Comparing clusterings using combination of the kappa statistic and entropy-based measure

Evženie Uglickich*[1], Ivan Nagy[1,2], and Dominika Vlčková[3]

[1]*Department of Signal Processing, The Czech Academy of Sciences, Institute of Information Theory and Automation*
Pod vodárenskou věží 4, 18208 Prague, Czech Republic, suzdalev@utia.cas.cz
[2]*Faculty of Transportation Sciences, Czech Technical University*
Na Florenci 25, 11000 Prague, Czech Republic, nagy@utia.cas.cz
[3]*Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University*
Břehová 7, 11519 Prague, Czech Republic, vlckodom@fjfi.cvut.cz

Abstract

The paper focuses on a problem of comparing clusterings with the same number of clusters obtained as a result of using different clustering algorithms. It proposes a method of the evaluation of the agreement of clusterings based on the combination of the Cohen's kappa statistic and the normalized mutual information. The main contributions of the proposed approach are: (i) the reliable use in practice in the case of a small fixed number of clusters, (ii) the suitability to comparing clusterings with a higher number of clusters in contrast with the original statistics, (iii) the independence on size of the data set and shape of clusters. Results of the experimental validation of the proposed statistic using both simulations and real data sets as well as the comparison with the theoretical counterparts are demonstrated.

Keywords: comparing clusterings, clusters agreement, $\kappa_{\max}$ statistic, normalized mutual information

1 Introduction

This paper deals with a task of the evaluation of the agreement of clusters resulting from different methods of the cluster analysis. The cluster analysis is a highly demanded branch of the data mining area, known also as unsupervised learning (Larose, 2005). It provides a considerable amount of algorithms directed at sorting data with similar attributes into groups called clusters, see e.g., (Jain, 2010; Han et al., 2011; Zaki and Meira, 2014), etc. Clustering is required in many application fields of multivariate data analysis, including, for example, but not limited to bioinformatics (MacCuish and MacCuish, 2010; MacCuish and MacCuish, 2014), social fields (Maione et al., 2018; Shiau et al., 2017), transportation sciences (Suzdaleva and Nagy, 2018; Tari and Hashemi, 2018), fault detection (Li and Hu, 2018; Rodríguez-Ramos et al., 2018), big data (Hidri et al., 2018) and many others.

Traditionally, clustering approaches are distinguished between hierarchical methods (divisive or agglomerative), e.g., (Guha et al., 2000; Hastie et al., 2016), etc., and partitioning methods, which are further divided (with probable overlapping) among

*Corresponding author. Department of Signal Processing, The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod vodárenskou věží 4, 18208 Prague, Czech Republic. Tel: +420 266 052 358, email: suzdalev@utia.cas.cz

- centroid-methods such as famous k-means developed more than 50 years ago (Steinhaus, 1956), but still used in many extensions overviewed in (Jain, 2010), k-medoids (Han et al., 2011), etc.;
- density-based methods, e.g., (Ng and Han, 1994), etc.;
- grid-based methods, e.g., (Saini and Rani, 2017), etc.;
- clustering high-dimensional data and constraint-based methods (Han et al., 2011);
- model-based clustering methods, such as, e.g., conceptual clustering (Han et al., 2011), neural network approach (Han et al., 2011), mixture-based clustering (Frühwirth-Schnatter, 2006; Nagy and Suzdaleva, 2017), etc.

Recent trends in clustering methods include also ensemble clustering (Hastie et al., 2016), semi-supervised clustering (Yin et al., 2010), etc., predominantly based on extension and improvements of numerical methods above.

When using different clustering methods for a set of data, the comparison of the obtained clusters is needed for their verification. Another challenge is the validation of new proposed clustering algorithms with the help of the evaluation of the cluster agreement against results of reliable theoretical counterparts, well-known in this area (for example, k-means (Jain, 2010), fuzzy c-means (Dunn, 1973), etc.).

Methods, which compare the clustering results were discussed in a series of papers. (Wagner and Wagner, 2007) categorized these studies as follows:

- Measures based on counting pairs in clusters, such as the Chi squared coefficient, Rand index (Rand, 1971) and adjusted Rand index (Hubert and Arabie, 1985), Fowlkes-Mallows index (Fowlkes and Mallows, 1983), Mirkin metric (Dongen, 2000), Jaccard index (Tan et al., 2005) and Partition difference (Li et al., 2004). They are used to compare clusterings, however, they are seriously limited in practice by (i) the independence assumed for the clusterings (e.g., the Chi squared coefficient and Fowlkes-Mallows index), (ii) the sensitivity to the fixed cluster number and size of clusters (the general Rand index and Partition difference) or (iii) the assumption on a generalized hypergeometric distribution with a fixed number of clusters and a fixed number of elements in each cluster (the adjusted Rand index). The Mirkin metric and Jaccard index are similar to the Rand index. Moreover, the Fowlkes-Mallows index was originally proposed for the comparison of hierarchical clusterings.
- Measures based on set overlaps, such as the $\mathcal{F}$-measure (Larsen and Aone, 1999), the asymmetric Meila-Heckerman measure, which can generalized to the symmetric maximum-match-measure (Meilă and Heckerman, 1999) and the Van Dongen measure (Dongen, 2000). The assymetry of these measures represents their serious drawback in the practical use. As reported by (Wagner and Wagner, 2007), they also ignore the non-overlapping elements of clusters.
- The entropy-based measures using the normalized mutual information between two clusterings, such as proposed by (Strehl and Ghosh, 2002), (Fred and Jain, 2003) or the variation of information according to (Meilă, 2007). According to (Wagner and Wagner, 2007), this group of measures is more promising for further research as well as for using in practice, as they do not suffer from limitations mentioned for two previous categories.

Another paper belonging to the information-based category was presented by (Meilă, 2005) introducing an axiomatic view of the clustering comparison, where the clusterings were taken as elements of a lattice and a criterion of their comparison was a function of pairs of elements in the lattice. The information theoretic measures for

clustering comparison were also extensively explored by (Vinh et al., 2010), where their important metric and normalization properties were discussed and proved.

Comparing the clustering results based on the maximization of the Cohen's κ coefficient (Cohen, 1960) was proposed in the paper of (Reilly et al., 2005). This method seems to be close to those based on counting pairs, however, it does not require any assumption of the clustering independence. One of the drawbacks mentioned by the authors is its lower successfulness for the comparison of model-based clusterings with the non-hyperspherical shape of clusters.

An extensive overview of methods was given by (Vendramin et al., 2010), who divided the existing 40 clustering validity criteria between optimization-based and difference-based methods, provided their complexity analysis and comparison. They also described an alternative methodology to compare clusterings, which enables getting rid of the simplified assumptions on the correctness of one of the algorithms, but on the contrary, trying to distinguish between better and worse clusterings within an application field. Unlike this, (Meila, 2016) evaluated clusterings based on the difference from the optimal partition of a data set and also provided the overview of the methods with regard to the above categorization.

A series of application related studies has been also found. For example, the paper of (Sirsikar and Wankhede, 2015) compared clustering algorithms applied to wireless sensor networks and proposed a new clustering algorithm bringing energy saving advantages. (Schütz et al., 2018) used the sum of squared errors between the original and clustered data as a simple comparative measure, which has been finally found as not a sufficient index to compare the clustering algorithms applied to building energy systems. (Asioli et al., 2018) dealt with the comparison of clustering approaches applied to the data from coffee consumers. In one of the recent publications, (Lee et al., 2018) note that the existing measures are known to be sensitive to the arbitrary shape of clusters and propose a new method based on a support vector data description. They show that the proposed measure is less sensitive to the cluster shape and more robust to the data noise and outliers.

Despite the considerable number of publications in this area, new methods are still proposed from time to time. This is obviously motivated by limitations of the existing methods that makes the choice of a suitable measure hard in practice. In this paper, we present an attempt of combining two of undemanding measures as a new method to compare clusterings with the same number of clusters with the aim to investigate whether their combination gives some advantages in the practical use. The $\kappa_{\max}$ statistic described by (Reilly et al., 2005) and the normalized mutual information presented by (Strehl and Ghosh, 2002) are chosen for using in the proposed method, which is then experimentally compared with each of the original techniques.

The remainder of the paper is organized as follows. Section 2 formulates a problem of comparing clusterings. Section 3 provides the theoretical background of both the original methods, while Section 4 introduces the new statistic combining them. Section 5 is devoted to the experimental validation of the proposed method based on simulations and real data sets. It also provides the discussion. Conclusions are given in Section 6.

2 Problem formulation

Let's have a set of data elements, which have been clustered by a pair of clustering algorithms. A clustering obtained with one of the algorithms is denoted by $A = \{A_1, A_2, \ldots, A_n\}$, where n is a number of clusters. All of the clusters are supposed to contain at least one element. Similarly, a clustering of another algorithm is denoted by $B = \{B_1, B_2, \ldots, B_m\}$ with m non-empty clusters. Generally, m is not necessarily equal to n. However, in the case $m \neq n$ the clusterings do not obviously agree. Thus, from a practical point of view, $m = n$, i.e., the same number of clusters obtained with a pair of algorithms will be assumed in this paper. None of the clusterings is understood as "more correct" for a given data set.

The task to be solved in the paper is to use the number of elements in the contingency table

$$T_{ij} = |A_i \cap B_j|$$

expressing the intersection of two clusterings A and B for evaluating the agreement of the clusterings. The denotation $i = \{1, \ldots, n\}$ means rows and $j = \{1, \ldots, m\}$ – columns of the table. It is clear that the capacity of the set $|A_i \cap B_j|$ for $i = j$ should be as large as possible, limited by the size of a smaller cluster (if any) in the considered pair of clusters.

The agreement of clusterings can be evaluated with the help of different approaches mentioned in Section 1. This paper proposes to use the combination of two of them to present the new modified method. For better understanding, the theoretical background of both of them is briefly given below.

3 Preliminaries

3.1 The $\kappa_{\max}$ statistic

(Reilly et al., 2005) proposed to use the $\kappa_{\max}$ statistic, which is the extension of the Cohen's κ coefficient (Cohen, 1960), for comparing clusterings. The main idea is to consider the permutation of the contingency table as a mapping of one of the clustering algorithms to another one and use the permutation with the maximum Cohen's κ coefficient. Then the $\kappa_{\max}$ statistic is calculated according to the following formulas (Umesh et al., 1989; Reilly et al., 2005)

$$\kappa_{\max}(A, B) = \frac{p_{\max} - p_{\exp}}{1 - p_{\exp}}, \tag{1}$$

where

$$p_{\max} = \frac{1}{N}\sum_{i=1}^{n} T_{ii}, \;\; p_{\exp} = p_r p_c, \;\; p_r = \frac{1}{N}\sum_{i=1}^{n} T_{ij}, \;\; p_c = \frac{1}{N}\sum_{j=1}^{m} T_{ij}, \tag{2}$$

N is the number of data in the clustered data set, T_{ii} and T_{ij} belong to the permuted contingency table, p_r is a sum of the permuted contingency table along rows and p_c is a sum along columns.

Values of the statistic close to 0 indicate very weak agreement between clusterings and 1 means conversely perfect match. According to (Reilly et al., 2005), the statistic works for any number of clusters.

3.2 The normalized mutual information

(Strehl and Ghosh, 2002) introduced the mutual information between two considered clusterings normalized using the geometric mean

$$\mathcal{NMI}(A, B) = \frac{\mathcal{I}(A, B)}{\sqrt{\mathcal{H}(A)\mathcal{H}(B)}}, \tag{3}$$

where the mutual information (see, e.g., (Wagner and Wagner, 2007))

$$\mathcal{I}(A, B) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(i, j) \log_2 \frac{p(i, j)}{p(i)p(j)} \tag{4}$$

expresses the uncertainty about the location of clusters in clustering A under condition that clusters in clustering B are known, and vice versa. Here, the probability that a data element belongs to the i-th cluster of clustering A, respectively to the j-th cluster of clustering B is defined by

$$p(i) = \frac{|A_i|}{N}, \;\; p(j) = \frac{|B_j|}{N} \tag{5}$$

and then

$$p(i,j) = \frac{|A_i \cap B_j|}{N} \tag{6}$$

gives the probability that a data element belongs both to cluster A_i of clustering A and to cluster B_j of B. The denominator in (3) includes the entropy of each of the clusterings A and B as follows:

$$\mathcal{H}(A) = -\sum_{i=1}^{n} p(i)\log_2 p(i), \quad \mathcal{H}(B) = -\sum_{j=1}^{m} p(j)\log_2 p(j) \tag{7}$$

that serves as a measure of the uncertainty regarding the belongingness of a randomly taken data element to clusters (Wagner and Wagner, 2007). The normalized mutual information (3) is defined on the interval from 0 to 1 with the interpretation similar to the $\kappa_{\max}$ statistic.

4 Comparing clusterings via combination of statistics

Analyzing data with the help of different-type clustering algorithms like model-based methods and k-means (Jain, 2010), for their comparing it is useful to apply measures suitable for various shapes of clusters. The use of the $\kappa_{\max}$ statistic according to (Reilly et al., 2005) is really fast and easy in practice. In $\kappa_{\max}$, the decisive information needed for evaluating the matching is rooted in the value of $p_{\max}$ from (2), which is a function of maximum diagonal elements of the permuted contingency table. As regards the value of $p_{\exp}$ in (1), it seems that it depends on the number of clusters and it has very close values in the case of pairs of clusterings with the same number of clusters, but with the different location of data elements among the clusters.

In (Reilly et al., 2005), the suitability of $\kappa_{\max}$ for arbitrary shape and size of clusters, which is characteristic in the case of mixture based clustering is not unambiguous. Thus, to verify the $\kappa_{\max}$ values, another measure can be taken. The normalized mutual information (Strehl and Ghosh, 2002) has been chosen due its belongingness to the entropy based measures not limited by the distribution assumption or cluster overlapping.

Experimentally, specific sensitive features have been discovered in the case of both measures (this will be shown later). Calculating $\mathcal{NMI}$, starting from substantial agreement between clusterings, the capacity of the intersection of any two clusters of the compared clusterings can be zero, which causes troubles with calculating the logarithm in (4). One of specific drawbacks of the $\kappa_{\max}$ statistic was related to decreasing its values in the case of the higher number of clusters. The values were decreasing faster than the values of the $\mathcal{NMI}$ statistic, however, otherwise these statistics agree in evaluation of cluster matching.

The minor drawbacks of these easily usable statistics not requiring any assumptions regarding clusterings have motivated to try to combine them. The combined statistic denoted by Λ is presented in the following way inspired by the logistic function, which has been empirically adjusted

$$\Lambda(A,B) = \frac{\exp\{z\}}{1+\exp\{\frac{p_{\max}}{\mathcal{NMI}(A,B)}\}}, \tag{8}$$

which uses the regression

$$z = \theta\psi', \tag{9}$$

$$\text{with the regression vector} \quad \psi = [1 \quad p_{\max} \quad \mathcal{NMI}(A,B)] \tag{10}$$

$$\text{and the regression coefficients} \quad \theta = [\exp\{-\frac{0.8}{n}\} \quad \frac{0.8}{n} \quad \frac{0.3}{n}], \tag{11}$$

which have been manually tuned. A simple example of calculating all the three statistics in a programming free and open source environment Scilab (www.scilab.org) is given below.

Let's compare two clusterings of simulated data shown in Figure 1. One of them is the result of the k-means algorithm (Jain, 2010), see Figure 1 (left). Another one is the cluster-weight model (CWM) based clustering according to (Ingrassia et al., 2015; Punzo and Igrassia, 2016; Dang et al., 2017; Mazza et al., 2018), see Figure 1 (right). The data set contains 650 data elements. Both the algorithms detected 3 clusters. Visually the agreement of the clusterings seems to be substantial. The $\kappa_{\max}$, $\mathcal{NMI}$ and Λ statistics are used to evaluate their agreement.

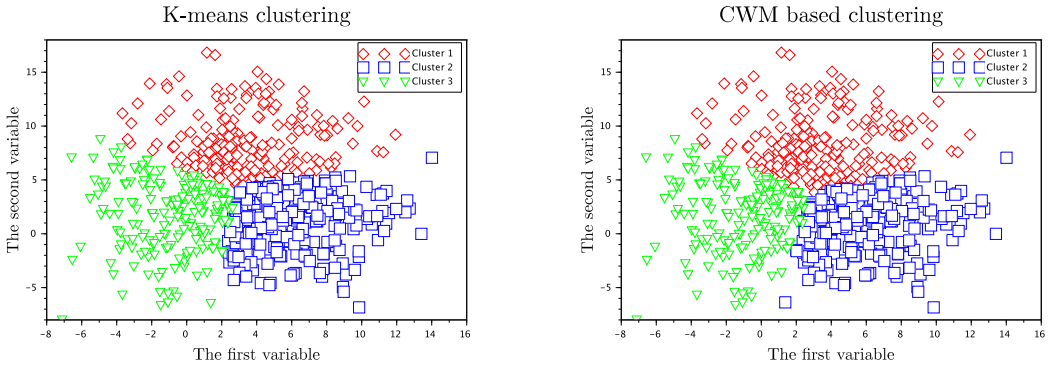

Figure 1: The k-means (left) and the model based (right) clusterings of simulated data

The contingency table of the clusterings is as follows:

$$T = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 133 & 0 & 93 \\ 2 & 1 & 86 & 138 \\ 3 & 70 & 129 & 0 \end{array}. \tag{12}$$

It should be permuted to find the maximum agreement on the main diagonal, which is here straightforward, i.e.,

$$T = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 133 & 93 & 0 \\ 2 & 1 & 138 & 86 \\ 3 & 70 & 0 & 129 \end{array}. \tag{13}$$

Using (1)-(2) and (13), the $\kappa_{\max}$ statistic is calculated as follows:

$$p_{\max} = \frac{133+138+129}{650} = 0.6153846,$$

$$p_{\exp} = [\frac{133+1+70}{650} \quad \frac{93+138}{650} \quad \frac{86+129}{650}] \cdot [\frac{133+93}{650} \quad \frac{1+138+86}{650} \quad \frac{70+129}{650}]' = 0.3334059,$$

$$\kappa_{\max}(\text{k-means, CWM}) = 0.4230141, \tag{14}$$

which evaluates the agreement of this pair of clusterings as moderate, see, e.g., https://www.statisticshowto.datasciencecentral.com/cohens-kappa-statistic.

The normalized mutual information (Strehl and Ghosh, 2002) does not require the optimal permutation of the table T. According to (3)-(7) and (12), it is obtained

$$\mathcal{H}(\text{k-means}) = -[\frac{133+93}{650} \quad \frac{1+86+138}{650} \quad \frac{70+129}{650}]$$

$$\times \log_2([\frac{133+93}{650} \quad \frac{1+86+138}{650} \quad \frac{70+129}{650}]') = 1.5825274, \tag{15}$$

$$\mathcal{H}(\text{CWM}) = -[\frac{133+1+70}{650} \quad \frac{86+129}{650} \quad \frac{93+138}{650}]$$
$$\times \log_2([\frac{133+1+70}{650} \quad \frac{86+129}{650} \quad \frac{93+138}{650}]') = 1.5830801. \tag{16}$$

The calculation of (4) can be shown here for individual clusters, e.g., for clusters 1 of both clusterings. It is

$$\mathcal{I}(\text{k-means}_1, \text{CWM}_1) = \frac{133}{650} \log_2 \frac{\frac{133}{650}}{\frac{133+93}{650}\frac{133+1+70}{650}} = 0.1855808. \tag{17}$$

It can be seen that calculating (4) via (6) for clusters 1 and 2, or 3 and 3 gives the zero in the logarithm. This can be treated by changing the zero values for small values close to zero, e.g., 0.00001. This minor numerical correction does not influence the result, but it enables calculating the logarithm. With this correction, the result is -0.0000004. After summing up all the values of (4) and using (3), the $\mathcal{NMI}$(k-means, CWM) measure is equal to 0.3863443. This is the fair agreement, which is slightly weaker than it was in the case of the $\kappa_{\max}$ statistic.

Further, calculating the Λ statistic is straightforward with the help of (8)-(11), (13), $p_{\max}$ and the $\mathcal{NMI}$ values, i.e.,

$$z = [\exp\{-\frac{0.8}{3}\} \quad \frac{0.8}{3} \quad \frac{0.3}{3}] \cdot [1 \quad 0.6153846 \quad 0.3863443]' = 0.9686653,$$
$$\text{and then} \quad \Lambda(\text{k-means, CWM}) = 0.4451777, \tag{18}$$

which again corresponds to the moderate agreement of the clusterings.

The statistics obtained slightly differ in their values. The statistical significance of the difference in the values is discussed in the subsequent section, which investigates the behavior of the statistics in order to find out whether the proposed combination keeps the properties and limitations of its origins.

5 Validation experiments

The validation experiments have been conducted in Scilab.

5.1 Experiments with a low fixed number of clusters

To investigate the behavior of the Λ statistic, a series of experiments has been conducted. Firstly, the aim of the experiments was to explore the agreement among all of the three statistics. For this end, 30 contingency tables corresponding to two clusterings with the fixed number of clusters (here equal to 3) were simulated 30 times, i.e., on the simulation space 30×30. To cover various degrees of clustering matching, the contingency tables were of the following types: (i) with a great random number of clustered elements on the main diagonal (i.e., belonging to the same clusters), which can be then evaluated as the almost maximum matching, (ii) with a dominating random number of elements in each row (i.e., belonging to one of the clusters), which allows the permutation to maximize the statistic, and (iii) the randomized contingency tables generated from the uniform distribution, where the low values of statistics can be expected. The number of the clustered data elements was random.

The obtained values of the $\kappa_{\max}$, $\mathcal{NMI}$ and Λ statistics were tested for a statistically significant difference. As the normality assumption was not satisfied, the Kruskal-Wallis test (Kruskal and Wallis, 1952) has been chosen to test whether the values of the three statistics originate from the same distribution. The p-values of the Kruskal-Wallis

tests can be found in Figure 2, where the average p-value was 0.34056, the maximum was 0.675 and the minimum 0.1346. The p-values were higher than the significance level of 0.05, which means that there is no significant difference among the values of the $\kappa_{\max}$, $\mathcal{NMI}$ and Λ statistics. It indicates that all of the statistics agree in evaluating the cluster matching for a fixed number of clusters equal to 3 and a random number of data.

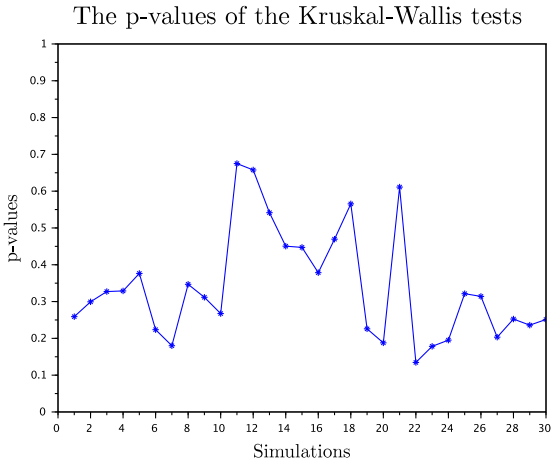

Figure 2: The p-values of the Kruskal-Wallis tests for 30x30 simulations

The expectations of the statistic values over all of the simulations are also demonstrated in Figure 3, where nevertheless it can be seen that, as a rule, $\kappa_{\max}$ has the highest values in comparison with $\mathcal{NMI}$ and Λ.

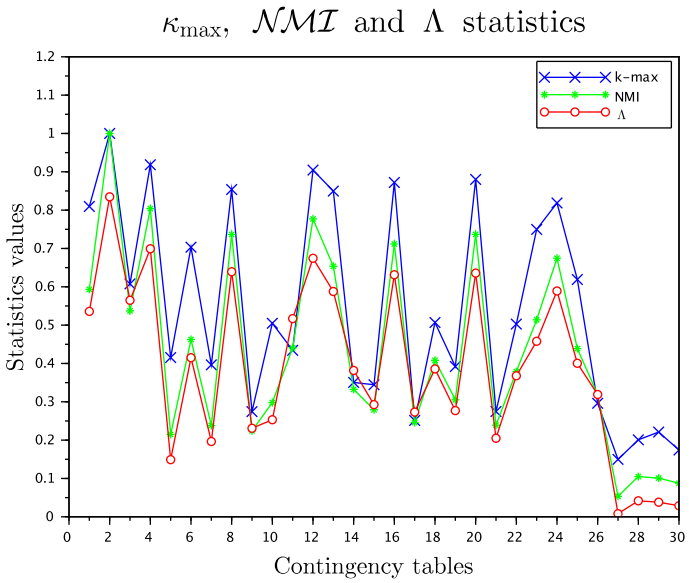

Figure 3: The mean values of the $\kappa_{\max}$, $\mathcal{NMI}$ and Λ statistics over all of the simulations of the contingency tables with 3 clusters and various agreement

5.2 Experiments with a growing number of clusters

Further, the experiments with a growing number of clusters have been conducted with the aim to turn out its influence on the statistics. In order to guarantee the same matching conditions for any number of clusters, the simulated contingency tables contained greater random numbers of elements on the main diagonal. The rest of clustered elements in the contingency tables were small random numbers. Using such simulations, the agreement of compared clusterings can be assumed from substantial (the statistic values are from 0.6 to 0.8) to near perfect (about 0.81 – 0.99). The number of clusters was growing from 3 to 15, which is the reasonable choice from a practical point of view. The number of data elements was random.

The evolution of mean values of all of the statistics calculated from 300 simulations of 13 contingency tables with the growing number of clusters from 3 to 15 is shown in Figure 4. It can be seen that on average, after 6 clusters values of the $\kappa_{\max}$ statistic started to fall sharply. The $\mathcal{NMI}$ also demonstrated the downward trend as the number of clusters was growing, however, not so abruptly. Due to the combination of both of them, the Λ statistic has remained on the assumed level of the statistic values. However, with the higher number of clusters about 13-15, the Λ statistic has approached too close to 1, which does not correspond to the assumed level of matching. It means that for such a high number of clusters none of the discussed statistics is effectively usable in this case. Probably, the average of the $\mathcal{NMI}$ and Λ will be successful in this case.

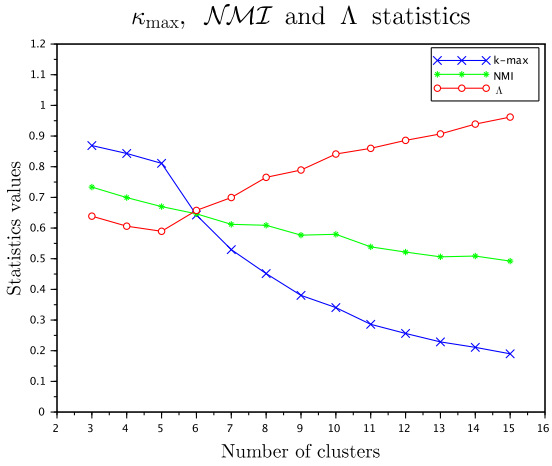

Figure 4: The mean values the $\kappa_{\max}$, $\mathcal{NMI}$ and Λ statistics with the growing number of clusters from 3 to 15 and the agreement from substantial to near perfect

The statistic values have been also obtained for individual numbers of clusters growing gradually. Their mean values over all the simulations are plotted in Figure 5. The figures show that under the same conditions of the substantial agreement, the $\kappa_{\max}$ statistic drops sharply with the number of clusters 7 and more. The $\mathcal{NMI}$ values are still acceptable for 9 clusters, but then they are also falling, although until the level of the moderate agreement. The Λ statistic is still within the assumed bounds in the case of 11 clusters, but subsequently it grows.

When using the "ideal" case with the perfect agreement (i.e., the contingency table, whose elements with the exception of the main diagonal are zeros), the following results are observed. For the number of clusters lower than 6, all the statistics start close to each other approaching the value of 1. For the higher number of clusters, $\kappa_{\max}$ drops again abruptly, $\mathcal{NMI}$ remains equal to 1 and Λ is approaching to 1 until the number of clusters is equal to 10 – 11. Subsequently, it remains just above 1. The growing number of clusters affects the evolution of the statistic values similarly in the case of slight agreement.

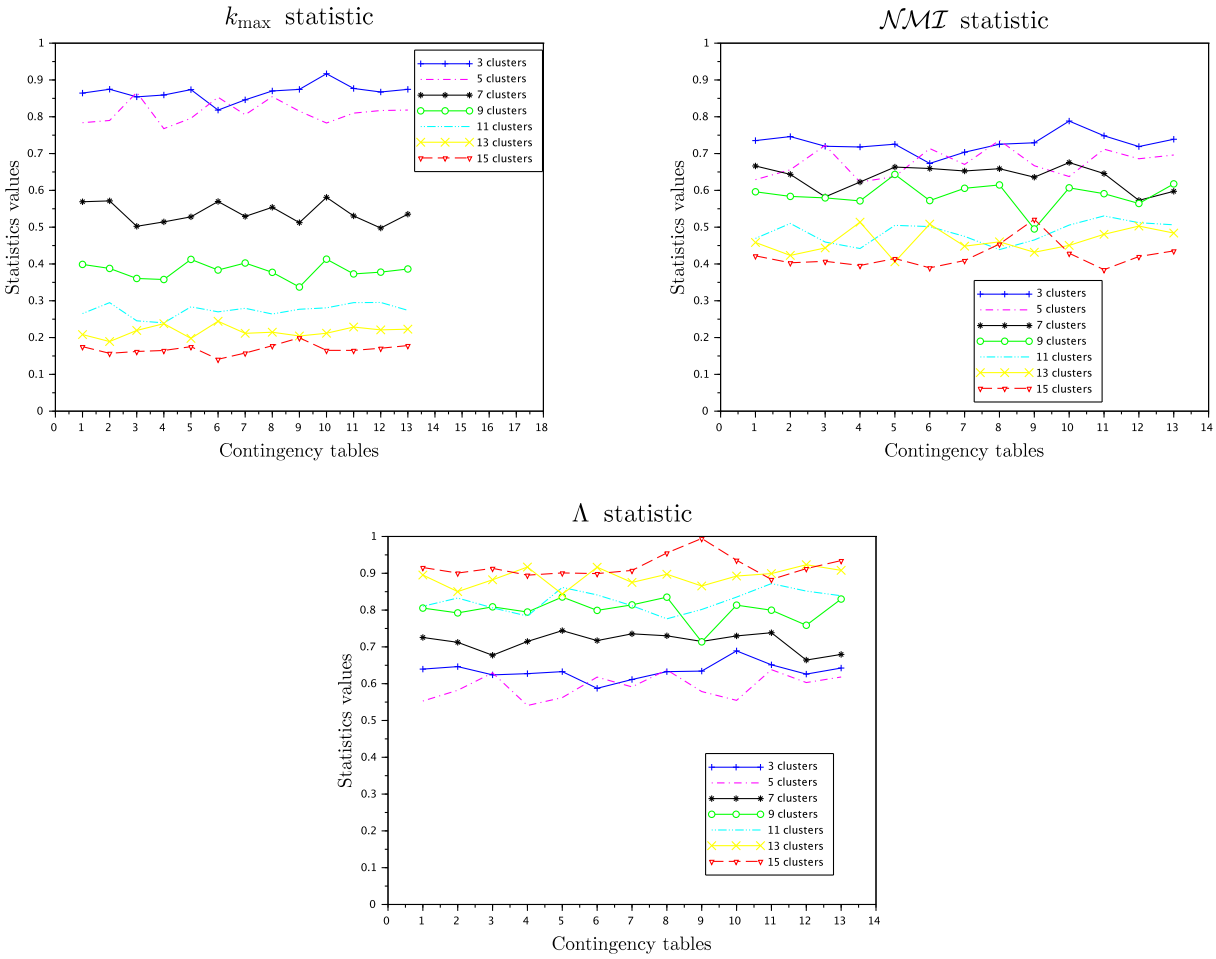

Figure 5: The values of the $\kappa_{\max}$, $\mathcal{NMI}$ and Λ statistics for the number of clusters from 3 to 15

Based on this part of experiments, it can be said that in practice not including the "ideal" matching, the proposed Λ statistic can be safely used until 11 clusters, $\mathcal{NMI}$ is suitable up to 9 clusters and $\kappa_{\max}$ should be utilized in the case of number of clusters lower than 6.

5.3 Arbitrary size of clusters

Here, the aim of the experiments was to investigate the influence of size of the data set to the statistic values. The fixed number of clusters equal to 5 has been chosen as reliable in the case of using any of the statistics. 30 random contingency tables have been simulated, which were corresponding to the low level of agreement from fair with the statistic values assumed about 0.21–0.40 to slight, where they are about 0.1– 0.20. The size of the data sets was growing in each table gradually from 100 to 3000 data elements. The simulations were repeated 100 times. No specific effect of the growing data size on the behavior of the statistics has been discovered: the results generally correspond to those obtained in Section 5.1 with the random data set size. The $\mathcal{NMI}$ and Λ show very close values and $\kappa_{\max}$ again has the highest ones, see Figure 6. For this specific case with the almost uniformly distributed contingency tables, the difference of the mean values of $\kappa_{\max}$ and the rest of statistics is statistically significant at the significance level 0.05, which has been tested in pairs by the two sample Kolmogorov-Smirnov test (Massey,

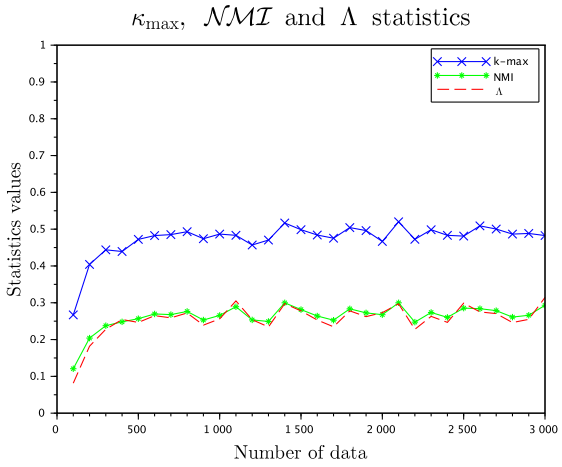

Figure 6: The mean values of the $\kappa_{\max}$, $\mathcal{NMI}$ and Λ statistics with the growing size of data sets from 100 to 3000 data elements

1951). The p-value obtained is 0.0001. The difference in mean values of $\mathcal{NMI}$ and Λ is not statistically significant with the p-value 0.32128.

The influence of an arbitrary size of individual clusters has been also tested, which gave similar results.

5.4 Arbitrary shape of clusters

The influence of shape of clusters has been tested on various real data sets. A series of experiments has been conducted with the number of clusters from 3 to 7. The statistics have agreed in all of the cases. Here, results of two of them are presented.

Hematological data

One of the tested data sets contained 81 anonymized hematological data elements, see, e.g., (Marinov et al., 2011). For two-dimensional visualization, the following variables were chosen: y_1 – precollection number of CD34+, $[\mu l]$, and y_2 – processed total blood volume (TBV), $[ml]$. The k-means (Jain, 2010) and fuzzy c-means (Dunn, 1973) algorithms have been used for clustering the data. The obtained clusterings to be compared are shown in Figure 7. The locations of three arbitrary-shaped clusters visually differ within the range from 17000 to 18000 $[ml]$ of TBV.

The contingency table in this case is

$$T = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 21 & 0 & 0 \\ 2 & 0 & 0 & 23 \\ 3 & 6 & 15 & 16 \end{array}. \tag{19}$$

For this pair of clusterings, $\kappa_{\max} = 0.6076618$, the mutual normalized information $\mathcal{NMI} = 0.5372067$ and the proposed Λ is equal to 0.5647885, which means that the agreement is evaluated as moderate by all of the statistics.

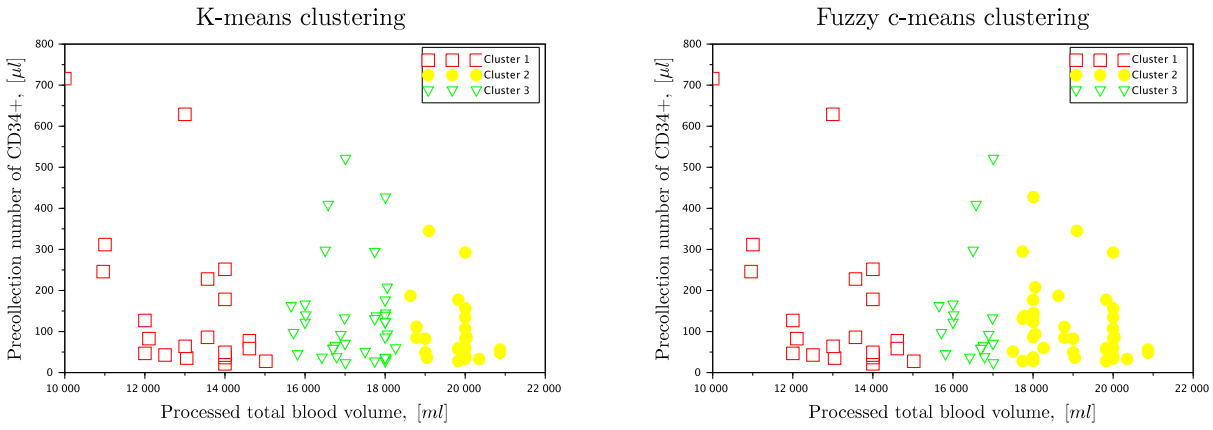

Figure 7: The k-means and fuzzy c-means clusterings to be compared

Epileptic seizure recognition data

Here, the epileptic seizure recognition data set publicly available at (Dua and Graff, 2019), see also (Andrzejak et al., 2001), has been used. As the explanatory variables, the data sets contains 178 electroencephalography (EEG) recordings within 11500 periods of time. 5 categories of the variables are available as follows: 1 – recording the seizure activity; 2 - recording the EEG signal from the brain area where the tumor was located; 3 – recording the EEG activity from the healthy brain area; 4 – recording the EEG signal when the patient had their eyes closed; 5 – recording the EEG signal when the patient had their eyes open. According to the description of the data set, it can be supposed that 5 clusters are probably to overlap substantially. Three clustering methods have been applied to the data set: two mixture-based clustering algorithms, namely, (i) the mixture model with normal components and the dynamic pointer (Nagy and Suzdaleva, 2017) and (ii) the mixture of uniform components and the dynamic pointer (Suzdaleva et al., 2017), and the k-means algorithm. It means that the agreement of six pair of clusterings (including the real realizations) can be evaluated with the help of the discussed statistics. The visual demonstration of clusters is not representative for such a high number of variables, so the contingency tables for the chosen explanatory variables will be shown. They have been obtained as follows, where $\mathcal{N}$ denotes the mixture-based clustering with normal components, $\mathcal{U}$ stands for uniform components, $\mathcal{R}$ corresponds to real categories and $\mathcal{KM}$ denotes k-means:

$T(\mathcal{N},\mathcal{U}) =$

	1	2	3	4	5
1	227	518	511	154	145
2	77	3337	1744	3	1
3	16	1642	727	0	0
4	18	1445	665	0	0
5	5	183	81	0	1

, $T(\mathcal{N},\mathcal{R}) =$

	1	2	3	4	5
1	1358	37	51	71	38
2	569	1428	1332	800	1033
3	53	678	714	270	670
4	308	130	171	1104	415
5	12	27	32	55	144

,

$T(\mathcal{N},\mathcal{KM}) =$

	1	2	3	4	5
1	18	52	650	68	767
2	82	132	2217	199	2532
3	40	64	999	109	1173
4	38	54	892	96	1048
5	1	9	117	7	136

, $T(\mathcal{U},\mathcal{R}) =$

	1	2	3	4	5
1	255	27	23	17	21
2	953	1487	1527	1572	1586
3	799	776	750	710	693
4	155	2	0	0	0
5	138	8	0	1	0

,

$$T(\mathcal{U},\mathcal{KM}) = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 2 & 10 & 154 & 16 & 161 \\ 2 & 105 & 186 & 3025 & 292 & 3517 \\ 3 & 70 & 107 & 1561 & 158 & 1832 \\ 4 & 0 & 3 & 64 & 7 & 83 \\ 5 & 2 & 5 & 71 & 6 & 63 \end{array}, \quad T(\mathcal{KM},\mathcal{R}) = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 33 & 40 & 32 & 40 & 34 \\ 2 & 70 & 60 & 56 & 60 & 65 \\ 3 & 985 & 982 & 995 & 923 & 990 \\ 4 & 104 & 100 & 95 & 97 & 83 \\ 5 & 1108 & 1118 & 1122 & 1180 & 1128 \end{array}. \tag{20}$$

The results of comparing these pairs of clusterings are given in Table 1, where it can be seen that the agreements of the clusterings compared are equivalent to chance with the exception of the mixture-based clustering with normal components compared with real categories of EEG recordings. Their agreement is evaluated as fair by all of the statistics.

Table 1: The $\kappa_{\max}$, $\mathcal{NMI}$ and Λ statistics

	$\kappa_{\max}$	$\mathcal{NMI}$	Λ
$\mathcal{N}$ vs. $\mathcal{U}$	0.0324529	0.0781617	0.0209253
$\mathcal{N}$ vs. $\mathcal{R}$	0.266087	0.183074	0.2403329
$\mathcal{N}$ vs. $\mathcal{KM}$	-0.0006052	0.0005609	0
$\mathcal{U}$ vs. $\mathcal{R}$	0.0345652	0.0578645	0.0468144
$\mathcal{U}$ vs. $\mathcal{KM}$	-0.0010054	0.0007411	0
$\mathcal{KM}$ vs. $\mathcal{R}$	0.0091304	0.0004312	0

The rest of experiments in this section have been conducted with the set of driving-related variables measured on a driven vehicle, which have been clustered by the already mentioned clustering algorithms giving different shapes of clusters. The values of the statistics obtained as the result of the comparison have been tested for a statistically significant difference by the Kruskal-Wallis test. The p-value obtained was 0.9099, which means that at the significance level of 0.05, there is no significant difference among the values of the $\kappa_{\max}$, $\mathcal{NMI}$ and Λ statistics. It means that the non-hyperspherical shape of clusters, which is characteristic for mixtures of uniform components, had no effect on the behavior of the statistics.

5.5 Discussion

The main aim of the presented study was to combine two existing methods of comparing clusterings and explore the behavior of the new combined statistic in order to understand whether it brings any improvements in practice.

As it was demonstrated in Sections 4 – 5.4, the aim has been successfully accomplished. The $\kappa_{\max}$ statistic and the mutual normalized information have been combined with the help of the empirically adjusted function based on the logistic one, which is mapping the values of both the statistics into the interval from zero to one. The behavior of the new method along with both original statistics has been extensively investigated under various conditions using the considerable number of experiments and compared among each other. The values of the proposed statistic agreed with both the original ones during the experimental verification, which has validated the new method.

As a result, the specific advantages brought by the proposed Λ statistic include: (i) the reliable use in practice in the case of low fixed number of clusters, (ii) the suitability to comparing clusterings with a higher number of clusters up to 11 in contrast with the $\kappa_{\max}$ statistic, whose values sharply decrease with the number of clusters greater than 6 and the $\mathcal{NMI}$, which also shows slight dropping values after 9 clusters, (iii) the independence on the cluster size and shape.

However, in the case of the number of clusters about 12 and higher, the proposed Λ statistic is not suitable, as its values increase higher than the assumed level of matching, which can be verified working with simulated contingency tables. The unsuitable behavior has been also shown by the $\kappa_{\max}$ statistic and the mutual normalized information. As it is mentioned in Section 5.2, the mean value of Λ and $\mathcal{NMI}$ seems to be suitable in this case.

Summarizing the experimental part of the work, it can be said that the proposed Λ statistic can be applied to comparing results of clustering algorithms including mixture-based clustering with different type of components independently on the number of data elements to be clustered and with arbitrary shape of clusters, whose number does not exceed 11 clusters.

The limitation of the approach is also the value 1 of $\mathcal{NMI}$ in the case of the perfect agreement. This issue is treated as follows. The perfect agreement means that the elements of the main diagonal of the contingency table correspond to individual clusters, which are matching, and the rest of the table are zero elements. However, to avoid the problem with calculating the logarithm in (4) for $\mathcal{NMI}$, the numerical correction of the contingency table is necessary in the form of replacing zero values by minor numbers, e.g., 0.000001. The result of comparing is not affected by this correction. However, beside avoiding the zero in the logarithm, it guarantees that the $\mathcal{NMI}$ measure in the case of perfect agreement will get the maximum value of 0.99999. This also helps fix the limitation in calculating Λ.

6 Conclusion

The paper proposes the new method for comparing clusterings with the same number of clusters based on the combination of the existing statistics such as $\kappa_{\max}$ and the mutual normalized information. The proposed method has been experimentally validated using simulations and real data sets. The agreement among all of the statistics has been statistically tested for various number of clusters, capacity of data sets and shape of clusters.

The open problem remains comparing clusterings with a high number of clusters, which needs another extensive study aimed at this specific problem. However, it should be noticed that in practice the number of clusters from 3 to 12 (also used in the application part of (Reilly et al., 2005)) is rather acceptable.

Acknowledgements

This work has been supported by the project SILENSE, project number ECSEL 737487 and MSMT 8A17006.

References

Andrzejak, R.G., Lehnertz, K., Rieke, C., Mormann, F., David, P., Elger, C.E., 2001. Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. Phys. Rev. E, 64, 061907.

Asioli, D., Berget, I., Næs, T., 2018. Comparison of different clustering methods for investigating individual differences using choice experiments. Food Research International, 111, 371-378.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 196037-196046.

Dang, U. J., Punzo, A., McNicholas, P. D., Ingrassia, S. and Browne, R. P., 2017. Multivariate response and parsimony for Gaussian cluster-weighted models. Journal of Classification, 34(1), 4-34.

van Dongen, S., 2000. Performance criteria for graph clustering and Markov cluster experiments. Technical Report INS-R0012, Centrum voor Wiskunde en Informatica.

Dua, D. and Graff, C., 2019. UCI Machine Learning Repository http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Dunn, J.C. 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Cybernetics and Systems, 3, 32-57.

Fowlkes, E. B., Mallows, C. L., 1983. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, 78, 553-569.

Fred, A., Jain, A., 2003. Robust data clustering. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 3, 128-136.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models, 2nd edition, Springer New York, 2006.

Guha, S., Rastogi, R., Shim, K., 2000. Rock: A robust clustering algorithm for categorical attributes. Inform. Systems. 25 (5), 345-366.

Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann.

Hastie, T., Tibshirani, R., Friedman, J., 2016. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) 2nd Ed., Springer.

Hidri, M. S., Zoghlami, M.A., Ayed, R.B., 2018. Speeding up the large-scale consensus fuzzy clustering for handling Big Data, Fuzzy Sets and Systems, 348, 50-74.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification, 2, 193-218.

Ingrassia S., Punzo A., Vittadini G. and Minotti S. C., 2015. The generalized linear mixed cluster-weighted model. Journal of Classification, 32(1): 85-113.

Jain, A. K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 31(8), 651-666.

Kruskal, W.H. and Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583-621.

Larose, D. T., 2005. Discovering Knowledge in Data. An Introduction to Data Mining, Willey.

Larsen, B. and Aone, C., 1999. Fast and effective text mining using linear time document clustering. In: Proceedings of the KDD, 16-29.

Lee, S.-H., Jeong, Y.-S., Kim, J-Y, Jeong, M.K., 2018. A new clustering clusters validity index for arbitrary shape of clusters. Pattern Recognition Letters, 112, 263-269.

Li, G. and Hu, Y., 2018. Improved sensor fault detection, diagnosis and estimation for screw chillers using density-based clustering and principal component analysis. Energy and Buildings, 173, 502-515.

Li, T., Ogihara, M., Ma, S., 2004. On combining multiple clusterings. In: Proceedings of the ACMConference on Information and Knowledge Management, 13, 294-303.

Maione, C., Nelson, D. R., Barbosa, R.M., 2018. Research on social data by means of cluster analysis. Applied Computing and Informatics, 2018, In press, https://doi.org/10.1016/j.aci.2018.02.003.

MacCuish, J.D. and MacCuish, N.E., 2010. Clustering in Bioinformatics and Drug Discovery. CRC Press.

MacCuish, J.D. and MacCuish, N.E., 2014. Chemoinformatics applications of cluster analysis. Wiley Interdisciplinary Reviews: Computational Molecular Science, 4(1), 34-48.

Marinov, I., Luxová, A., Tkáčová, V., Gašová, Z., Pohlreich, D., Cetkovský, P. (2011). Comparison of three single platform methods for CD34+ hematopoietic stem cell enumeration by flow cytometry. Clinical laboratory. 57(11-12): 1031-5.

Massey, F.J.Jr., 1951. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 46(253), 68-78.

Mazza A., Punzo A. and Ingrassia S., 2018. flexCWM: a flexible framework for cluster-weighted models. Journal of Statistical Software, 86(2), 1-30.

Meilă, M., 2005. Comparing clusterings: an axiomatic view. In: ICML'05 Proceedings of the 22nd international conference on Machine learning, 577-584.

Meilă, M., 2007. Comparing clusterings – an information based distance. Journal of Multivariate Analysis, 98, 873-895.

Meila, M., 2016. Criteria for comparing clusterings. In: Handbook of Cluster Analysis, Hennig, C. (Ed.), Meila, M. (Ed.), Murtagh, F. (Ed.), Rocci, R. (Ed.), New York: Chapman and Hall/CRC, 619-637.

Meilă, M., Heckerman, D., 1999. An experimental comparison of model-based clustering methods. In: Proceedings of the Conference on Knowledge Discovery and Data Mining, 16-22.

Nagy, I. and Suzdaleva, E., 2017. Algorithms and Programs of Dynamic Mixture Estimation. Unified Approach to Different Types of Components. SpringerBriefs in Statistics. Springer International Publishing, 2017.

Ng, R. and Han, J., 1994. Efficient and effective clustering method for spatial data mining. In: Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), Santiago, Chile, 144-155.

Punzo, A. and Igrassia, S., 2016. Clustering bivariate mixed-type data via the cluster-weight model. Computational Statistics, 31(3), 989-1013.

Rand, W., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.

Reilly, C., Wang, Ch., Rutherford, M., 2005. A rapid method for the comparison of cluster analyses. Statistica Sinica 15, 19-33.

Rodríguez-Ramos, A., da Silva Neto, A.J., Llanes-Santiago, O., 2018. An approach to fault diagnosis with online detection of novel faults using fuzzy clustering tools, Expert Systems with Applications, 113, 200-212.

Saini, S., and Rani, P., 2017. A survey on STING and CLIQUE grid based clustering methods. International Journal of Advanced Research in Computer Science, 8(5), 1510-1512.

Shiau, W.-L., Dwivedi, Y. K., Yang, H. S., 2017. Co-citation and cluster analyses of extant literature on social networks. International Journal of Information Management, 37(5), 390-399.

Sirsikar, S. and Wankhede, K., 2015. Comparison of clustering algorithms to design new clustering approach. Procedia Computer Science, 49, 147-154.

Schütz, T., Schraven, M. H., Fuchs, M., Remmen, P., Müller, D., 2018. Comparison of clustering algorithms for the selection of typical demand days for energy system synthesis, Renewable Energy, 129(A), 570-582.

Steinhaus, H., 1956. Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci. IV (C1.III), 801-804.

Strehl, A. and Ghosh, J., 2002. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3, 583-617.

Suzdaleva, E., Nagy, I., Pecherková, P., Likhonina, R., 2017. Initialization of recursive mixture-based clustering with uniform components. In: Proceedings of the 14th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2017), p. 449-458.

Suzdaleva, E. and Nagy, I., 2018. An online estimation of driving style using data-dependent pointer model. Transportation Research Part C. 86C, 23-36.

Tan, P.-N., Steinbach, M., Kumar, V., 2005. Introduction to Data Mining, Pearson.

Tari, F. G. and Hashemi, Z., 2018. Prioritized K-mean clustering hybrid GA for discounted fixed charge transportation problems. Computers & Industrial Engineering, 126, 63-74.

Umesh, U. N., Peterson, R.A., Sauber M. H., 1989. Interjudge agreement and the maximum value of kappa. Educational and Psychological Measurement. 49, 835-850.

Vendramin, L., Campello, R. J. G. B., Hruschka, E. R., 2010. Relative clustering validity criteria: a comparative overview. In: Statistical Analysis and Data Mining, 3(4), 209-235.

Vinh, N. X., Epps, J., Bailey, J., 2010. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. Journal of Machine Learning Research, 11, 2837-2854.

Wagner, S. and Wagner, D., 2007. Comparing Clusterings - An Overview. Technical Report 2006-04.

Yin, X., Chen, S., Hu, E., Zhang., D., 2010. Semi-supervised clustering with metric learning: an adaptive kernel method. Pattern Recognition, 43(4), 1320-1333.

Zaki, M.J., Meira, Jr. W., 2014 Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.