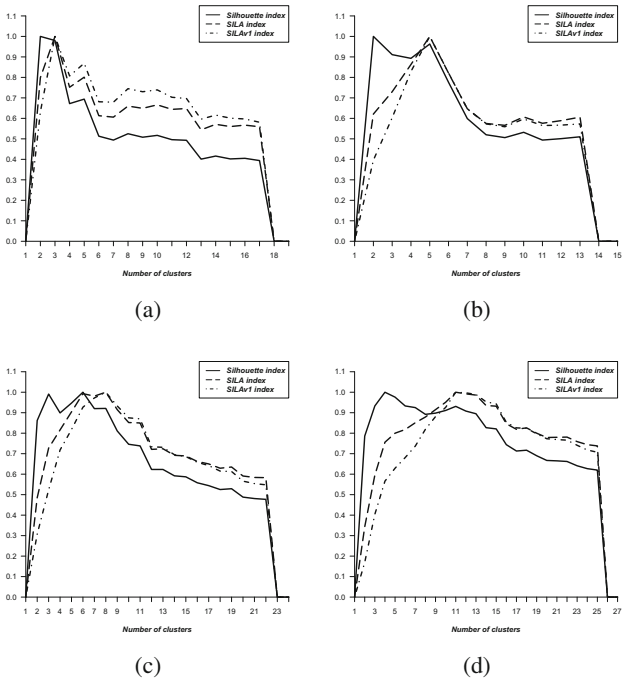

Fig. 2. Variations of the *Silhouette*, *SILA* and $SILAv1$ indices with respect to the number of clusters for 2-dimensional datasets: (a) *Data* 1, (b) *Data* 2, (c) *Data* 3, and (d) *Data* 4 partitioned by the *PAM* method.

4.2 Experiments

The experimental analysis is designed to evaluate the performance of the new indices. In these studies, the partitional *PAM* method as the underlying clustering method was adopted to clustering of the datasets. First of all, the *Silhouette*, *SILA* and $SILAv1$ indices are analyzed. For this purpose, the 2-dimensional *Data* 1, *Data* 2, *Data* 3 and *Data* 4 datasets have been clustered by the *PAM* algorithm. As shown in Fig. 1, these datasets create groups of clusters, which are far away from each other and their sizes are very different. As mentioned above, the number of clusters is the key configuration parameter of clustering methods and it is usually varied from $K_{min} = 2$ to $K_{max} = \sqrt{n}$. It is assumed that the value of the validity indices is equal 0 for $K = 1$. In Fig. 2 the comparison of the variations of the *Silhouette*, *SILA* and $SILAv1$ indices with respect to the number of clusters is presented for the artificial datasets. It is also noticeable that the *SILA* and $SILAv1$ indices provide the correct number of clusters for all the artificial datasets. In addition, the value of the *SILAv1* index more decreases than the value of *SILA* for the small number of clusters, i.e. when the number $K < c^*$ (where c^* is the right number of clusters). This means that the additional component $A(\mathbf{x})$ used in the $SILAv1$ index more reduces the value of the

index than the value of the *SILA* index. On the other hand, when the number of clusters $K > c^*$ the component $A(\mathbf{x})$ can increase the values of these indices slightly (see Fig. 2). On the contrary, the *Silhouette* index incorrectly selects all partitioning schemes and mainly provides the greatest values when the number of clusters $K = 2$. Next, the $Simplified\ Silhouette$, *SimSILA*, and $SimSILAv1$ indices are analyzed. As with the previous studies, four artificial datasets, i.e. *Data* 1, *Data* 2, *Data* 3 and *Data* 4 have been clustered by the *PAM* algorithm. The comparison of the variations of the $Simplified\ Silhouette$, *SimSILA* and $SimSILAv1$ indices with respect to the number of clusters is presented in Fig. 3. Despite the fact that the differences of distances between clusters are large, the *SimSILA* and $SimSILAv1$ indices provide the correct partitioning for all these data. It can be noted that the component $\hat{A}(\mathbf{x})$ strongly reduces values of the $SimSILAv1$ index when the number of clusters $K < c^*$. This is due to the fact that the exponent q in $\hat{A}(\mathbf{x})$ is calculated by the formula (4). Generally, the component $\hat{A}(\mathbf{x})$ improvements the results especially when the clustering algorithm combines clusters into larger ones and differences of distances between clusters are large. Then the influence of the separability measure is significant and consequently, it can strongly affect the value the index. On the other hand, when

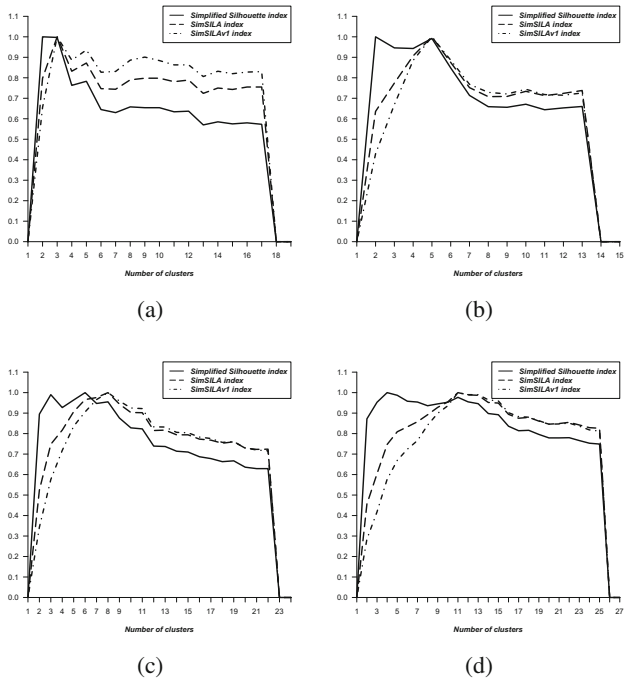

Fig. 3. Variations of the $Simplified\ Silhouette$, *SimSILA* and $SimSILAv1$ indices with respect to the number of clusters for 2-dimensional datasets: (a) *Data* 1, (b) *Data* 2, (c) *Data* 3, and (d) *Data* 4 partitioned by the *PAM* method.

$K > c^*$ the values of these new indices are increased slightly. It can be noted that the *Simplified Silhouette* and the *Silhouette* indices incorrectly select the number of clusters, whereas the new indices provide the right results for all the artificial datasets. The next experiments are related to the real datasets. As outlined above, the real datasets are numeric data: *Diabetes*, *Ecoli*, *Glass*, *Iris*, *Spectf*, *Wine*. In the experimental process, these datasets have been clustered by the *PAM* algorithm. Moreover, for the evaluation of the clustering validity, the six indices have been used. Table 2 shows the comparison of these indices taking into account the number of clusters, which is the configuration parameter for the clustering algorithm. In addition, the Table also includes results from previous experiments related to the artificial data. From the Table 2, it can be noted that for the real datasets the best results are achieved by the *SILA*, *SILAv1*, *SimSILA* and *SimSILAv1* indices. Moreover, for the *Glass* and *Iris* data, the results of the *SimSILAv1* index are better in comparison with other indices. Based on these results, it can be concluded that for all the experiments carried out on artificial and real data the best clustering results are selected by using these new indices.

Table 2. Comparison of the number of clusters obtained when using the *PAM* algorithm in conjunction with the *SIL*, *SILA*, *SILAv1*, *SimSil*, *SimSILA* and *SimSILAv1* indices. *N* denotes the actual number of clusters in the datasets.

Datasets	*N*	Number of clusters					
		SIL	*SILA*	*SILAv*1	*SimSIL*	*SimSILA*	*SimSILAv*1
Data 1	3	2	3	3	2	3	3
Data 2	5	2	5	5	2	5	5
Data 3	8	6	8	8	6	8	8
Data 4	11	4	11	11	4	11	11
Diabetes	2	2	2	2	2	2	2
Ecoli	8	4	4	4	4	4	4
Glass	6	2	2	7	2	7	7
Iris	3	2	2	2	2	2	3
Spectf	2	2	2	2	2	2	2
Wine	3	2	3	3	2	3	3

5 Conclusions

In this paper new indices called *SimSILA* and *SimSILAv*1 are proposed, which are the modification of the *Simplified Silhouette* index. As mentioned above, neither the *Simplified Silhouette* index nor the *Silhouette* index performs well when there are large differences of distances between clusters in a dataset. Similarly to the modification of the *Silhouette* index, the change of the *Simplified Silhouette* relies on the application of the additional component, which improves

the performance of the index. This additional component contains a measure of cluster compactness and reduces the high values of the index caused by large differences between clusters. In these conducted experiments, several datasets were used, where the number of clusters varied within a wide range. Moreover, the PAM clustering algorithm was selected for clustering of all the artificial and the real datasets. It has been noticeable that the $SILA$, $SILAv1$, $SimSILA$ and $SimSILAv1$ indices have provided the best results. However, the $Simplified$ $Silhouette$ index is much less computationally expensive than the $Silhouette$ index. From this perspective, the $SimSILA$ and $SimSILAv1$ indices have the competitive performance to the $SILA$ and $SILAv1$ indices in the selection of the right clustering results. All the presented results confirm the very high efficiency of the newly proposed indices.

References

1. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Prez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recogn. **46**, 243–256 (2013)
2. Bilski, J., Smoląg, J.: Parallel architectures for learning the RTRN and Elman dynamic neural networks. IEEE Trans. Parallel Distrib. Syst. **26**(9), 2561–2570 (2015)
3. Bilski, J., Wilamowski, B.M.: Parallel learning of feedforward neural networks without error backpropagation. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2016. LNCS (LNAI), vol. 9692, pp. 57–69. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39378-0_6
4. Bologna, G., Hayashi, Y.: Characterization of symbolic rules embedded in deep DIMLP networks: a challenge to transparency of deep learning. J. Artif. Intell. Soft Comput. Res. **7**(4), 265–286 (2017). https://doi.org/10.1515/jaiscr-2017-0019
5. Bradley, P., Fayyad, U.: Refining initial points for k-means clustering. In: Proceedings of the Fifteenth International Conference on Knowledge Discovery and Data Mining, pp. 9–15. AAAI Press, New York (1998)
6. Chang, O., Constante, P., Gordon, A., Singana, M.: A novel deep neural network that uses space-time features for tracking and recognizing a moving object. J. Artif. Intell. Soft Comput. Res. **7**(2), 125–136 (2017). https://doi.org/10.1515/jaiscr-2017-0009
7. Cpałka, K., Rebrova, O., Nowicki, R., Rutkowski, L.: On design of flexible neuro-fuzzy systems for nonlinear modelling. Int. J. Gen. Syst. **42**(6), 706–720 (2013)
8. Cpałka, K., Rutkowski, L.: Flexible Takagi-Sugeno fuzzy systems. In: Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, IJCNN (2005)
9. Devi, V.S., Meena, L.: Parallel MCNN (PMCNN) with application to prototype selection on large and streaming data. J. Artif. Intell. Soft Comput. Res. **7**(3), 155–169 (2017). https://doi.org/10.1515/jaiscr-2017-0011
10. Fränti, P., Rezaei, M., Zhao, Q.: Centroid index: cluster level similarity measure. Pattern Recogn. **47**(9), 3034–3045 (2014)
11. Gabryel, M.: A bag-of-features algorithm for applications using a NoSQL database. Inf. Softw. Technol. **639**, 332–343 (2016)

12. Gabryel, M., Grycuk, R., Korytkowski, M., Holotyak, T.: Image indexing and retrieval using GSOM algorithm. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2015. LNCS (LNAI), vol. 9119, pp. 706–714. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19324-3_63
13. Gałkowski, T.: Kernel estimation of regression functions in the boundary regions. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2013. LNCS (LNAI), vol. 7895, pp. 158–166. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38610-7_15
14. Galkowski, T., Pawlak, M.: Nonparametric estimation of edge values of regression functions. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2016. LNCS (LNAI), vol. 9693, pp. 49–59. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39384-1_5
15. Hruschka, E.R., de Castro, L.N., Campello, R.J.: Evolutionary algorithms for clustering gene-expression data. In: Fourth IEEE International Conference on Data Mining, ICDM 2004, pp. 403–406. IEEE (2004)
16. Jain, A., Dubes, R.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1988)
17. Ke, Y., Hagiwara, M.: An English neural network that learns texts, finds hidden knowledge, and answers questions. J. Artif. Intell. Soft Comput. Res. **7**(4), 229–242 (2017). https://doi.org/10.1515/jaiscr-2017-0016
18. Lago-Fernández, L.F., Corbacho, F.: Normality-based validation for crisp clustering. Pattern Recogn. **43**(3), 782–795 (2010)
19. Lichman, M.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2013). http://archive.ics.uci.edu/ml
20. Liu, H., Gegov, A., Cocea, M.: Rule based networks: an efficient and interpretable representation of computational models. J. Artif. Intell. Soft Comput. Res. **7**(2), 111–123 (2017). https://doi.org/10.1515/jaiscr-2017-0008
21. Meng, X., van Dyk, D.: The EM algorithm - an old folk-song sung to a fast new tune. J. Roy. Stat. Soc. Ser. B (Methodol.) **59**(3), 511–567 (1997)
22. Murtagh, F.: A survey of recent advances in hierarchical clustering algorithms. Comput. J. **26**(4), 354–359 (1983)
23. Pal, N.R., Bezdek, J.C.: On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Syst. **3**(3), 370–379 (1995)
24. Park, H.S., Jun, C.H.: A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. **36**(2), 3336–3341 (2009)
25. Rohlf, F.: Single-link clustering algorithms. In: Krishnaiah, P.R, Kanal, L.N. (eds.) Handbook of Statistics, vol. 2, pp. 267–284 (1982)
26. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
27. Rutkowski L, Cpałka K.: Compromise approach to neuro-fuzzy systems. In: Sincak, P., Vascak, J., Kvasnicka, V., Pospichal, J. (eds.) Intelligent Technologies - Theory and Applications. New Trends in Intelligent Technologies. Frontiers in Artificial Intelligence and Applications, vol. 76, pp. 85–90 (2002)
28. Rutkowski, L., Cpałka, K.: A neuro-fuzzy controller with a compromise fuzzy reasoning. Control Cybern. **31**(2), 297–308 (2002)
29. Saha, S., Bandyopadhyay, S.: Some connectivity based cluster validity indices. Appl. Soft Comput. **12**(5), 1555–1565 (2012)
30. Sameh, A.S., Asoke, K.N.: Development of assessment criteria for clustering algorithms. Pattern Anal. Appl. **12**(1), 79–98 (2009)

31. Serdah, A.M., Ashour, W.M.: Clustering large-scale data based on modified affinity propagation algorithm. J. Artif. Intell. Soft Comput. Res. **6**(1), 23–33 (2016). https://doi.org/10.1515/jaiscr-2016-0003
32. Shieh, H.-L.: Robust validity index for a modified subtractive clustering algorithm. Appl. Soft Comput. **22**, 47–59 (2014)
33. Starczewski, A.: A new validity index for crisp clusters. Pattern Anal. Appl. **20**(3), 687–700 (2017)
34. Starczewski, A., Krzyżak, A.: A modification of the silhouette index for the improvement of cluster validity assessment. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2016. LNCS (LNAI), vol. 9693, pp. 114–124. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39384-1_10
35. Starczewski, A., Krzyżak, A.: Improvement of the validity index for determination of an appropriate data partitioning. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2017. LNCS (LNAI), vol. 10246, pp. 159–170. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59060-8_16
36. Wu, K.L., Yang, M.S., Hsieh, J.N.: Robust cluster validity indexes. Pattern Recogn. **42**, 2541–2550 (2009)
37. Vendramin, L., Campello, R.J., Hruschka, E.R.: Relative clustering validity criteria: a comparative overview. Stat. Anal. Data Min. **3**(4), 209–235 (2010)
38. Zhao, Q., Fränti, P.: WB-index: a sum-of-squares based index for cluster validity. Data Knowl. Eng. **92**, 77–89 (2014)