http://doi.org/10.35784/iapgos.2596

A STEP TOWARDS THE MAJORITY-BASED CLUSTERING VALIDATION DECISION FUSION METHOD

Taras Panskyi, Volodymyr Mosorov
Lodz University of Technology, Institute of Applied Computer Science, Lodz, Poland

Abstract. *A variety of clustering validation indices (CVIs) are aimed at validating the results of clustering analysis and determining which clustering algorithm performs best. Different validation indices may be appropriate for different clustering algorithms or partition dissimilarity measures; however, the best suitable index to use in practice remains unknown. A single CVI is generally unable to handle the wide variability and scalability of the data and cope successfully with all the contexts. Therefore, one of the popular approaches is to use a combination of multiple CVIs and fuse their votes into the final decision. This work aims to analyze the majority-based decision fusion method. Thus, the experimental work consisted of designing and implementing the NbClust majority-based decision fusion method and then evaluating the CVIs performance with different clustering algorithms and dissimilarity measures to discover the best validation configuration. Moreover, the authors proposed to enhance the standard majority-based decision fusion method with straightforward rules for the maximum efficiency of the validation procedure. The result showed that the designed enhanced method with an invasive validation configuration could cope with almost all data sets (99%) with different experimental factors (density, dimensionality, number of clusters, etc.).*

Keywords: clustering, clustering validation index, decision fusion method

KROK W KIERUNKU METODY FUZJI DECYZJI OPARTEJ NA WIĘKSZOŚCI DLA WALIDACJI WYNIKÓW KLASTERYZACJI

Streszczenie. *Różnorodne indeksy walidacji klasteryzacji (CVI) mają na celu walidację wyników analizy skupień i określenie, który algorytm klasteryzacji działa najlepiej. Różne indeksy walidacji mogą być odpowiednie dla różnych algorytmów klasteryzacji lub miar niepodobieństwa podziału; jednak najlepszy walidacyjny indeks do zastosowania w praktyce pozostaje nieznany. Pojedynczy CVI na ogół nie jest w stanie poradzić sobie z dużą zmiennością i skalowalnością danych oraz z powodzeniem poradzić sobie we wszystkich kontekstach. Dlatego jednym z popularnych podejść jest użycie kombinacji wielu CVIs i połączenie ich głosów w ostateczną decyzję. Celem tej pracy jest analiza metody fuzji decyzji opartej na większości. W związku z tym prace eksperymentalne polegały na zaprojektowaniu i wdrożeniu metody NbClust fuzji decyzji opartej na większości, a następnie ocenianie wydajności CVIs za pomocą różnych algorytmów klasteryzacji i miar niepodobieństwa w celu odkrycia najlepszej konfiguracji walidacji. Ponadto autor zaproponował rozszerzenie standardowej metody fuzji decyzji oparta na większości o proste reguły dla maksymalnej efektywności procedury walidacji. Wynik pokazał, że zaprojektowana ulepszona metoda z inwazyjną konfiguracją walidacji może poradzić sobie z prawie wszystkimi zbiorami danych (99%) z różnymi eksperymentalnymi parametrami (gęstość, wymiarowość, liczba klastrów itp.).*

Słowa kluczowe: klasteryzacja, indeks walidacji klasteryzacji, metoda fuzji decyzji

Introduction

Clustering is a process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have a high natural association among themselves while remaining relatively distinct from each other [3]. In general, the essence of cluster analysis assumes that little or nothing is known about the grouping structure which underlies the data set. The operational objective, in this case, is to discover the grouping data structure which is frequently described as a problem of finding “natural groups”.

Many methods for cluster analysis have been developed in recent years and many of these methods have shortcomings and limitations in their practical use. It is difficult to provide a clear categorization of clustering methods because these categories may overlap so that a method may have features from several categories. Nevertheless, the major fundamental clustering methods can be classified into the following categories [9]: hierarchical methods formed [27, 47, 60], partitioning methods [15, 56, 62], density-based methods [10, 34], graph-based methods [1, 67], and grid-based methods [14, 17].

Different clustering algorithms usually lead to different partitions of data; even for the same algorithm, the selection of different input parameters may greatly affect the clustering results. Thus, effective evaluation standards and criteria are critically important to give the researcher confidence regarding the clustering results. The procedure of evaluating the correctness of clustering results is called cluster validation [32] and for a long time, it has been recognized as one of the vital problems essential to the success of data clustering.

It is usual to classify the cluster validation techniques under two groups — internal and external [35, 53]. External validation indices use external information not presented in the data to estimate the extent to which the clustering structure discovered by a clustering algorithm matches a certain external structure. On the other hand, internal indices evaluate the correctness of the clustering structure without reference to external information. Both external and internal validation indices are crucial for many application scenarios. However, there are still scenarios where clustering validation indices have limitations in estimating the correctness of clustering results. Examples include the case when external criteria are not available and internal indices are not robust enough. Moreover, despite the vast amount of expert endeavors spent on this issue, there is no consistent and conclusive solution to cluster validation. The multitude of different validation approaches creates an added difficulty, since results obtained using different methods cannot be compared in the same validation framework. Also, the relationship between different validation indices is not clear and has not been fully established.

A variety of indices aimed at validating the results of clustering analysis and determining which clustering algorithm performs best. However, the choice of the best or the most appropriate clustering validation index is strikingly similar to the dilemma of comparing and selecting the best classifiers in pattern recognition, where the no free lunch theorem rules that there is no universally best classifier [44]. Moreover, given the fact that different validation indices may be appropriate for different clustering algorithms or partition dissimilarity measures, the best suitable index to use in practice remains unknown. In the recent work by Gurrutxaga et al. [33], the authors accepted that there is no single way of establishing the quality of a partition by selecting the optimal validation index which would be more robust than the rest in all contexts and under different conditions. Therefore, Yera et al. [69] suggest using decision fusion validation strategies to obtain a more stable behavior which would make it possible for the user to avoid having to choose a different validation index for each particular environment.

In this work, the authors drew inspiration from the works of Arbelaitz et al. [2], Gurrutxaga et al. [33], Yera et al. [69], and the decision fusion method developed by Charrad et al. [16]. Since Charrad et al. [16] did not substantiate the use of a particular data, clustering algorithm, and a dissimilarity measure for the majority-based decision fusion voting procedure, the authors provide a comparative study of these factors. In the first part of the article, the authors propose to analyze Charrad's majority-based decision fusion method (MBDFM) using a *non-invasive configuration*. A non-invasive configuration relies on selecting the best MBDFM clustering algorithm and dissimilarity measure that works correctly in every experimental environment. Thereby, the authors try to achieve the best possible results for Charrad's MBDFM without changing the internal validation algorithm, but only its input validation parameters.

In the second part of this work, the authors propose to use an *invasive configuration* of Charrad's MBDFM. The authors hypothesize that the non-invasive configuration of MBDFM can be better than the default configuration of MBDFM (*k*-means algorithm and Euclidean distance), however, the authors suspect that this may not be enough to achieve the task, that is the revealing of the largest number of "true" clusters in the experimental setup. *What if the non-invasive configuration of MBDFM even with the best clustering algorithm and dissimilarity measure does not provide the expected satisfactory results*? The underlying idea of an invasive configuration is to change the MBDFM by interfering and modifying the internal Charrad's validation algorithm. Finally, the authors will show the difference between the MBDFM voting approaches with default, non-invasive and invasive configurations.

According to Arbelaitz et al. [2], there is no standard terminology and formalization for clustering validity indices; therefore, in this article, the abbreviation of CVI will be used for Cluster Validity Index. The next section discusses other works related to CVI comparison, in particular the examples of decision fusion methods found in the literature.

Since testing for revealing the data structure is the main objective of this article, the problem of choosing between the attribute space and the problem of discovering the optimal number of clusters will not be considered.

1. Related works

Most of the works that compare CVIs use the same approach: a set of CVIs is used to estimate the number of clusters in a set partitioned by several algorithms. Despite this widely used approach, most of the works are not comparable since they differ with respect to the compared CVIs, used data sets, or analysis results.

The paper published by Milligan and Cooper [52] compared 30 CVIs. The experiments were conducted using hierarchical clustering algorithms. They used 108 synthetic data sets with a varying number of non-overlapped clusters (2, 3, 4, or 5), dimensionality (4, 6, or 8), and cluster sizes. The same tabular format was used by Dubes [23]. Bezdek et al. [7] published a paper comparing 23 CVIs based on 3 runs of the EM algorithm and 12 synthetic data sets. Another study that compared 15 CVIs was performed by Dimitriadou et al. [22], based on 100 runs of the *k*-means algorithm for 162 data sets with binary attributes. Recently, Brun et al. [12] have compared 8 CVIs using several clustering algorithms: *k*-means, fuzzy *c*-means, SOM, single-linkage, complete-linkage, and EM, using 600 synthetic data sets with varying dimensionality (2 or 10), cluster shape, and number of clusters (2 or 4). Shim et al. [64] followed the Milligan and Cooper experiment but added certain CVIs or extended the study. Other CVI comparisons can be found where new CVIs are proposed; however, the experiments are usually limited to similar data sets comparing 5 or 10 CVIs [18, 37, 71]. The exception is a work by Arbelaitz et al. [2] based on the same Milligan and Cooper CVI framework, but with an extensive set of configurations (dissimilarity measure, data density, noise, overlapping clusters, etc.).

Since there is no universal CVI to always make a correct decision, many authors [44, 69] agreed to use multi-criteria solutions to reach the best and adequate results. Multi-criteria solutions assume the adoption of several CVIs to achieve greater certainty and correctness of clustering results. Bezdek and Pal [8] suggested a combined decision-based fusion strategy for all CVIs used in a validation process. This research includes the following decision-making methods and their rules: the mean, median, and mode rules. According to those rules, the final validation decision is made by the mean, median, or mode of CVIs that participated in the voting procedure. Later, the authors [44] showed a comparison of different fusion techniques of multiple CVIs. Moreover, Kryszczuk and Hurley [44] pointed that the best-performing scheme was the mean-rule decision fusion scheme. The recent work by Yera et al. [69] also discusses the use of decision fusion strategies for cluster validation purposes. The authors suggested two types of voting strategies: Global Voting and Selective Voting. Global Voting is a simple vote that fuses the decisions of all CVIs presented in the study. Selective Voting uses a limited group of CVIs for decision fusion. Moreover, three approaches were developed, each limiting the group of voting CVIs based on the following criteria: the global performance of the CVIs, their factor dependence success rate, and the impact the CVIs have on the results. The Yera et al. [69] decision fusion techniques presented above have certain critical disadvantages. The Global Voting approach was not even used in the comparative study due to the weakness of the archived results. This type of decision fusion technique is similar to the decision-making methods developed by Kryszczuk and Hurley. The Selective Voting technique with the global performance of CVIs did not meet the authors' expectations either, since the best vote could not beat the success rate of the best individual CVI. The Selective Voting technique with the factor dependence success rate of the CVIs beat the overall success rate of the best individual CVI. However, the improvement was slightly more robust than the best CVI involved in the voting procedure. Furthermore, the work uses a limited number of experimental factors (three clusters, at least 100 numbers of objects in each cluster, three dimensions, etc.) which can significantly affect the presented decision fusion approaches. Finally, these decision fusion strategies require weighting CVIs votes, which reduces the precision of estimated CVIs decisions.

Charrad et al. [16] suggested that other decision-making fusion methods should be used which a group of CVIs may use to seek a satisfying solution; namely, the authority rule and the majority rule. The authority rule refers to groups that have a leader, i.e. the main CVI which has the authority to make the ultimate decision for the entire group. Although the method can generate a final decision quickly, it does not encourage the maximization of the strengths of individual CVIs in the group [46]. The majority rule depends on an individual decision of each CVI, where the final decision is made by the majority of the total CVIs votes. This method delivers fast solutions and follows a clear rule of using independent CVIs in the validation process.

In light of the decision fusion techniques presented above, taking all of the above multi-criteria methods into consideration, the MBDFM has been chosen as a major scheme for further analysis and improvement.

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

2. Tools for multi-criteria decision fusion clustering validation

A significant amount of software is available for data clustering validation purposes. Interestingly, most of the sophisticated software on clustering validation is open-source software, which is freely available at different Web sites. On the other hand, most of the commercial software comprises implementations of simpler and more classical validation solutions. This is because open-source software is often in the form of research prototypes, created by researchers, which reflect more recent advances in the clustering validation field. The clustering validation procedure has been implemented as packages in many software applications, such as *SAS*, *RapidMiner*, *MATLAB*, and *R*. However, only the *R* programming environment offers a large number of unique CVIs and different validation approaches. Moreover, in addition to the object-oriented nature of the language, implementing the CVIs within the *R* statistical programming framework provides the additional advantage in that it can interface with numerous clustering algorithms in existing *R* packages, and accommodate new algorithms as they are created and coded into R libraries.

There are several *R* packages that perform clustering validation and are available from https://www.r-project.org/. Indeed, Milligan and Cooper [52] examined thirty CVIs, with simulated artificial data, where the number of clusters was known beforehand. Eleven CVIs among them are available in the *R cclust* package [21], eight CVIs in the *clusterSim* package [68], two CVIs in *clv* [54]. The *clValid* package [11] includes 3 internal CVIs, 4 stability CVIs (special versions of internal indices), and 2 biological CVIs. The *cl_validity()* function in the *clue* package [42] performs validation for both partitioning and hierarchical methods using 3 CVIs, and the *fcclusterIndex()* function in package *e1071* [50] has built-in 7 fuzzy CVIs. The *cluster.stats()* function in the *fpc* package [41] uses 8 CVIs for clustering validation purposes. The package *NbClust* [16] gathered the 26 CVIs, several clustering algorithms with corresponding dissimilarity measures together to provide an exhaustive list of CVIs. Currently, NbClust is only one package that offers such a variety of CVIs, however, some indices examined in the Milligan and Cooper study were not implemented due to a lack of detailed CVI's explanation. Moreover, unlike the rest packages, the NbClust is quite flexible and offers the possibility of a fairly broad change of input parameters for further clustering validation purposes. Therefore, Charad's NbClust is chosen as the basic MBDFM for further modifications. For clarification purposes, instead of Charad's NbClust MBDFM the abbreviation MBDFM will be used in the remainder of this article.

3. Majority-based decision fusion method notation

This note studies a method of the CVI's decision formation and aims to explain the stylized fact that the support for one out of k clusters at stake often shows a high degree of CVI's heterogeneity and persistent cross-sectional variance that is only partly explained in clustering conditions. An intuitive explanation of this stylized fact is that each CVI may show a tendency to conform to the vote of what it perceives to be the best opinion. As postulated before that the behavior of each CVI in formulating the overall decision could be described by the majority rule. Although the authors do not model the voting process of CVI's in detail, but only show that the reduced forms are consistent with an explicit clustering validation foundation. The clustering validation theory does not explain how each CVI votes to conform to the majority. Moreover, the theory does not describe the growing or reducing CVI's tendency towards consensus in the decision scheme. Therefore, the first objective is to formulate the simple model of the majority-based decision fusion rule.

The CVIs decision formation process is defined as follows. Suppose the clustering algorithm run over the data set $\mathbb{X}$ with a set of m different values for the k parameter $K = \{k_1, k_2 \dots, k_m\}$, and let $\mathcal{P} = \{P_1, P_2, \dots, P_m\}$ be the k partitions. Suppose the clustering algorithm reveals the k_{est} to be the best number of clusters with the corresponding P_{est} partition. The true number of clusters k_{true} is known beforehand. Let CVI= $\{CVI_1, CVI_2, \dots, CVI_n\}$ be a set of n clustering validation indices CVI_i, where $i = 1 \dots n$ which are to be analyzed. Each CVI_i return the value of $CVI_i(P)$ for the proposed partition over all the partition $\mathcal{P}$. Moreover, the returned $CVI_i(P)$ value indicates the specific k_j used as an input parameter for the clustering algorithm. Each $CVI_i(P)$ value is counted as a vote for a particular k_j. Thereafter, the method counts the CVIs votes for each k_j and forms the CVI's decision groups DEC $= \{Dec_1, Dec_2, \dots, Dec_k\}$, where Dec_j is a sum of CVI's votes for particular k_j. From among the CVI's decision groups is formed the biggest group or the majority $Maj_{est} = max_{Dec\in\mathbb{N}}\{Dec_1, Dec_2, \dots, Dec_k\}$ that should identify the "true" estimated number of clusters $k_j = k_{est}$. However, only if the Maj_{est} identifies the $k_{est} = k_{true}$ clusters, the majority justifies that the estimated number of clusters is the "true" one, and hence $Maj_{est} = Maj_{true}$.

4. Experimental setup

Before analyzing and modifying the majority-based decision fusion method the experimental setup should be outlined. In this section, the authors describs the experiment setup including 24 CVIs: *Calinski-Harabasz index* [13], *J-index* [24], *pseudo T-squared* [25], *C-index* [43], *F-ratio* [6], *CCC criterion* [61], *Ptbiserial index* [51], *DB index* [19], *Frey index* [30], *Harigan index* [40], *Ratkowsky* and *Lance index* [57], *Scott* and *Symons index* [63], *Marriot index* [48], *Ball* and *Hall index* [4], *TrCovW* and *TraceW indices* [52], *Friedman* and *Rubin ind*ices [31], *McClain* and *Rao i*ndex [49], *KL index* [45], *Silhouette in*dex [59], *Dunn index* [26], *Halkidi indices* [38, 39]. Since the CVIs are compared in a wide variety of configurations, an experiment with several factors has been designed.

The authors' proposal follows, to a certain extent, the traditional problem of estimating the number of clusters in a data set, which was described well in Arbelaitz et al. [2]: to run a clustering algorithm over a data set with a set of different values for the k input parameter, to obtain a set of different partitions, and to evaluate each particular CVI for all obtained partitions. The detected number of clusters in the target partition yielding satisfactory results is considered a decision of the CVI for that particular data set. However, the decision is considered successful only if it justifies that the estimated number of clusters is "true".

Eight agglomerative hierarchical clustering algorithms were used to compute partitions from the data sets: *Ward*, *single-linkage*, *complete-linkage*, *average-linkage*, *mcquitty*, *median*, and *centroid*. The *k*-means, one of the most commonly adopted partitioning algorithms, has also been used. These clustering algorithms are well known; moreover, it is easy to obtain different partitions by modifying the input parameter that controls the number of clusters of an output partition. Each clustering algorithm will be used to compute a set of partitions with the number of clusters ranging from 2 to 10. From the perspective of dissimilarity measures, the comparison analysis will also be conducted. Five dissimilarity measures for each particular clustering algorithm will also be used: *Euclidea*n, *maximum, Manhattan*, *Canberr*a, and *Minkowski* distances.

To evaluate the performance of the 24 CVIs, 90 artificially generated data sets will be created. Most of the synthetic numerical data sets will be generated using the mixture models of the Gaussian distribution but with different parameters. Furthermore, 10 benchmark data sets (the true number of clusters is known a priori from the literature) drawn from the literature sources as well as from available *UCI* and *Kaggle* repositories will also be analyzed (see Table 2). The synthetic data sets were created to cover a large number of factor combinations such as: the number of clusters (K), the minimum (n_{min}) and maximum (n_{max}) number of objects in a data set, cluster density (den), and dimensionality (dim). The values of the parameters used to create the synthetic data sets are shown in Table 1.

Table 1. Values of the parameters used for generating the synthetic data sets

Parameter	Value
n_{min}	100
n_{max}	6000
K	2…10
dim	2…4
den	1…4

Since 90 synthetic data sets will be created, 4050 configurations have been obtained by multiplying this value by 5 partition dissimilarity measures and 9 clustering algorithms. In the case of benchmark data sets, the experiment is based on 450 configurations — 10 data sets, 9 clustering algorithms, and 5 partition dissimilarity measures. Considering the synthetic and benchmark data sets and taking into account the different number of partitions computed for each data set, each of the 24 CVIs should be computed for 40500 partitions.

Data sets

In synthetic numerical data sets, the clusters are non-overlapping represented as multivariate finite mixtures. The synthetic data sets were created without introducing overlapping, noise, or missing data objects. Imprecise and noisy data with overlapping clusters could be distorted as compared to human intuition [55]; therefore, noise and overlap level factors are excluded from this experimental setup.

Table 2. Characteristics of the benchmark data sets

Data set	Number of clusters
Steinley [65]	5
G2-set [29]	2
Unbalance 1 [58]	3
Unbalance 2 [58]	5
Square1 [36]	4
Triangle1 [36]	4
AD_5_2 [5]	5
AD_10_2 [5]	10
Haberman-survival (Kaggle)	4
Iris (UCI)	3

The "true" number of clusters in synthetic data sets ranges from 2 to 10 depending on the set, with cluster sizes from 50 to 3000 data objects per cluster. Furthermore, 90 synthetic data sets were generated with an uneven number of clusters per data set, namely: 4 data sets with 2 clusters, 8 data sets with 3 clusters, 22 data sets with 4 clusters, 20 data sets with 5 clusters, 10 data sets with 6 and 7 clusters, 6 data sets with 8 and 9 clusters, 4 data sets with 10 clusters. Afterward, about half of them (47%), which is 42 data sets, were generated with 4 or 5 clusters.

5. Common majority-based decision fusion method dubious scenarios

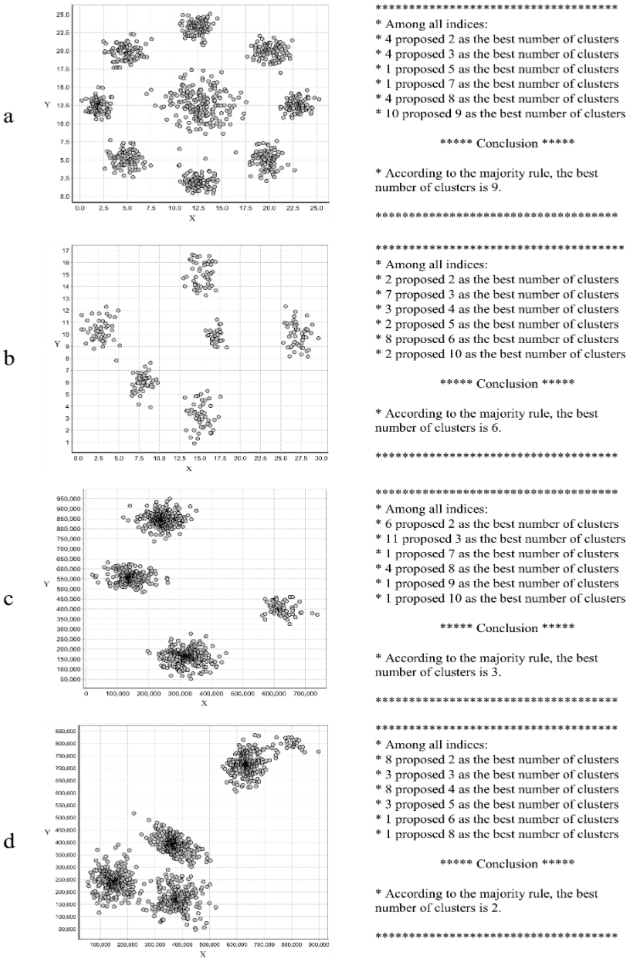
Passing through the testing phase repeatedly, there is not always a clear distinguishing line between all the majority situations. Taking into consideration the experimental results, the most common 4 cases of MBDFM controversial situations are presented. All scenarios were obtained by the default clustering validation configuration, i.e. *k*-means clustering and Euclidean distance.

Scenario 1: The decision is made by the relative CVI's majority, and the nearest alternative is 50% votes less than half of the majority one. The relative majority points to the "true" number of clusters. The example of Scenario 1 and the accompanying data set presented to enhance reader understanding of CVIs voting is shown in Figure 1a.

Scenario 2a: Situation when the decision is taken by the CVI's relative majority, and the nearest alternative is 50% votes more than half of the majority one. The relative majority points to the "true" number of clusters. The example of Scenario 2a is shown in Figure 1b. This scenario requires additional MBDFM verification for the final statement. In a validation configuration presented in Figure 1b, an almost equal number of CVIs voted for the 3 and 6 clusters. This scenario shows the controversial situation with no clear-cut majority. However, following the hard logic, the relative CVI's majority points to the 6 clusters to be the "true" ones.

Scenario 2b: The scenario where the relative CVI's majority points to the incorrect number of clusters. The nearest alternative group of CVIs in turn shows the „true" number of clusters. The example of the Scenario 2b case is shown in Figure 1c. The most critical of all the previous scenarios requires complete and precise MBDFM verification. Due to the limited facility of the classifier (Euclidean distance) and the crisp nature of the *k*-means algorithm, the results are completely misleading. This scenario shows the majority of CVIs voted for 3 clusters, and the nearest alternative voted for 4 numbers of clusters to be the "true" ones. The MBDFM is data-dependent since different CVIs behave differently on different data sets in various environments. The majority evaluation works on the fundamental assumption that the clustering algorithm works correctly. If this assumption does not hold, there could be a "fake" majority that identifies a false "true" number of clusters. Moreover, the lack of knowledge of the "true" number of clusters has a detrimental effect on clustering quality. Clustering in real-life applications is executed in a black-box fashion. The analyst is usually unable to correctly determine the "true" number of clusters beforehand. Therefore, Scenario 2 has been divided into two sub-cases.

Scenario 3: The data sets for which no majority prevails. In situations when no majority exists and two equal groups of CVIs voted for a different number of clusters to be the "true" ones, however, only one group of CVIs is correct and the other is misleading. The example of Scenario 3 is shown in Figure 1d. The presented scenario is a natural CVIs bias in favor of the status quo. However, according to the MBDFM, the "true" number of clusters among two equal groups of CVIs, is the one that is the first in the list. In this case, the function decided that the majority of CVIs voted for 2 as the best number of clusters despite the same number of CVIs cast the vote for 4 clusters.



```
*************************************
* Among all indices:
* 4 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 4 proposed 8 as the best number of clusters
* 10 proposed 9 as the best number of clusters

            ***** Conclusion *****

* According to the majority rule, the best
number of clusters is 9.

*************************************
```

```
*************************************
* Among all indices:
* 2 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 3 proposed 4 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 8 proposed 6 as the best number of clusters
* 2 proposed 10 as the best number of clusters

            ***** Conclusion *****

* According to the majority rule, the best
number of clusters is 6.

*************************************
```

```
*************************************
* Among all indices:
* 6 proposed 2 as the best number of clusters
* 11 proposed 3 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 4 proposed 8 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters

            ***** Conclusion *****

* According to the majority rule, the best
number of clusters is 3.

*************************************
```

```
*************************************
* Among all indices:
* 8 proposed 2 as the best number of clusters
* 3 proposed 3 as the best number of clusters
* 8 proposed 4 as the best number of clusters
* 3 proposed 5 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 1 proposed 8 as the best number of clusters

            ***** Conclusion *****

* According to the majority rule, the best
number of clusters is 2.

*************************************
```

Fig. 1. Two-dimensional plot of synthetic data sets used in the experiment broken down by the default MBDFM configuration: (a) corresponds to the Scenario 1 situation with nine globular-form "true" clusters. (b) corresponds to the Scenario 2a situation with six globular-form "true" clusters. (c) corresponds to the Scenario 2b situation with four globular-form "true" clusters, however, according to the default MBDFM configuration, the revealed number of clusters is equal to three. (d) corresponds to the Scenario 3 situation with four globular and elongate form "true" clusters, however, according to the default MBDFM configuration the revealed number of clusters is equal to two

Summarizing all the above scenarios and CVI's results, doubts are expressed whether the MBDFM with default configuration is feasible in all clustering situations. Moreover, even a clear-cut data grouping structure (compact and far from other clusters) can easily deceive the 24 CVIs. Interestingly, there are less than 15% (15 data sets) of all the cases in the experimental setup, where the absolute majority indicates the "true" number of clusters. Thus, the cause must be sought in the NbClust majority rule, in the expediency, the classifier, and clustering algorithm, but do not question the quality and the correctness of each of the 24 CVIs. The MBDFM with the default validation configuration should maintain the basis for further clustering improvement. Thus, a non-invasive configuration of MBDFM is proposed to replace the default configuration, for all controversial scenarios.

6. Non-invasive validation configuration

6.1. Clustering method

Recent research focuses on clustering analysis to understand the strengths and weaknesses of various clustering algorithms in terms of data factors. As has been mentioned before, certain data characteristics may strongly affect clustering analysis, including high dimensionality, noise, types of attributes, and scales [66]. That being said, the authors have studied the clustering validation procedure by answering the question: *How to choose the best clustering algorithm appropriate for the MBDFM*? Considering that there are numerous clustering algorithms proposed in the literature, especially after an algorithm boom in the data mining area, it is arguable which clustering algorithm is the most suitable for the MBDFM.

Fig. 2 shows the percentage of correct guesses achieved by all 24 CVIs, which are sorted by the success score. Notice that this percentage refers to 194400 configurations: 100 synthetic and benchmark data sets, 9 clustering algorithms, 9 partitions, and 24 CVIs. All presented clustering algorithms are examined with the Euclidean dissimilarity measure as a distance metric. Correct guesses are considered as CVI votes that identify the "true" number of clusters. In brief, the Boolean value for each CVI is obtained; a correct CVI guess corresponds to 1 and an invalid one to 0. The sum of correct guesses forms the majority. Furthermore, only scenarios 1 and 2a are in favor of proper situations, and scenarios 2b and 3 are considered to be invalid. Indeed, Scenario 2a is initially classified along with Scenario 1 as a part of the group that forms the overall success score. Moreover, after the comparative analysis, all Scenario 1 data sets identify the true number of clusters correctly, and therefore, the MBDFM does not require any additional adjustment.

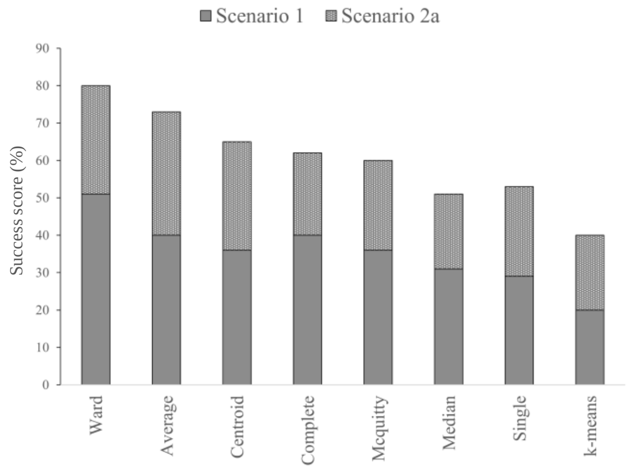


Fig. 2. Overall success score for data sets broken down by a clustering algorithm

Although we can find a clear pattern, it seems that the overall comparative results are severely affected by the used clustering algorithm. Assuming that all the potential candidates from Scenario 2a become a fully-fledged majority that point to the "true" number of clusters, most CVIs obtain their worst results for the *k*-means algorithm, i.e. 40% (Scenario 1 – 20% + Scenario 2a –20%), while Ward shows the highest success score of 80% (Scenario 1 – 51% + Scenario 2a – 29%). If we focus on *k*-means and Ward only, this factor shows drastically different results and an accuracy difference of 40%. On the other hand, the result for the centroid algorithm of 65% (Scenario 1 – 36% + Scenario 2a – 29%), the complete-linkage algorithm of 62% (Scenario 1 – 40% + Scenario 2a – 22%), and the mcquitty algorithm of 60% (Scenario 1 – 36% + Scenario 2a – 24%) reduce the differences between the CVIs decisions and balance the overall success score.

Table 3. Overall success-failure score (%) of the majority-based decision fusion method for all data sets with the Euclidean dissimilarity measure broken down by clustering algorithms

	Scenario 1	Scenario 2a	Scenario 2b	Scenario 3
Ward	51%	29%	9%	11%
Single	29%	24%	47%	0%
Complete	40%	22%	31%	7%
Average	40%	33%	22%	5%
Mcquitty	36%	24%	36%	4%
Median	31%	20%	40%	9%
Centroid	36%	29%	26%	9%
k-means	20%	20%	49%	11%

The situation becomes more interesting and clearer after the analysis of all scenarios (see Table 3). According to the previous definition of success, failure is defined as an incorrect decision of the CVI. Thus, the failure score is a total sum of CVI scores that form Scenario 3 situations, or wrong and "fake" majority (Scenario 2b) that differs from the "true" number of clusters. With respect to the failure score, Scenario 3 with the Ward clustering algorithm has reached (11%); the next are the median and centroid clustering algorithms with (9%) each and the last is the single-linkage clustering algorithm (0%). The results of Scenario 2b are the most valuable and controversial at the same time. As can be seen, *k*-means shows the highest failure score (49%) and Ward – the lowest (9%) one. Summing up, according to the presented results, the Ward clustering algorithm is the only obvious rational choice for further validation research.

6.2. Dissimilarity measure

Using an explicit dissimilarity measure to guide the validation process is a very popular approach, adopted by many widely-used clustering algorithms. Unfortunately, there are no definitive rules on which measure to choose for a particular problem. Dissimilarity measures should be considered in the context of the study where they are to be used, including the nature of data and the type of analysis. However, certain general guidelines do exist, i.e. the nature of data should strongly influence the choice of the dissimilarity measure; the choice of dissimilarity measure should depend on the scale of the attributes; the clustering algorithm should influence the choice of the dissimilarity measure. It could be considered a fatal defect in the validation procedure if too many dissimilarity measures have to be taken into consideration; however, it might be felt that a wide variety of possible measure choices is an advantage making the validation procedure usefully flexible.

It is hard to choose the most appropriate dissimilarity measure for a given clustering task without a preliminary experiment. Various dissimilarity measures presented in this article can be considered for use with all the presented clustering algorithms that are flexible enough not to be tied to a particular measure. It makes it possible to choose carefully based on the available domain knowledge and to verify the effects of several candidate measures experimentally.

Instead, a comparative study of 5 dissimilarity measures has to be conducted in the clustering verification process. The analysis will be focused on an appropriate choice of a dissimilarity measure in Ward's algorithm since the rest of the clustering algorithms have previously been rejected.

As in the case of selecting the clustering algorithm, scenarios 1 and 2a are the proper situations and scenarios 2b and 3 are considered to be incorrect. Figure 3 shows that the selected partition dissimilarity measure moderately affects the behavior of the CVIs, not as extremely as in the case of the clustering algorithm. Two of the presented dissimilarity measures, i.e. Minkowski and Euclidean, follow the overall pattern with 80% of correct guesses. The maximum dissimilarity measure shows slightly better results with an 81% success score. Furthermore, the Canberra dissimilarity measure yields extremely good results – 85% of correct guesses.

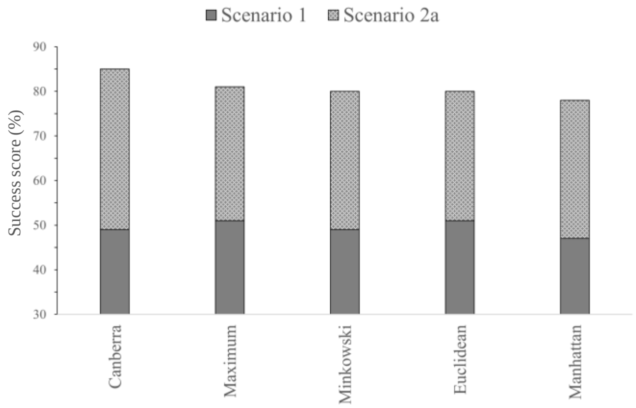


Fig. 3. The overall score for data sets broken down by a dissimilarity measure

In terms of the failure score, the results show that the difficulty imposed by the bias situations (Scenario 3) could be seen to a relatively small extent (9% – 11%) in all dissimilarity measures (see Table 4). The Canberra dissimilarity measure notifies the 5% of the detected Scenario 3 cases. Considering the contribution of Scenario 2b situations to the overall results, the Euclidean, maximum, and Minkowski dissimilarity measures should be mentioned as ones with the lowest (9%) failure scores. Mostly, false decisions were made using the Manhattan dissimilarity measure (13%). The Canberra dissimilarity measure shows slightly better results (10%) than Manhattan one.

Table 4. Overall success-failure score (%) of the majority-based decision fusion method for all data sets with the Ward clustering algorithm broken down by a dissimilarity measure

	Scenario 1	Scenario 2a	Scenario 2b	Scenario 3
Canberra	49%	36%	10%	5%
Maximum	51%	30%	9%	10%
Minkowski	49%	31%	9%	11%
Euclidean	51%	29%	9%	11%
Manhattan	47%	31%	13%	9%

In conclusion, the experiments show sufficiently moderate evidence for choosing a dissimilarity measure that is significantly better than the rest. However, the Ward clustering algorithm with the Canberra dissimilarity measure is recommended as the best for non-invasive MBDFM configuration settings.

7. Invasive validation configuration

Despite all attempts to improve MBDFM using the non-invasive configuration by way of altering the dissimilarity measure and the clustering algorithm, the best overall success score remains at the 85% with the Ward clustering algorithm in conjunction with the Canberra dissimilarity measure. Moreover, attempts to improve the result, the validation of the CVIs number were also carried out. Unfortunately, adding 2 more CVIs to 26 or subtracting 2 CVIs to obtain 22 did not change the overall success score. Of course, if the number is significantly changed from 24 to 10 CVIs, the result will also change. However, then another problem appears, namely the expediency and application correctness of each CVI individually. The smaller the number of CVIs the greater responsibility and user trust lies with each of them. In this context, the first step is to justify and select the best group of CVIs, and only then hold the voting procedure. The disadvantage of this approach is the very fact of CVIs division into best and worst. Furthermore, changing the input data can dramatically turn the situation and the best ones may become worst and vice versa. Taking all of the above into consideration, the main way to improve the results is to modify the MBDFM using the invasive configuration.

The non-invasive MBDFM relies on selecting the best clustering algorithm and dissimilarity measure to ensure optimal validation results. These two input MBDFM parameters could be tuned by the researcher. Moreover, the researcher could tune the third parameter – cluster range. The "true" number of clusters k_{true} are located across the range $K = \{2, \dots, 10\}$. However, the cluster range should be changed with great caution, as ill-considered change can lead to critical consequences. Such a change could eliminate the "true" number of clusters from the validation procedure and further the researcher, without knowing it, will look for the "true" number of clusters in a range of knowingly fake ones. All the following partitions will reveal the fake number of clusters, moreover, all of the CVIs will be forced to vote for the fake number of "true" clusters which will lead to erroneous MBDFM decisions. In real-world validation issues, the "true" number of clusters is unknown a priori and the researcher without knowing this fact forcibly restricts the CVI's possible decisions in the frames of cluster range. Whether the CVI is good or its decision is far from optimal, it should cast the vote only for the particular number of clusters in the prescribed cluster range. The role of the initial cluster range is extremely high as a broad cluster range gives more freedom to each of CVIs in their voting, however, the validation procedure is becoming fuzzy. On the other hand, the excessive compression of cluster range gives less freedom of CVI's votes cast for a particular cluster, but the validation procedure is becoming crisp.

The authors found a clear pattern in cluster range modification which allows to safely reduce it. The cluster range remains the same at the beginning of the MBDFM procedure. Figure 1d describes the operation of MBDFM with decision groups of CVIs that cast votes for the particular cluster. The authors noticed, that none of the CVIs cast a vote for 9 and 10 clusters. None of the 24 CVIs presented in the experimental setup even with possessed decision capabilities doesn't vote for these clusters. Thereby, the upper bound of the cluster range could be safely reduced – $K = \{2, \dots, 8\}$. Moreover, if none of the CVIs cast votes for clusters in the lower bound, it allows reducing the cluster range on the other side. This procedure will efficiently distribute the CVI's decisions across the optimized cluster range. The authors hypothesized, if none of the CVIs cast votes for clusters in the upper or lower bound of the cluster range, it is reliably confirmed that these clusters cannot become the candidates to be the "true" ones. However, it is not always possible to optimize the cluster range. If one of the CVIs cast one single vote for the particular cluster in the lower or upper bound of the cluster range, it cannot be taken lightly (see Figure 1b, 1c). That is, at this stage, all CVI's votes are taken into account and the cluster range is reduced by cutting its lower and upper bounds i.e. the clusters for which none of the CVIs cast a vote.

It should also be noted that only the lower and upper bound of the cluster range should be reduced without dividing the cluster range into two or more subranges. Figure 1c shows that none of the CVIs cast votes for the 4, 5, and 6 clusters. However, these clusters are located within the initial cluster range, and subtracting them leads to the cluster range division into $K_1 = \{2 \dots 3\}$ and $K_2 = \{7, \dots, 10\}$ subranges. Analyzing the cluster subranges separately gives the researcher two or more "true" clusters where only one is "true" and the others – fake "true". Moreover, Figure 1c will show the erroneous CVI results, since the researcher will subtract the chance of CVIs to vote for the four clusters to be the "true" one. Therefore, in this stage, the initial cluster range should remain integral, even if none of the CVIs cast a single vote to the cluster located within the cluster range. The cluster range optimization procedure applies to the clusters without any votes cast by CVIs only in the lower and upper bound of the initial cluster range.

The cluster range optimization procedure is not always possible to conduct. Figure 1a shows that CVI's votes are cast for clusters that completely cover the initial cluster range $K = \{2, \dots, 10\}$. This, in turn, does not allow to carry out optimization procedure in the way described above, however, gives the chance to consider another cluster range optimization strategy. This strategy is based on the CVI's votes cast for the relative majority $Maj_{est}^{\wedge}$ and its nearest alternative Maj_{est}^{*}. The relative majority $Maj_{est}^{\wedge}$ estimated by the MBDFM with the non-invasive configuration, could point to the "true" number of clusters (Scenario 1, 2a) but also may indicate the fake number (Scenario 2b). The nearest alternative Maj_{est}^{*} is the second-largest decision CVI's group which is closest to the "true" number of clusters. If the first strategy optimizes the cluster range by means of cutting its lower and upper bound and subtracting the cluster with no CVI's vote, the second one is entirely based on the CVI's majority and its nearest alternative. The authors' hypothesizes, that $Maj_{est}^{\wedge}$ and Maj_{est}^{*} with its corresponding $k_{est_j^{\wedge}}$ and $k_{est_j^{*}}$ should become the upper/lower bound of K. In Figure 1b in the second optimization step, the cluster range will become $K = \{3, \dots, 6\}$, where the relative majority $Maj_{est}^{\wedge}$ will become its upper bound with corresponding $k_{est_5^{\wedge}} = 6$ and the nearest alternative Maj_{est}^{*} will become its lower bound with corresponding $k_{est_2^{*}} = 3$ clusters. That is, at this stage, all CVI's votes that were excluded from the validation procedure by means of cluster range reduction will forcibly cast the votes only for clusters within the new optimized cluster range. The authors assume that all CVI's votes excluded from the initial cluster range will strengthen the final decision – the majority that points to the "true" number of clusters in the optimized cluster range. Furthermore, even if MBDFM with a non-invasive configuration will return the final decision of fake "true" number of clusters (Scenario 2b), the optimization procedure will help the CVI's votes to steer their decisions in the direction of the "true" number of clusters. This optimization strategy assumes that the "true" number of clusters should be located within the new cluster range (with Maj_{est}^{*} and $Maj_{est}^{\wedge}$ the upper/lower bound of K), and all CVI's votes that were excluded from the optimized cluster range will be forcibly asked to cast their votes only for clusters within the new range K.

Each CVI formulates a vote that favors one of the 9 clusters at stake. The dynamic process that characterizes each CVI vote formation is based on the idea that the CVI's majority reveals the particular clusters probabilistically. The formation process of the CVI's votes strongly depends on the cluster range. It is assumed that the optimization procedures described above could change the CVI's votes in favor of the "true" number of clusters, even if before the optimization procedures some of CVIs could cast their votes for the fake number of "true" clusters.

These optimization procedures are carefully collected and written in the form of the majority MBDFM rule (*i*), which aims at revealing the biggest number of "true" clusters in the experimental setup. Using Scenario 1 2a and 2b cases from non-invasive MBDFM configuration as our input data, the enhanced majority rule can be written as follows.

Rule (*i*):

1. Run the MBDFM with a non-invasive configuration. Let DEC= $\{Dec_1, Dec_2, \dots, Dec_k\}$ and Scenario 1, 2a, or 2b is considered.
2. Reveal the number of clusters $k_{est_j^{\wedge}}$ and $k_{est_j^{*}}$ that corresponds to the relative majority $Maj_{est}^{\wedge}$ and its nearest alternative Maj_{est}^{*} respectively.
3. Optimize the cluster range by means of reducing the upper and lower bound of K when some of the decision groups Dec_i do not reveal any number of clusters. If no majority prevails, use rule (*ii*).
4. Repeat step (2). If MBDFM reveals the absolute majority $Maj_{est}^{\wedge}$, assume it corresponds to the "true" number of clusters, then $k_{est_j^{\wedge}} = k_{true}$, otherwise, if the absolute majority was not achieved move to the next step.
5. Change the cluster range, where $Maj_{est}^{\wedge}$ and Maj_{est}^{*} and its corresponding $k_{est_j^{\wedge}}$ and $k_{est_j^{*}}$ becomes the upper/lower bound of K.
6. Rep`eat step (2). If the MBDFM reveals $k_{est_j^{\wedge}} = k_{true}$, the estimated number of clusters is the "true" one; otherwise, if the $k_{est_j^{\wedge}} \neq k_{true}$, the majority $Maj_{est}^{\wedge}$ identifies the fake "true" number of clusters. Moreover, if no majority prevails, rule (*i*) did not give the expected results and should not be used in the MBDFM invasive configuration.

The overall success score is fully justified and confirmed, since rule (*i*) of MBDFM with an invasive configuration applied to the Scenario 1, 2a, and 2b cases (49% for scenario 1, 36% for scenario 2a, 10% for scenario 2b) approves the 95% of correct guesses. All Scenario 2b controversial situations (10%) have been solved in favor of the "true" number of clusters. The rule (*i*) shows sufficiently strong evidence to adopt it into the MBDFM's default clustering validation decision scheme to enhance the NbClust majority voting procedure.

Nevertheless, there is a group (Scenario 3) of about 5% of all data sets that seems to show the questionable bias situations even under the MBDFM invasive validation configuration applied with the rule (i). It should be emphasized that Scenario 3 mirrors the situation where no majority prevails when the two biggest decision groups $Dec_{est}^{1} = Dec_{est}^{2}$ have equal votes cast by CVIs. In this case, only one decision group among them specifies the "true" number of clusters. Scenario 3 situations require not only the confirmation of the relative majority correctness (as in Scenario 2a data sets) but thorough analysis and modification of the decision-making scheme in general. Due to the work limitations which cannot embrace every data case, the contentious situations, therefore, could appear for other data sets not examined in the experimental setup.

The MBDFM with an invasive configuration and rule (*i*) should be applied to Scenarios 1,2a and 2b, where the majority of CVI's votes prevails. However, rule (*i*) could not be used in the case of Scenario 3. The strategy of optimization the cluster range, by means of excluding the CVIs that cast no votes in the upper or lower bound of the initial cluster range, remains the same. However, the strategy of optimizing the cluster range by means of revealing the majority and its nearest alternative will be modified. Rule (*ii*) is created to cope with scenario 3 situations. The authors hypothesize, that Dec_{est}^{1} and Dec_{est}^{2} with their corresponding $k_{est_j^{\wedge 1}}$ and $k_{est_j^{\wedge 2}}$ should become the upper/lower bound of K. In Figure 1d this optimization procedure will produce the new cluster range $K = \{2, \dots, 4\}$, where the Dec_{est}^{1} will become its lower bound with corresponding $k_{est_1^{\wedge 1}} = 2$ and the Dec_{est}^{2} will become its upper bound with corresponding $k_{est_3^{\wedge 2}} = 4$ clusters. All CVI's votes excluded from the initial cluster range will be forcibly asked to cast their votes for clusters in the optimized range. The optimization procedure in Scenario 3 cases mainly aimed at revealing the majority between two equal groups of CVIs. Therefore, rule (*ii*) aims to effectively imbalance the votes divided between the two biggest groups of CVIs and tips the scales in favor of the decision group that points to the "true" number of clusters. Finally, the optimized cluster range applied in a rule (*ii*) should strengthen the final CVI's decision – the majority with the corresponding "true" number of clusters.

Rule (*ii*) along with rule (*i*) make up an integral part of the validation procedure and, therefore, both of them become appropriate tools for the invasive MBDFM configuration scheme. Furthermore, for the best revealing of the "true" number of clusters, the MBDFM with invasive configuration should be performed based on the validation results produced by the non-invasive configuration. That is, the MBDFM with non-invasive configuration should always precede the MBDFM with invasive configuration. Finally, the invasive configuration aims to strengthen the non-invasive to efficiently cope with all data sets presented in the experimental majority-based validation procedure.

Using Scenario 3 cases from the non-invasive MBDFM configuration as our input data, the enhanced majority rule can be written as follows.

Rule (*ii*):

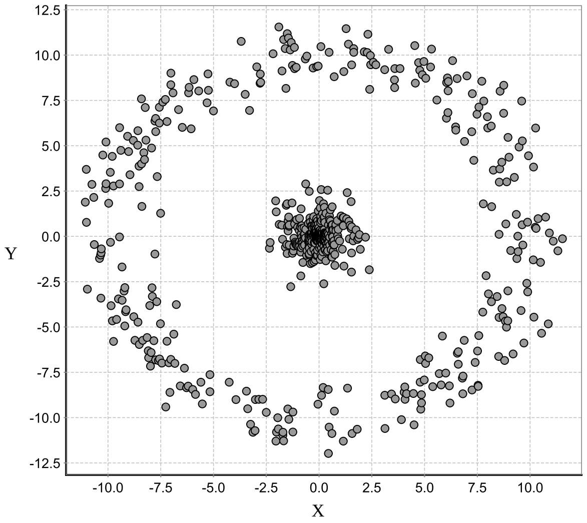
1. Run the MBDFM with a non-invasive configuration. Let DEC= $\{Dec_1, Dec_2, \ldots, Dec_k\}$ and Scenario 3 is considered.
2. Reveal the number of clusters $k_{est_j^{\wedge 1}}$ and $k_{est_j^{\wedge 2}}$ that corresponds to the two equal groups of CVIs, Dec_{est}^1 and Dec_{est}^2 respectively.
3. Optimize the cluster range by means of reducing the upper and lower bound of K when some of the decision groups Dec_i do not reveal any number of clusters.
4. Repeat step (2). If MBDFM reveals an absolute majority, assume it corresponds to the "true" number of clusters, then $k_{est_j^{\wedge}} = k_{true}$, otherwise, if no majority prevails, move to the next step. If the MBDFM reveals the relative majority $Maj_{est}^{\wedge}$ use step (5) of rule (*i*).
5. Change the cluster range, where Dec_{est}^1 and Dec_{est}^2 and its corresponding $k_{est_j^{\wedge 1}}$ and $k_{est_j^{\wedge 2}}$ becomes the upper/lower bound of K.
6. Repeat step (2). If the MBDFM reveals $k_{est_j^{\wedge}} = k_{true}$, the estimated number of clusters is the "true" one; otherwise, if the $k_{est_j^{\wedge}} \neq k_{true}$, the majority $Maj_{est}^{\wedge}$ identifies the fake "true" number of clusters. Moreover, if no majority prevails, rule (*ii*) did not give the expected results and should not be used in this MBDFM invasive configuration.

With regard to Scenario 3 situations, rule (*ii*) had a strong impact on the CVI majority voting procedure. The number of successes is considerably increased when rule (*ii*) is adopted to the MBDFM with an invasive configuration. In particular, the overall success score of 95% without rule (*ii*) is exceedingly improved to 99%.

Another remarkable and surprising fact is that 1% of experimental data sets (1 data set) show wrong results even with the application of MBDFM with the invasive configuration. This is due to a more complex data group structure presented in the data set.

Figure 4 shows the informally called the "crater" data set that misled rule (*i*) and the MBDFM with the invasive configuration in general. This synthetic "toy" data set consists of 2 clusters, one of them being a globular form dense cluster and the other being a ring cluster that surrounds the first one. This case corresponds to the Scenario 2c situations. This scenario was not included in the main list of the most frequent dubious scenarios. This case cannot be included in Scenario 2a, since the relative majority $Dec_3 = Maj_{est}^{\wedge} \neq Maj_{true}$ and the number of corresponding clusters $k_{true} \neq k_{est_3^{\wedge}} = 6$ did not point to the "true" number of clusters. Moreover, this case cannot be included in Scenario 2b, since the nearest alternative group of CVIs and the number of corresponding clusters $k_{est_2^*} = 3$ did not show the „true" one $Dec_2 = Maj_{est}^* \neq Maj_{true}$. This case shows that neither the relative majority $Maj_{est}^{\wedge}$ nor the neighboring alternative Maj_{est}^* indicated the correct "true" number of clusters $k_{true} = 2$. In this particular data set, the MBDFM with an invasive configuration did not provide the expected results, moreover, the method reveals the misleading fake number of "true" clusters.

The behavior of MBDFM becomes unpredictable for a number of reasons which are not directly related to the correctness of the proposed approach. For such cases, it is necessary to separately select a clustering algorithm and a dissimilarity measure that would be well adapted to such data. For such data sets [20] suggested using the special clustering algorithms. These types of clustering algorithms are robust to noise and the "touching problems" [72] including the "neck problem" [70] and the ‘‘adjacent problem'' [28]. Moreover, some of CVIs may not be able to cope with such data a priori.



```
************************************************************************
* Among all indices:
* 5 proposed 2 as the best number of clusters
* 6 proposed 3 as the best number of clusters
* 8 proposed 6 as the best number of clusters
* 2 proposed 7 as the best number of clusters
* 2 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters
         ***** Conclusion *****
* According to the majority rule, the best number of clusters is 6

*************************************************************************
```

Fig. 4. Two-dimensional plot of synthetic "toy" data set used in the experiment broken down by the non-invasive MBDFM. The data set shows two "true" clusters of different densities, however, according to the non-invasive MBDFM configuration the revealed number of clusters is equal to six

8. The performance of validation approaches

In this section, the authors will show the difference between the MBDFM voting approaches with default, non-invasive and invasive configurations. Moreover, the authors will compare the MBDFMs to the individual CVI (Silhouette) with *k*-means and Ward clustering with Euclidean and Canberra dissimilarity measures. The authors [21] claim that the Silhouette is the best individual CVI that achieves the best overall validation results for synthetic and real data sets broken down by the number of clusters, dimensionality, cluster overlap, and density experimental factors. Therefore, with full confidence in accordance with results presented by Arbelaitz et al. [21], the Silhouette has been chosen as the best individual CVI for clustering validation comparison reasons. Table 5 lists the overall success-failure score (%) of the MBDFMs compared to the individual CVI.

Table 5. Overall success-failure score (%) of the MBDFM with non-invasive, invasive, and default configurations compared with the result revealed by the individual Silhouette CVI

	Success score (%)		Failure score (%)		
Scenario:	1	2a	2b	2c	3
1. Silhouette with *k*-means clustering and Euclidean distance	26%	16%	41 %	10 %	7 %
2. Silhouette with *k*-means clustering and Canberra distance	21%	20%	39 %	9 %	11 %
3. Silhouette with Ward clustering and Euclidean distance	28%	22%	36 %	7 %	7 %
4. Silhouette with Ward clustering and Canberra distance	23%	27%	43 %	3 %	4 %
5. MBDFM with default configuration	20%	20%	49 %	0 %	11 %
6. MBDFM with non-invasive configuration	49%	36%	10 %	0 %	5 %
7. MBDFM with invasive configuration	99%	0%	0 %	1 %	0 %

As it can be observed, the MBDFM with default configuration cannot beat the Silhouette index for all data sets in a different configuration. The best success score has been achieved using the Silhouette CVI with the Ward clustering algorithm and Canberra or Euclidean distance 50%. The Silhouette CVI with the *k*-means clustering and Canberra distance has achieved the smallest success score – 41% (Scenario 1 – 21% + Scenario 2a – 20%). Moreover, the MBDFM with default configuration achieved an equal success score in comparison with the Silhouette CVI with *k*-means clustering and Euclidean distance – 40%. In conclusion, the Silhouette CVI with different configurations achieved a higher or equal individual success score than MBDFM with the default configuration.

Considering the MBDFM with non-invasive and invasive configurations, both of these decision fusion methods beat the overall success score of the Silhouette CVI for all data sets presented in an experimental setup. In particular, the improvement over this CVI and the best configuration was 35% for the MBDFM with the non-invasive configuration and 49% for the MBDFM with the invasive configuration and adapted rules. The analysis showed that the design of decision fusion strategies requires careful choice of the validation configurations. Finally, the MBDFM with default configuration showed no improvement in performance, whereas both voting methods MBDFM with non-invasive and invasive configurations showed to perform better than single Silhouette CVI.

9. Conclusions

The experimental results demonstrated the appealing performance of MBDFMs in searching and justifying the “true” number of clusters and thus confirmed the potential approach of integrating MBDFMs into the clustering framework. The MBDFMs and overall clustering validation schema could be iterative and researchers seek a “true” number of clusters each time. Depending on the task’s requirements or/and the level of acceptance with the validation results, the MBDFM with default configuration can be quite satisfactory. However, for detailed and sophisticated data analysis, the authors propose a more refined MBDFM with invasive configuration, where the information of all previously uncovered “true” clusters by means of MBDFM with non-invasive configuration will be used as background knowledge to derive a precise final decision. Moreover, if researchers wish to determine the pros and cons of other existing or novel CVIs, clustering algorithms, or data sets in the future, this benchmarking framework can be applied to make a thorough comparison.

In light of the results achieved, the authors consider that MBDFMs are a successful path to obtain the best partition for each context, which is the key issue in the data clustering field. Thus, the authors believe that new contributions on MBDFMs clustering validation can help to reduce the uncertainty about the suitability of the partitions generated by the algorithms. This work also raises some questions and, therefore, suggests some future work. The authors consider that, even though they performed an extensive comparison, there is still room for extending it to include more CVIs, data sets, clustering algorithms, dissimilarity measures, cluster range, high dimensionality, etc. In this context noise and overlap would appear to be the most interesting factors to analyze in greater depth. Moreover, the work is limited to binary crisp CVI’s decisions, so a fuzzy CVI’s comparison would be a natural continuation.

References

[1] Akoglu L., Tong H., Koutra D.: Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery 29(3), 2015, 626–688.

[2] Arbelaitz O., Gurrutxaga I., Muguerza J., Pérez J., Perona I.: An extensive comparative study of cluster validity indices. Pattern Recognition 46(1), 2013, 243–256.

[3] Bailey K.D.: Typologies and Taxonomies: An introduction to classification techniques (quantitative applications in the social sciences). SAGE Publications, Thousand Oaks 1994.

[4] Ball G.H., Hall D.J.: ISODATA, a Novel Method of Data Analysis and Pattern Classification. Stanford Research Institute 1965.

[5] Bandyopadhyay S., Maulik U: Nonparametric genetic clustering: comparison of validity indices. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 31(1), 2001, 120–125.

[6] Beale E.M.L.: Cluster Analysis. Scientific Control Systems, London 1969.

[7] Bezdek J., Li W., Attikiouzel Y., Windham M.: A geometric approach to cluster validity for normal mixtures. Soft Computing – A Fusion of Foundations, Methodologies and Applications 1(4), 1997, 166 –179.

[8] Bezdek J., Pal N.: Some new indexes of cluster validity. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics) 28(3), 1998, 301–315.

[9] Berkhin P.: A Survey of Clustering Data Mining Techniques. Grouping Multidimensional Data. Springer, Berlin 2006.

[10] Braune C., Besecke S., Kruse R.: Density Based Clustering: Alternatives to DBSCAN, Partitional Clustering Algorithms. Springer, Cham 2014.

[11] Brock G., Pihur V., Datta S., Datta S.: clValid: An R Package for Cluster Validation. Journal of Statistical Software 25(4), 2008, 1–22.

[12] Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E.: Model-based evaluation of clustering validation measures. Pattern Recognition 40(3), 2007, 807–824.

[13] Calinski T., Harabasz J.: A dendrite method for cluster analysis. Communications in Statistics – Theory and Methods 3(1), 1974, 1–27.

[14] Cannataro M., Congiusta A., Mastroianni C., Pugliese A., Talia D., Trunfio P.: Grid-Based Data Mining and Knowledge Discovery. Intelligent Technologies for Information Analysis. Springer, Berlin 2004.

[15] Celebi M.: Partitional clustering algorithms. Springer, Cham 2015.

[16] Charrad M., Ghazzali N., Boiteau V., Niknafs A.: NbClust: AnRPackage for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software 61(6), 2014, 1–36.

[17] Cho K., Lee J.: Grid-Based and Outlier Detection-Based Data Clustering and Classification. Communications in Computer and Information Science. Springer, Berlin 2011.

[18] Chou C., Su M., Lai E.: A new cluster validity measure and its application to image compression. Pattern Analysis and Applications 7(2), 2004, 205–220.

[19] Davies D., Bouldin D.: A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1(2), 1979, 224–227.

[20] Deng M., Liu Q., Cheng T., Shi Y.: An Adaptive Spatial Clustering Algorithm Based On Delaunay Triangulation. Computers, Environment and Urban Systems 35, 2011, 320–332.

[21] Dimitriadou E.: cclust: Convex Clustering Methods and Clustering Indexes. R package version 0.6-18, 2014.

[22] Dimitriadou E., Dolňicar S., Weingessel A.: An examination of indexes for determining the number of clusters in binary data sets. Psychometrika 67(1), 2002, 137–159.

[23] Dubes R.: How many clusters are best? – An experiment. Pattern Recognition 20(6), 1987, 645–663.

[24] Duda R., Hart P: Pattern classification and scene analysis. Wiley, New York 1973.

[25] Duda R, Hart P., Stork D.: Pattern classification. Wiley, New York 2001.

[26] Dunn J.: Well-Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybernetics 4(1), 1974, 95–104.

[27] Embrechts E., Gatti C., Linton J., Roysam B.: Hierarchical Clustering for Large Data Sets. Advances in Intelligent Signal Processing and Data Mining. Springer, Berlin 2013.

[28] Estivill-Castro V., Lee I.: Argument Free Clustering For Large Spatial Point-Data Sets Via Boundary Extraction From Delaunay Diagram. Computers, Environment and Urban Systems 26, 2002, 315–334.

[29] Fränti P., Mariescu-Istodor R., Zhong C.: XNN Graph, Lecture Notes in Computer Science, 10029, 2016, 207–217.

[30] Frey T., van Groenewoud H.: A Cluster Analysis of the D 2 Matrix of White Spruce Stands in Saskatchewan Based on the Maximum-Minimum Principle. The Journal of Ecology 60(3), 1972, 873–886.

[31] Friedman H., Rubin J.: On Some Invariant Criteria for Grouping Data. Journal of the American Statistical Association 62(320), 1967, 1159–1178.

[32] Granichin O., Volkovich Z., Toledano-Kitai D.: Cluster Validation. Intelligent Systems Reference Library. Springer, Berlin 2015.

[33] Gurrutxaga I., Muguerza J., Arbelaitz O., Pérez J., Martín J.: Towards a standard methodology to evaluate internal cluster validity indices. Pattern Recognition Letters 32(3), 2011, 505–515.

[34] Halim Z., J. Khattak J.: Density-based clustering of big probabilistic graphs. Evolving Systems 10, 2019, 333–350.

[35] Halkidi M., Batistakis Y., Vazirgiannis M.: On Clustering Validation Techniques. Journal of Intelligent Information Systems 17(2/3), 2001, 107–145.

[36] Handl J., Knowles J.: Multi-Objective Clustering and Cluster Validation. Studies in Computational Intelligence. Springer, Berlin 2006.

[37] Halkidi M., Vazirgiannis M.: A density-based cluster validity approach using multi-representatives. Pattern Recognition Letters, 29(6), 2008, 773–786.

[38] Halkidi M., Vazirgiannis M.: Clustering validity assessment: finding the optimal partitioning of a data set. Proceedings 2001 IEEE International Conference on Data Mining. IEEE, San Jose 2001.

[39] Halkidi M., Vazirgiannis M., Batistakis Y.: Quality Scheme Assessment in the Clustering Process. Lecture Notes in Computer Science. Springer, Berlin 2000.

[40] Hartigan J.A.: Clustering Algorithms. John Wiley & Sons, New York 1975.

[41] Hennig C.: Methods for merging Gaussian mixture components. Advances in Data Analysis and Classification 4, 2010, 3–34.

[42] Hornik K.: A CLUE for CLUster Ensembles. Journal of Statistical Software 14(12), 2005, 1–25.

[43] Hubert L., Levin J.: A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin 83(6), 1976, 1072–1080.

[44] Kryszczuk K., Hurley P.: Estimation of the Number of Clusters Using Multiple Clustering Validity Indices. Lecture Notes in Computer Science, Springer, Berlin 2010.

[45] Krzanowski W., Lai Y.: A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. Biometrics 44(1), 1988, 23–34.

[46] Lu J., Zhang G., Ruan D., Wu F.: Multi-objective group decision making: methods, software and applications with fuzzy set techniques. Imperial College Press, London 2007.

[47] Maalel W., Zhou K., Martin A., Elouedi Z.: Belief Hierarchical Clustering, Belief Functions: Theory and Applications. Lecture Notes in Computer Science. Springer, Cham 2014.

[48] Marriott F.: Practical Problems in a Method of Cluster Analysis. Biometrics 27(3), 1971, 501–514.

[49] McClain J., Rao V.: CLUSTISZ: A Program to Test for the Quality of Clustering of a Set of Objects. Journal of Marketing Research 12(4), 1975, 456–460.

[50] Meyer D., Dimitriadou E., Hornik K., Weingessel A., Leisch F.: E1071: Misc Functions of the Department of Statistics, Probability Theory Group. R package version 1.6-8, 2017.

[51] Milligan G.: An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45(3), 1980, 325–342.

[52] Milligan G., Cooper M.: An examination of procedures for determining the number of clusters in a data set. Psychometrika 50(2), 1985, 159–179.

[53] Nerurkar P., Pavate A., Shah M., Jacob S.: Performance of Internal Cluster Validations Measures for Evolutionary Clustering. Advances in Intelligent Systems and Computing. Springer, Singapore 2018.

[54] Nieweglowski L.: clv: Cluster Validation Techniques. R package version 0.3-2.1, 2014.

[55] Oliveira J., Pedrycz W.: Advances in fuzzy clustering and its applications. John Wiley & Sons Ltd, Chichester 2007.

[56] Peng Q., Wang Y., Ou G., Tian Y., Huang L., Pang W.: Partitioning Clustering Based on Support Vector Ranking. Lecture Notes in Computer Science. Springer, Cham 2016.

[57] Ratkowsky D.A., Lance G.N.: A Criterion for Determining the Number of Groups in a Classification. Australian Computer Journal 10(3), 1978, 115–117.

[58] Rezaei M., Fränti P.: Set Matching Measures for External Cluster Validity. IEEE Transactions on Knowledge and Data Engineering 28(8), 2016, 2173–2186.

[59] Rousseeuw P.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 1987, 53–65.

[60] Roux M.: A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms. Journal of Classification 35(2), 2018, 345–366.

[61] Sarle W.S.: Cubic Clustering Criterion, SAS Technical Report A-108. SAS Institute Inc, Cary 1983.

[62] Saemi B., Hosseinabadi A., Kardgar M., Balas V., Ebadi H.: Nature Inspired Partitioning Clustering Algorithms: A Review and Analysis. Advances in Intelligent Systems and Computing. Springer, Cham 2017.

[63] Scott A., Symons M.: Clustering Methods Based on Likelihood Ratio Criteria. Biometrics 27(2), 1971, 387–397.

[64] Shim Y., Chung J., Choi I.: A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm. International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). IEEE, Vienna 2005.

[65] Steinley D., Henson R.: OCLUS: An Analytic Method for Generating Clusters with Known Overlap. Journal of Classification 22(2), 2005, 221–250.

[66] Tan P., Steinbach M., Kumar V.: Introduction to data mining. Pearson, 2005.

[67] Vathy-Fogarassy A., Abonyi J.: Graph-Based Clustering and Data Visualization Algorithms. Springer, London 2013.

[68] Walesiak M., Dudek A.: clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package version 0.43-4, 2014.

[69] Yera A., Arbelaitz O., Jodra J., Gurrutxaga I., Pérez J., Muguerza J.: Analysis of several decision fusion strategies for clustering validation. Strategy definition, experiments and validation. Pattern Recognition Letters 85, 2017, 42–48.

[70] Zahn C.: Graph-Theoretical Methods For Detecting And Describing Gestalt Clusters. IEEE Transactions on Computers C-20, 1971, 68–86.

[71] Žalik K., Žalik B.: Validity index for clusters of different sizes and densities. Pattern Recognition Letters 32(2), 2011, 221–234.

[72] Zhong C., Miao D., Wang R.: A Graph-Theoretical Clustering Method Based On Two Rounds Of Minimum Spanning Trees. Pattern Recognition 43, 2010, 752–766.

M.Sc. Eng. Taras Panskyi
e-mail: tpanski@kis.p.lodz.pl

T. Panskyi received his M.Sc. from the Lviv Polytechnic National University, Institute of Telecommunications, Radioelectronics, and Electronic Engineering. Currently, he is a Ph.D. student at the Lodz University of Technology, Institute of Applied Computer Science. His areas of interest are data clustering, clustering validation indices, clusterability, etc. He has published more than 20 technical articles.

http://orcid.org/0000-0002-0416-8711

Prof. D.Sc. Eng. Volodymyr Mosorov
e-mail: w.mosorow@kis.p.lodz.pl

V. Mosorov received his Ph.D. in 1998 from the Lviv Polytechnic National University, Ukraine. He received his habilitation degree from AGH University of Science and Technology in Krakow Poland in 2009. Currently, he holds a position as an associate professor at the Institute of Applied Computer Science of Lodz University of Technology, Poland. His research interests include data mining, clustering, etc. He has published more than 110 technical articles.

http://orcid.org/0000-0001-6016-8671

otrzymano/received: 11.03.2021 *przyjęto do druku/accepted: 7.06.2021*