Performance and Convergence Analysis of Modified C-Means Using Jeffreys-Divergence for Clustering

Ayan Seal[1,2]*, Aditya Karlekar[3], Ondrej Krejcar[2,4], Enrique Herrera-Viedma[5,6]

[1] PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, 482005 (India)
[2] Center for Basic and Applied Research, Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, Hradec Kralove, 50003 (Czech Republic)
[3] Hitkarini College of Engineering and Technology, Jabalpur, 482005 (India)
[4] Malaysia Japan International Institute of Technology, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur (Malaysia)
[5] Department Computer Science and Artificial Intelligence, University of Granada, 18071 Granada (Spain)
[6] Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah, 21589 (Saudi Arabia)

Received 27 September 2020 | Accepted 20 March 2021 | Published 29 April 2021

LA UNIVERSIDAD
EN INTERNET

Abstract

The size of data that we generate every day across the globe is undoubtedly astonishing due to the growth of the Internet of Things. So, it is a common practice to unravel important hidden facts and understand the massive data using clustering techniques. However, non- linear relations, which are essentially unexplored when compared to linear correlations, are more widespread within data that is high throughput. Often, non-linear links can model a large amount of data in a more precise fashion and highlight critical trends and patterns. Moreover, selecting an appropriate measure of similarity is a well-known issue since many years when it comes to data clustering. In this work, a non-Euclidean similarity measure is proposed, which relies on non-linear Jeffreys-divergence (JS). We subsequently develop c- means using the proposed JS (J-c-means). The various properties of the JS and J-c-means are discussed. All the analyses were carried out on a few real-life and synthetic databases. The obtained outcomes show that J-c-means outperforms some cutting-edge c-means algorithms empirically.

Keywords

C-mean, Clustering, Convergence, Jeffreys-Divergence, Jeffreys-Similarity Measure.

DOI: 10.9781/ijimai.2021.04.009

I. Introduction

Machine learning considers clustering to be an important issue. It is normally used to reveal some existing hereditary structure by analyzing a set of data items or patterns. The aim of clustering is to split data into groups so that data in the same groups are similar and data items in different groups are not capable of comparison in the same sense. Clustering is the subject of active research for varying areas, including marketing [1], biology [2], libraries [3], insurance [4], city planning [5], and earthquake studies [6]. Common clustering algorithms include Gaussian Mixture models [7], hierarchical clustering [8], Hidden Markov models [9], self-organizing maps [10], and c- means clustering [11]. Hierarchical clustering constructs a multi- level hierarchy of groups by making a tree, which is known as a cluster tree. Gaussian mixture model forms groups, which would be considered as a mixture of multivariate normal density components. The self-organizing map takes the help of neural networks for learning the topology and data structure in the form of distribution. Hidden Markov models use observed data for recovering the sequence of states.

The performance of a clustering algorithm always relies upon data items or their features, choice of the initial cluster centers, similarity measures, objective function, and clustering algorithms [12]–[14]. In this study, the *c*-means algorithm is implemented on synthetic and real-life databases, so everything is similar except the similarity measure. In other words, the use of different similarity measures is studied because the selection of proper similarity measures is an important issue in clustering and it helps to find the cluster structure in data [15] properly. However, Euclidean distance is one of the widely accepted similarity measures even though a large number of researches are going on around the world to introduce non-linearity in similarity measures for data clustering [15], [16]. In recent times, Euclidean distance in *c*-means is replaced using different non-linear metrics. From this, some do not obey triangle inequality property [17]–[21]. The objective of instigating non-linearity is to detect a more accurate boundary between two clusters. A. Banerjee et al. initiated general Bregman divergence as a distance metric in the *c*-means to augment its effectiveness [17]. This method in reality unified the divergence measures, for which the first moment was used as cluster

* Corresponding author.

E-mail address: ayanseal30@ieee.org

representative ensuring a gradual depreciating of the objective function in the iterative relocation technique. The interested reader can go through [12], [22]–[27] to know the use of various divergence-based similarity measures in clustering.

II. Clustering

This section presents the formal definition of clustering. A concise overview of conventional *c*-means is also discussed, given that the performance comparison is made between the conventional *c*-means and the modified one.

A. Basic Principle

The method of dividing *n*´dimensional *m* data-points or their features, $A[= a_1, a_2, ..., a_n]$, in R^n into 'c' groups of homogeneous data-points, $G[= (G_1, G_2, ..., G_c)]$ to increase association strength within the cluster, is known as clustering. However, association strength will be low or weak between different clusters. Then

$$G_i \neq \phi \quad for \quad i = 1, \ldots, c,$$
$$G_i \cap G_j = \phi \quad for \quad i = 1, \ldots, c; j = 1, \ldots, c \quad and \ i \neq j,$$
$$\cup_{i=1}^{c} G_i = G$$

B. The C-Means Algorithm

It is certainly a well-known clustering technique because it is easy to implement. Sometimes, it is applied in the pre-processing step to finding the knowledge by analyzing data [28]. It partitions data into 'c' distinct groups by reducing the entire intra-cluster variance, beginning with an arbitrarily selected group of the centroid from each group. Each centroid should effectively denote the central location of a group. The ideal value of 'c' leads to the highest separation (distance) and is an unknown priori. It has to be approximated from the database itself. The c-means intends to reduce total intra-cluster variance, or, the squared error function, *E*, which could be computed using Eq. (1).

$$E = \sum_{j=1}^{m} \sum_{i=1}^{c} | \mathrm{a}_j - \mathrm{g}_i|^2 \tag{1}$$

where $|a_j - g_i|^2$ is a similarity measure between the cluster center, g_i and a data-object, a_j.

The *c*-means algorithm consists of the given steps:

Step 1: Select 'c' initial cluster centers $g_1, g_2, ..., g_c$ arbitrarily from the m data-points $A[= a_1, a_2, ..., a_n]$.

Step 2: Designate data-point a_j , *j*=1, 2, ..., m to cluster center g_i, $i \in 1, 2, ..., c$ iff $||a_j - g_i|| \leq ||a_j - g_k||$, $k = 1, 2, ..., c$, $\&i \neq c$. Ties are broken randomly.

Step 3: Find new cluster centers $\mathrm{g}_1^+, \mathrm{g}_2^+, \ldots, \mathrm{g}_c^+$, by Eq. (2).

$$\mathrm{g}_i^+ = \frac{1}{m_i} \sum_{\mathrm{a}_j \in G_i} \mathrm{a}_j \, , i = 1,2, \ldots, c \tag{2}$$

where m_i is the count of data-objects in cluster G_i.

Step 4: If $\mathrm{g}_i^+ = \mathrm{g}_i \forall i = 1, 2, ..., c$ then stop. If not, go to Step 2.

Note that if Step 4 does not terminate then the algorithm executes for a predetermined fixed number of epochs.

This work focuses to introduce JS, which is inherited from the concept of Jeffreys-divergence [29]. Several characteristics of this similarity measure are studied. The entire experiment set is executed on some synthetic and real-life benchmark databases. These simulation outcomes show that c-means utilizing JS performs better than a traditional c-means algorithm and along with *c*-means with various other divergences in certain situations. Our assertion is confirmed through a statistical analysis of the results obtained.

III. Jeffreys-Similarity Measure (JS) and its Properties

The definition of JS and its properties are discussed in this section.

Definition 3.1. Let J_n be a set of all positive definite matrices of size $n \times n$ and Jeffreys-divergence is a similarity measure defined over J_n, which could be computed by Eq. (3).

$$\partial_{Jeffreys}(P, Q) = (P - Q)(\log(|P|) - log(|Q|)) \tag{3}$$

where $|P|$=determinant of *P*.

Consider a real positive vector $\mathrm{a} = (a_1, a_2, ..., a_n) \in \mathbb{R}_+^n$. Let us define a one-to-one function $\psi: \mathbb{R}_+^n \longrightarrow J_n$ such that $\psi(\mathrm{a}) = \mathrm{diag}(a_1, a_2, ..., a_n)$. The definition of JS is as follows:

Definition 3.2. The JS function $d_{Jeffreys}: \mathbb{R}_+^n \times \mathbb{R}_+^n \longrightarrow \mathbb{R}_+ \cup \{0\}$ between any two $a, b \in \mathbb{R}_+^n$ is defined by applying Eq. (4).

$$d_{Jeffreys}(\mathrm{a,b}) = \partial_{Jeffreys}(\psi(\mathrm{a}), \psi(\mathrm{b})) \tag{4}$$

The JS measure, $d_{Jeffreys}$, is well-stated because ψ is a one-to-one function by definition. Some of the following properties are stated here as $\partial_{Jeffreys}$ divergence is defined on J_n.

Proposition 3.1. $d_{Jeffreys}(a, b) = d_{Jeffreys}(b, a)$

Proof: $d_{Jeffreys}(a, b) = \partial_{Jeffreys}(\psi(a), \psi(b)) = \partial_{Jeffreys}(\psi(b), \psi(a)) = d_{Jeffreys}(b, a)$

Proposition 3.2.

$d_{Jeffreys}(a, b) \geq 0$ and $d_{Jeffreys}(a, b) = 0$ iff $a = b$

Proof: $d_{Jeffreys}(a, b) = \partial_{Jeffreys}(\psi(a), \psi(b)) \geq 0$ and $d_{Jeffreys}(a, b) = 0$ iff $\partial_{Jeffreys}(\psi(a), \psi(b)) = 0$ iff $\psi(a) = \psi(b)$ iff $a = b$

So, $d_{Jeffreys}$ is a similarity measure on $\mathbb{R}_+^n$, which could be thought as $d_{Jeffreys}(a, b) = \sum_{i=1}^{n} \partial_{Jeffreys}(a_i, b_i)$. Now, its time to investigate some of the properties of JS.

Theorem 3.1. The JS is not a Bregman divergence.

Proof: If JS was a Bregman divergence $d_{Jeffreys}(a, b)$ would have been strictly convex in *a*. However, our objective is to prove that $d_{Jeffreys}(a, b)$ is not convex in *a*. We know that the JS, $d_{Jeffreys}$, could also be expressed by Eq. (5).

$$d_{Jeffreys}(\mathrm{a,b}) = \sum_{i=1}^{n} (a_i - b_i)(\log(a_i) - \log(b_i)) \tag{5}$$

The expression below can be acquired if the derivative of both sides of Eq. (5) is taken with respect to a_i. $\frac{\partial d_{Jeffreys}}{da_i} = 1 - \frac{b_i}{a_i} + \log(a_i) - \log(b_i)$ $\frac{\partial^2 d_{Jeffreys}}{da_i da_j} = 0$ when $i \neq j$ otherwise,

$$\frac{\partial^2 d_{Jeffreys}}{\partial a_i^2} = \frac{b_i}{a_i^2} + \frac{1}{a_i}$$

We have, $\frac{\partial^2 d_{Jeffreys}}{\partial a_i^2} < 0$ for the values in the range of $\{-\infty, -1\} \cup \{0, 1\}$. So, $d_{Jeffreys}(a, b)$ is not convex in *a*. So, it is demonstrated that JS measure is not a Bregman divergence.

Theorem 3.2. $d_{Jeffreys}(x \circ a, x \circ a) = x d_{Jeffreys}(a, b)$ for $x \in \mathbb{R}_+^n$, where $x \circ a$ depicts the Hadamord product between *a* and *x*.

Proof: It is known that $(x \circ a) = (x_1 a_1, x_2 a_2, ..., x_n a_n)$. So,

$\delta_{Jeffreys}(x_i a_i, x_i b_i) = (x_i a_i - x_i b_i)(\log(x_i a_i) - \log(x_i b_i)) = x_i (a_i - b_i)$
$(\log x_i + \log a_i - \log x_i - \log b_i) = x_i (a_i - b_i)(\log a_i - \log b_i)$
$\sum_{i=1}^{n} \delta_{Jeffreys}(x_i a_i, x_i b_i) = \sum_{i=1}^{n} x_i \, \delta(a_i, b_i)$ implying
$d_{Jeffreys}(x \circ a, x \circ b) = x \, d_{Jeffreys}(a, b)$

Theorem 3.3. JS is f-divergence.

Proof: If a divergence expression can be made through the following

$$\phi(t) = a\,\phi(\tfrac{\mathrm{b}}{\mathrm{a}}), \ where \ t = \tfrac{\mathrm{b}}{\mathrm{a}}$$

then that divergence is known as f-divergence. The JS between $a \in \mathbb{R}_+^n$ and $b \in \mathbb{R}_+^n$ is given by

$d_{Jeffreys}$(a,q) $=\sum_{i=1}^{n}(a_i - b_i)(\log(a_i) - \log(b_i))$

putting $t_i = \frac{b_i}{a_i}$

$d_{Jeffreys}$(a,b) $=\sum_{i=1}^{n}(a_i - b_i t_i)(\log(a_i) - \log(a_i t_i)$

$=\sum_{i=1}^{n} a_i\,(1 - t_i)(\log(a_i) - \log(a_i) - \log(t_i))$

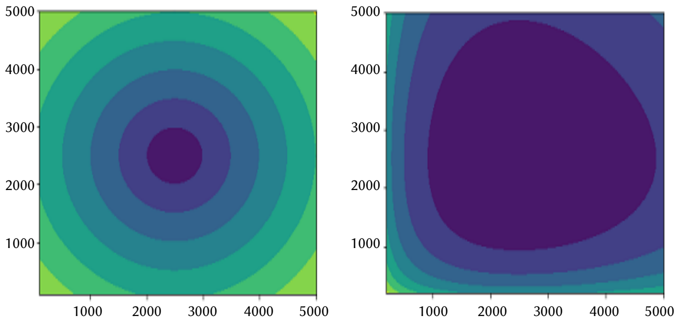
$=\sum_{i=1}^{n} a_i\,(1 - t_i)(-\log(t_i))$

$=\sum_{i=1}^{n} x_i\,(1 - t_i)(\log(\frac{1}{t_i}))$

$\sum_{i=1}^{n}\phi\,(t) =\sum_{i=1}^{n}x_i\,\phi(\frac{b_i}{a_i})$

Since, $d_{Jeffreys}$ $(a,\ b)$ can be expressed as $\sum_{i=1}^{n}x_i\,\phi(\frac{b_i}{a_i})$. Thus, JS is f-divergence.

Remark 3.1: We may consider another imperative facet of JS. Fig. 1 portrays the contour plot of the norm-balls in $\mathbb{R}^2$ everywhere over the point (5000,5000) for Euclidean distance (Fig. 1a) and JS (Fig. 1b). We can also observe from Fig. 1 that the norm-ball of Euclidean distance is similar to concentric circles, on the other hand, JS is similar to some extent to askew ovals. It is further evident from Fig. 1b that contour lines confine together as they come near the origin i.e. (0,0). Thus, we conclude that the J-divergence between two points is greater when they come in the vicinity of the origin and it reduces when their distance from the origin increases. While on the contrary, the Euclidean distance within two points remains constant regardless of their location. For instance, the J-divergence and the Euclidean distance between (3,3) and (5,5) are 2.043 and 2.82 respectively and for points (1003,1003) and (1005,1005) they are 0.0079 and 2.82 respectively. At times, the attribute in question might prove beneficial in situations where the clusters have varying sizes and densities.



a) Contour plot of Euclidean distance

b) Contour plot of JS

Fig. 1. Contour plot of norm ball for the Euclidean distance and JS.

IV. Proposed Method

A. The C-Means with JS

Consider a given set of vector, $A = \{a_1, a_2, ..., a_m\}$, in $\mathbb{R}^n_+$. Our objective is to divide A into 'c' disjoint groups where, the value of 'c' could be any value between 2 and 'm'. This problem can be formalized using the following form.

M : *minimize* $\psi\,(W,\ G) = \sum_{j=1}^{m}\sum_{i=1}^{c} w_{ij}\ d_{Jeffrey}\ (a_j,\ g_i)$, subject to the constraints

$$\sum_{i=1}^{c} w_{ij} = 1 \tag{6a}$$

$$w_{ij} \in \{1,0\}\forall j \in \{1,\dots,m\}, \forall i \in \{1,\dots,c\} \tag{6b}$$

$$G = \{g_1, g_2, \dots, g_c\}, g_i \in \mathbb{R}^n_+, \forall i \in \{1,\dots,c\} \tag{6c}$$

Two following heuristics steps are given in order to solve M.

Initialization:

The 'c' number of vectors have to pick randomly from A and called them as cluster centers, which are denoted as

$$G^{(0)} = \{g_1^{(0)}, g_2^{(0)}, \dots, g_c^{(0)}\}$$

Iterative Steps:

- Set $W^{(z+1)} = argmin_W\,\psi\,(W, G^{(z)})$ subject to constraints 6a and 6b are satisfied. In other words, each vector a_i is assigned to its nearest cluster center.
- Set $G^{(z+1)} = argmin_G\,\psi\,(W^{(z+1)}, G)$ subject to constraint 6c is satisfied.
- Set $z = z + 1$ until convergence.

Criterion for stopping:

We cease iteration in cases where the cost function reduces experiences alteration i.e.

$\psi\,(W^{(z+1)}, G^{(z)}) = \psi\,(W^{(z)}, G^{(z)})$ or $\psi\,(W^{(z+1)}, G^{(z+1)}) = \psi\,(W^{(z+1)}, G^{(z)})$. An informal program code of J-c-means is given in algorithm 1.

```
Algorithm 1 J-c-means([A] m × n, c)
Input: a set of vector, $A = \{a_1, a_2, ..., a_m\}$, $a_i \in \mathbb{R}^n$.
Output: a partition, $M = \{A_1, A_2, ..., A_c\}$, of A together with the
  centroids $g_1, g_2, ..., g_c$ of each cluster.
Initialization: select $g_1, g_2, ..., g_c$ in A at random
while terminating condition has not been met do
    for $i = 1$ to $c$ do
        $A_i \leftarrow 0$
    end for
    for $j = 1$ to $m$ do //updating the class membership of the vectors
        $\omega\,(a_j) \leftarrow argmin_{i\in\{1,2,\dots,c\}}\ d_{Jeffreys}\,(a_j, g_i)$
        $A_{\omega(a_j)} \leftarrow A_{\omega(a_j)} \cup \{a_j\}$
    end for
    for $i = 1$ to $c$ do //updating centroids
        $m_i \leftarrow \sum_{j=1}^{n} 1\,(a_j \in A_i)$
        $g_i \leftarrow \frac{1}{m_i}\sum_{j=1}^{n} a_j\ 1\,(a_j \in A_i)$
    end for
    return $M, g_1, g_2, ..., g_c$
end while
```

B. Convergence of J-C-Means Algorithm

Theorem 3.1. The J-c-means monotonically decreases the inertia $\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{c} w_{ij}\ d_{Jeffreys}\,(a_j,\ g_i)$

Proof: Let $\phi\,(A^u) = \frac{1}{m}\sum_{i=1}^{c}\sum_{j=1}^{m} d_{Jeffreys}\,(a_j,\ g_i)$, where $A^{(u)}$ is the recent group $A_1^{(u)}$, ..., $A_c^{(u)}$ with the centre of the clusters $g_1^{(u)}$, ..., $g_c^{(u)}$and assignation function $\omega^{(u)}$, then $\phi(A^u) \geq \sum_{i=1}^{c}\sum_{a_j \in A_i^{(u)}} d_{Jeffreys}(a_j, g^{(u)}_{\omega^{(u+1)}(a_j)})$ because $\omega(a_j)$ minimizes the quantity $d_{Jeffreys}\,(a_j, g_i)$ over all $i \in \{1, ..., c\}$.

$\phi(A^u) \geq \sum_{i=1}^{c}\sum_{a_j \in A_i^{(u)}} d_{Jeffreys}\ (a_j,\ g_i^{(u+1)})$ because $g_i^{(u+1)}$ minimizes the quantity $d_{Jeffreys}\,(a_j,\ g_i)$ over all $a_j \in A_i$.

Therefore, $\phi\,(A^{(u)}) \geq (A^{(u+1)})$.

Corollary: The J-c-means stops after a finite amount of time.

There are only finite number of partitions $\binom{m}{c}$. Thus, the sequence $\phi\,(A^{(u)})_{u \in N}$ has a finite number of values i.e. there exist u such that $\phi\,(A^{(u+1)}) = \phi\,(A^{(u)})$.

Remark 4.1: The above corollary does not say anything about how fast the J-c-means converges. There is an exponential bound $\binom{m}{c}$. The time required for the above mentioned algorithm to converge depends on the initialization. However, some heuristic can be found in the literature.

V. Experiments

A. Database Description

All the experiments are performed on some synthetic databases: 2_blobs, 3_blobs, 5_blobs, and 10_blobs and real-word databases: Iris, Glass, Cleveland, Bank Note Authentication, Appendicitis, Breast Cancer Wisconsin, and Mammography. These real-world databases are acquired from the Keel Repository [30] and UCI Machine Learning Repository [31].

B. Cluster Validity Index

The fundamental question that requires to be responded to in clustering is: how good a clustering technique is. The concept of goodness is quantified by validity indexes. The notion of these indexes may be explained mathematically. We may consider c-partitions namely, $A_1, A_2, ..., A_c$ of A, found by a clustering technique and the valuations of their respective validity indexes are $Z_1, Z_2, ..., Z_c$. The $Z_{h1} \geq Z_{h2} \geq ... \geq Z_{hc}$ will represent that $A_{h1} \uparrow A_{h2} \uparrow ... \uparrow A_{hc}$, for a particular permutation $h1, h2, ..., hc$ of {1, 2, ..., c}, where $A_i \uparrow A_j$ depicts that partition A_i is a better clustering than A_j [32]. Validity indexes can be categorized into two sets namely, internal validity index and external validity index. Two external validation indexes namely, Normalized Mutual Information (NMI) [33] and Adjusted Rand Index (ARI) [34] are considered in this work to measure the performance of the c-means algorithms by varying distance metrics. NMI will typically be utilized as an index that can compare the performance of two data-point groups. Meanwhile, ARI is seen as an index for cluster validation. Both metrics show the mismatch in terms of two data clustering of an allotted arrangement of data points. The highest value (1) and the lowest value (0) indicate no mismatch and complete mismatch respectively. Both metrics use the ground truth to compute the efficiency of a clustering algorithm. Three internal evaluation schemes, for example, the Silhouette index (SI) [35], Dunn index (DI) [32], and Davies Boulden Index (DBI) [32] are further employed in this research to explore the cohesiveness of the obtained clusters. These indexes estimate the similarity between a data point with the corresponding group called cohesion and disunion between different groups known as separation. The domain of SI lies within −1 and +1, in which a greater value illustrates that the data point is excellently suited with its corresponding cluster and weakly paired to neighboring clusters. A higher DI and lower DBI demonstrate a more favorable grouping.

C. Computational Protocols

Five sets of experiments were performed on the aforementioned databases through c-means-E: c-means with Euclidean distance [36], c-means-S: c-means with S-distance [37], c-means-W: Weighting in c-means [38], c-means-M: Minkowski weighted c-means [33], and c-means-P: the proposed c-means. Performance comparison: We consider the same arbitrarily selected centroids for all the algorithms while calculating ARI, NMI, SI, DI, and DBI values to make results consistent. The performance of a clustering algorithm does not rely on the better extraction of inceptive set centroids. Nevertheless, it relies upon the clustering technique. The exact methodology is administered tenfold on each database. Then Wilcoxon's rank-sum is executed to determine whether two dependent data-points from populations have the exact distribution on the acquired values of ARI, NMI, SI, DI, and DBI using the above-mentioned methods.

VI. Results and Discussion

Fig. 2 shows the clustering results. Table I shows the mean ARI, NMI, SI, DI, and DBI values obtained by the methods presented in section V-C on synthetic and real-life databases. However, the first two i.e. ARI and NMI are external clustering validity indexes for which actual class labels are required to match with the predicted class labels. database 2_blobs consists of two clusters having the same density and same size. However, one is close to the origin and the other is away from the origin. It is evident from Table I that the suggested c-means-P on 2_blobs defeats other algorithms mentioned in section V-C because nearly all of the ARI and NMI values are close to the greatest value i.e. 1. Moreover, c-means-P returns a higher expected value of ARI and NMI values over other algorithms, which depicts the efficiency of c-means-P. The proposed c-means-P outperforms due to askew oval figures of contour norm-balls of the J-divergence as considered in Remark 3.1. The proposed method also works well for the databases 3_blobs to 5_blobs, which contain clusters having the same size and same density. However, some noise is introduced to them. Still, the performance of the proposed method is good as J-divergence is invariant to the Hadamard product. The performance of all the methods on some real-life databases is noted in Table I. These outcomes depict that the proposed method c-means-P is the best among all the methods discussed in this study. The values of three internal clustering evaluation indexes namely, SI, DI, and DBI for the same databases are included in Table I. Although, actual class labels are not required in this case. The received results further validate the efficiency of the c-means-P over other methods discussed in section V-C due to the values obtained by c-means-P approach nearer to ideal values in comparison to values generated by methods other than the proposed one. The non-parametric Wilcoxon's rank-sum is also performed for comparing c-means-P over other methods presented in section V-C using the p-values achieved from ARI, NMI, SI, DI, and DBI. Table II reports the estimated p-values. We can very well observe that the generated outcomes advice that we discard the null hypothesis for a 5% level of significance. It may be proposed that substantial proof is presented using data available with us to comment that c-means-P algorithm surpasses other methods discussed in section V-C.

VII. Conclusion

In this work, a similarity measure on $\mathbb{R}^n_+$ is presented based on Jeffreys-divergence. Different JS properties are also elaborated. The conventional c-means algorithm is altered, where Euclidean distance is substituted with the similarity measure introduced. A theoretical evaluation of the JS and c-means was also conducted by outlining the convergence proof. Research on complexity metrics promises to be an area of research with potential when it comes to field clustering. It should be explored in future work. We focused on the evaluation of multiple database properties to find information. This can be used to design proper clustering algorithms. JS can be used for the Fuzzy c-means type algorithm.

Acknowledgment

This work is partially supported by the project "Prediction of diseases through computer assisted diagnosis system using images captured by minimally-invasive and non-invasive modalities", Computer Science and Engineering, PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur India (under ID: SPARC-MHRD-231). This work is also partially supported by the project Grant Agency of Excellence, Univer-sity of Hradec Kralove, Faculty of Informatics and Management, Czech Republic (under ID: UHK-FIM-GE-2204-2021).

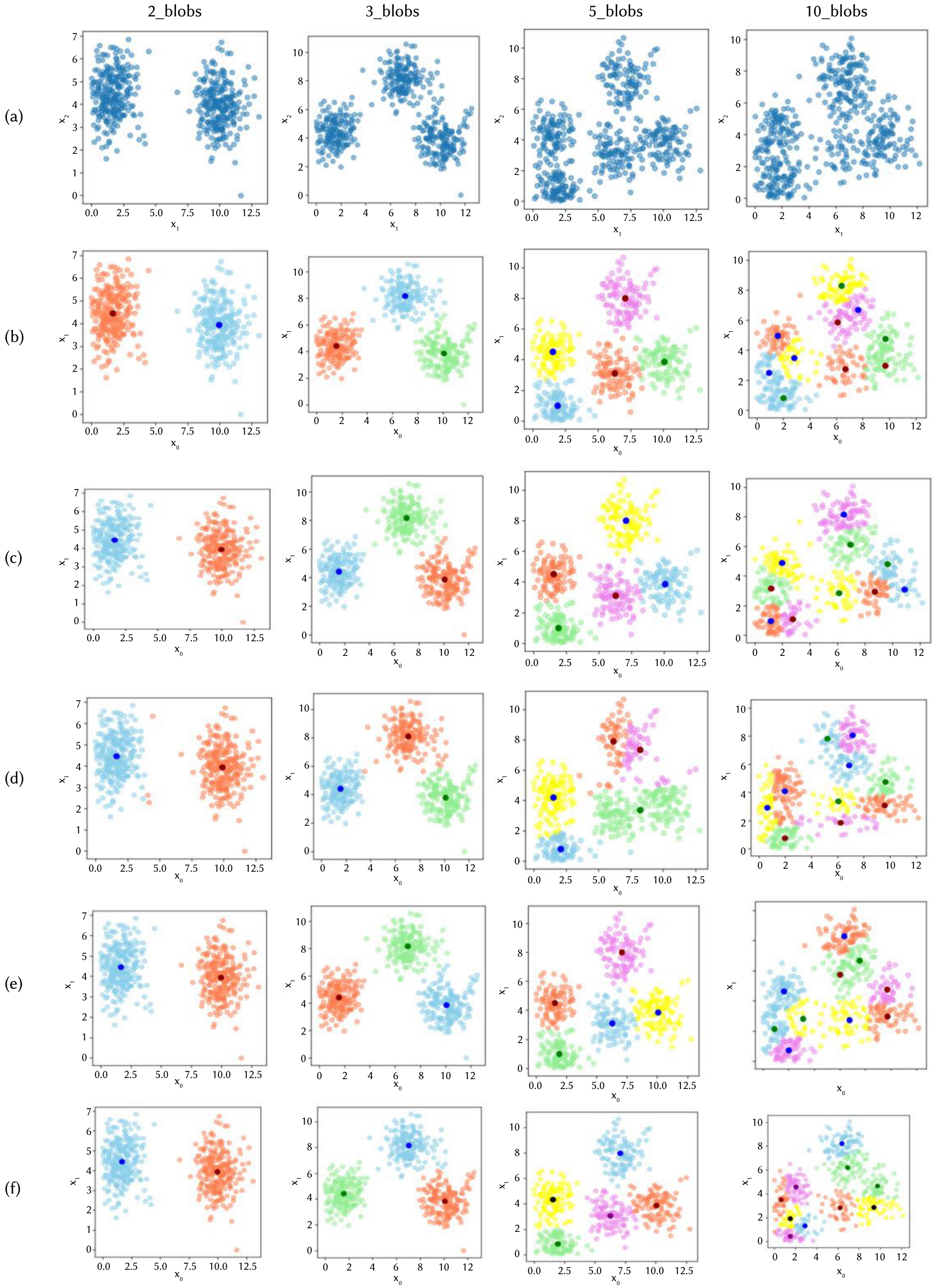


Fig. 2. (a): Original structure of #_blobs. Result of clustering corresponding #_blobs with (b): *c*-means-E, (c): *c*-means-M (d): *c*-means-S (e): *c*-means-W and (f): *c*-means-P.

TABLE I.: The Values of ARI, NMI, SI, DI and DBI for Synthetic and Real-life Databases

	Database	*c*-means-E	*c*-means-S	*c*-means-W	*c*-means-M	*c*-means-P
ARI	2_blobs	**1.0000000**	0.9760961	**1.0000000**	**1.0000000**	**1.0000000**
	3_blobs	**0.9820360**	0.9582248	0.9760904	0.9701813	**0.9820360**
	5_blobs	0.8986230	0.6279446	0.6663589	0.8941665	**0.9028355**
	10_blobs	0.430297	0.3859427	0.4502888	0.4185428	**0.4686642**
	Iris	**1.0000000**	**1.0000000**	**1.0000000**	0.9600667	**1.0000000**
	Glass	0.3008098	0.6361220	0.5476595	0.6276935	**0.7149283**
	Cleveland	0.0162569	**0.3369418**	0.0465415	0.0505246	0.1416707
	Bank Note Authentication	0.0485381	0.0404252	0.0485381	0.0488371	**0.0491855**
	Appendicitis	0.4843330	0.4320631	0.4843330	0.4978654	**0.5360417**
	Breast Cancer Wisconsin	0.4914245	0.5286179	0.4914245	0.4914245	**0.5666393**
	Mammography	0.0905026	0.0930185	0.0905026	0.0821258	**0.1275994**
NMI	2_blobs	**1.0000000**	0.9530566	**1.0000000**	**1.0000000**	**1.0000000**
	3_blobs	0.9541820	0.9391844	0.9362472	0.9503597	**0.9666142**
	5_blobs	**0.8883229**	0.7297619	0.7692242	0.8801728	**0.8883229**
	10_blobs	0.6232038	0.5858008	0.6035275	0.6132328	**0.6336725**
	Iris	**1.0000000**	**1.0000000**	**1.0000000**	0.9404430	**1.0000000**
	Glass	0.5075728	0.6577571	0.6441501	0.6832619	**0.7325871**
	Cleveland	0.0183458	0.1175260	0.0375054	0.0472017	**0.3864150**
	Bank Note Authentication	0.0303241	0.0245671	0.0303241	0.0312593	**0.0327895**
	Appendicitis	0.3999936	0.3690075	0.3809936	0.4048908	**0.4401108**
	Breast Cancer Wisconsin	0.4671655	0.4863613	0.4671655	0.4671655	**0.5163683**
	Mammography	0.0846832	0.0846832	0.0846832	0.0846832	**0.1298267**
SI	2_blobs	**0.7949160**	0.7874309	**0.7949160**	**0.7949160**	**0.7949160**
	3_blobs	**0.6932280**	0.6861834	**0.6932280**	0.6916559	**0.6932280**
	5_blobs	0.5711057	0.4126689	0.4446950	0.5759132	**0.5759364**
	10_blobs	0.3645875	0.2902406	0.3306109	0.3527441	**0.3857648**
	Iris	**0.5824192**	**0.5824192**	**0.5824192**	0.5818419	**0.5824192**
	Glass	0.2909336	0.1899170	0.3491109	0.2386990	**0.3928576**
	Cleveland	0.2076061	-0.026441	0.2657142	0.2390776	**0.2808949**
	Bank Note Authentication	0.4308310	0.4293403	0.4308310	0.4310046	**0.4310995**
	Appendicitis	0.4137615	0.4127611	0.4136630	0.4086627	**0.4137615**
	Breast Cancer Wisconsin	**0.6972643**	0.6741518	0.6910678	**0.6972643**	**0.6972643**
	Mammography	0.1243098	**0.5419065**	**0.5419065**	**0.5419065**	**0.5419065**
DI	2_blobs	**1.9040153**	1.3735754	1.3735754	**1.9040153**	**1.9040153**
	3_blobs	1.7047981	1.6202096	1.7047981	1.7047981	**1.7663088**
	5_blobs	**1.2592171**	0.6447496	0.6709407	**1.2592171**	**1.2592171**
	10_blobs	0.9048197	0.4620280	0.8099586	0.8140677	**1.2805434**
	Iris	**2.0197395**	**2.0197395**	**2.0197395**	1.9596349	**2.0197395**
	Glass	0.4010836	0.2986482	0.4644115	0.5325171	**0.6286726**
	Cleveland	0.5363801	0.5889773	0.5371508	0.5005224	**0.6011315**
	Bank Note Authentication	**1.5469099**	1.5013920	**1.5469099**	**1.5469099**	**1.5469099**
	Appendicitis	**1.0011285**	**1.0011285**	**1.0011285**	1.0017089	**1.0011285**
	Breast Cancer Wisconsin	**1.3494101**	1.1806848	**1.3494101**	1.3005589	**1.3494101**
	Mammography	**1.3974134**	**1.3974134**	**1.3974134**	1.1162343	**1.3974134**
DBI	2_blobs	**0.144604**	**0.144604**	**0.144604**	0.147063	**0.144604**
	3_blobs	0.157400	0.159603	**0.1565898**	**0.1565898**	**0.1565898**
	5_blobs	0.348582	**0.122678**	0.2365076	0.123528	0.1236736
	10_blobs	0.1068881	0.121510	0.162423	0.109133	**0.10423103**
	Iris	**0.167358**	0.16801707	**0.167358**	0.167373	**0.167358**
	Glass	0.532517	0.398066	0.46441156	0.271434	**0.2093515**
	Cleveland	0.383664	2.071026	0.33105801	0.4016002	**0.320763**
	Bank Note Authentication	0.4371350	0.436876	0.43713506	0.439077	**0.436666**
	Appendicitis	**0.516156**	0.5261876	**0.516156**	0.516380	**0.516156**
	Breast Cancer Wisconsin	0.268049	0.257680	**0.2522018**	**0.2522018**	**0.2522018**
	Mammography	**0.311799**	0.860389	0.34720557	**0.311799**	**0.311799**

TABLE II. P-Values Generated from ARI, NMI, SI, DI and DBI for Wilcoxon's Rank-sum Test for Comparing J-C-Means With other Algorithms

	Database	*c*-means-E	*c*-means-S	*c*-means-W	*c*-means-M
ARI	2_blobs	0.0010	1.5938E-06	0.0010	0.0010
	3_blobs	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
	5_blobs	5.3477E-06	6.1582E-06	4.0402E-06	4.7314E-06
	10_blobs	0.0099	0.0059	0.0099	0.0485
	Iris	0.0128	0.0088	4.0402E-06	4.0167E-04
	Glass	0.0046	0.01038	0.0017	0.0046
	Cleveland	1.4851E-04	1.8267E-04	0.0022	0.0211
	Bank Note Authentication	0.02547	6.0243E-06	4.5506E-06	0.01485
	Appendicitis	0.0325	6.0243E-06	0.0165	0.03681
	Breast Cancer Wisconsin	3.2899E-06	3.2899E-06	4.7314E-06	3.2899E-06
	Mammography	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
NMI	2_blobs	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
	3_blobs	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
	5_blobs	5.3477E-06	6.1582E-06	4.0402E-06	4.7314E-06
	10_blobs	0.04698	1.8165E-04	0.04097	0.04272
	Iris	0.0146	0.03812	5.7206E-06	4.0167E-04
	Glass	0.0013	0.0036	0.0013	0.0013
	Cleveland	1.4851E-04	1.8267E-04	0.0017	0.0058
	Bank Note Authentication	1.5938E-06	6.0243E-06	1.5938E-06	0.04339
	Appendicitis	0.0325	6.0243E-06	0.0125	0.0125
	Breast Cancer Wisconsin	1.5938E-06	1.5938E-06	1.5938E-06	2.4282E-06
	Mammography	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
SI	2_blobs	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
	3_blobs	0.00586	1.5938E-06	0.0332	0.0039
	5_blobs	6.1582E-06	6.1582E-06	0.1058	4.7314E-06
	10_blobs	0.01855	1.8165E-04	0.0211	0.0211
	Iris	0.0474	0.008812	0.0131	0.0469
	Glass	0.0451	1.8165E-04	0.0451	0.0451
	Cleveland	1.4851E-04	1.8267E-04	0.0204	0.04725
	Bank Note Authentication	2.4282E-06	0.04429	1.5938E-06	4.7682E-06
	Appendicitis	1.5938E-06	0.0010	0.03681	0.0165
	Breast Cancer Wisconsin	0.0010	2.1650E-06	2.1650E-06	2.1650E-06
	Mammography	1.5938E-06	1.5938E-06	1.5938E-06	1.5938E-06
DI	2_blobs	0.0010	0.0010	0.0010	0.0010
	3_blobs	0.0215	0.0215	0.0215	0.0215
	5_blobs	0.0014	0.045	0.0089	0.0078
	10_blobs	0.0339	0.0339	0.0339	0.07539
	Iris	0.02891	0.02891	0.02891	0.02891
	Glass	0.0339	0.0339	0.0339	0.0339
	Cleveland	0.03438	0.03438	0.03438	0.03438
	Bank Note Authentication	0.0010	0.0010	0.0010	0.0010
	Appendicitis	0.02547	0.0075	0.0125	0.0056
	Breast Cancer Wisconsin	0.0020	0.0020	0.0020	0.0020
	Mammography	0.0020	0.0020	0.0020	0.0020
DBI	2_blobs	1.5938E-05	1.5938E-06	1.5938E-06	1.5938E-06
	3_blobs	0.0486	1.5938E-06	0.0332	0.0039
	5_blobs	5.3477E-06	6.1582E-06	4.0402E-06	4.7314E-06
	10_blobs	0.0450	5.8006E-06	1.8267E-06	5.7729E-06
	Iris	0.0015	0.0321	9.6624E-06	2.5597E-06
	Glass	0.04722	1.8165E-06	0.01523	0.04772
	Cleveland	1.4851E-06	1.8267E-06	0.03845	0.0199
	Bank Note Authentication	2.4282E-06	0.04429	1.5938E-06	4.7682E-06
	Appendicitis	1.5938E-06	0.0014	0.03681	0.0013
	Breast Cancer Wisconsin	3.2899E-06	3.2899E-06	3.2899E-06	3.2899E-06
	Mammography	3.2899E-06	3.2899E-06	1.5938E-06	1.5938E-06

References

[1] H. Wattimanela, U. Pasaribu, S. Indratno, A. Puspito, "Eartquakes clustering based on the magnitude and the depths in molluca province," in *AIP Conference Proceedings*, vol. 1692, 2015, p. 020021, AIP Publishing.

[2] J. Yang, J. Cao, R. He, L. Zhang, "A unified clustering approach for identifying functional zones in suburban and urban areas," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WK- SHPS)*, April 2018, pp. 94–99.

[3] T. Pitchayaviwat, "A study on clustering customer sugges tion on online social media about insurance services by using text mining techniques," in *2016 Management and Innovation Technology International Conference (MITicon)*, Oct 2016, pp. MIT–148–MIT–151.

[4] R. Suresh, I. Anand, B. Vianesh, H. R. Mohammed, "Study of clustering algorithms for library management system," in *2018 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)*, March 2018, pp. 221–224.

[5] A. Naik, D. Reddy, P. K. Jana, "A novel clustering algorithm for biological data," in *2011 Second International Conference on Emerging Applications of Information Technology*, Feb 2011, pp. 249–252.

[6] K. C. Gull, A. B. Angadi, C. G. Seema, S. G. Kanakaraddi, "A clustering technique to rise up the marketing tactics by looking out the key users taking facebook as a case study," in *2014 IEEE International Advance Computing Conference (IACC)*, Feb 2014, pp. 579–585.

[7] J. Li, A. Nehorai, "Gaussian mixture learning via adaptive hierarchical clustering," *Signal Processing*, vol. 150, pp. 116–121, 2018.

[8] R. Abe, S. Miyamoto, Y. Endo, Y. Hamasuna, "Hierarchical clustering algorithms with automatic estimation of the number of clusters," in *17th World Congress of International Fuzzy Systems Association*, 2017.

[9] S. Ghassempour, F. Girosi, A. Maeder, "Clustering multivariate time series using hidden markov models," *International Journal of Environmental Research and Public Health*, vol. 11, pp. 2741–2763, 2014.

[10] M. Pacella, A. Grieco, M. Blaco, "On the use of self-organizing map for text clustering in engineering change process analysis: A case study," *Computational Intelligence and Neuroscience*, p. 11, 2016.

[11] V. Schellekens, L. Jacques, "Quantized compressive k-means," *IEEE Signal Processing Letters*, vol. 25, no. 8, 2018.

[12] K. K. Sharma, A. Seal, "Spectral embedded generalized mean based k-nearest neighbors clustering with s-distance," *Expert Systems with Applications*, p. 114326, 2020.

[13] K. K. Sharma, A. Seal, "Outlier-robust multi-view clustering for uncertain data," *Knowledge-Based Systems*, vol. 211, p. 106567, 2021.

[14] K. K. Sharma, A. Seal, "Multi-view spectral clustering for uncertain objects," *Information Sciences*, vol. 547, pp. 723–745, 2021.

[15] L. Bottou, Y. Bengio, "Convergence properties of the k-means algorithms," in *Advances in neural information processing systems*, 1995, pp. 585–592.

[16] A. Karlekar, A. Seal, O. Krejcar, C. Gonzalo-Martin, "Fuzzy k-means using non-linear s-distance," *IEEE Access*, vol. 7, pp. 55121–55131, 2019.

[17] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, "Clustering with bregman divergences," *Journal of machine learning research*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[18] S. Chakraborty, S. Das, "k- means clustering with a new divergence-based distance metric: Convergence and performance analysis," *Pattern Recognition Letters*, vol. 100, pp. 67–73, 2017.

[19] L. Legrand, E. Grivel, "Jeffrey's divergence between moving-average models that are real or complex, noise-free or disturbed by additive white noises," *Signal Processing*, vol. 131, pp. 350–363, 2017.

[20] K. K. Sharma, A. Seal, "Modeling uncertain data using monte carlo integration method for clustering," *Expert systems with applications*, vol. 137, pp. 100–116, 2019.

[21] A. Seal, A. Karlekar, O. Krejcar, C. Gonzalo-Martin, "Fuzzy c- means clustering using jeffreys-divergence based similarity measure," *Applied Soft Computing*, vol. 88, p. 106016, 2020.

[22] F. Nielsen, R. Nock, "Total jensen divergences: Definition, properties and clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 2016–2020.

[23] F. Nielsen, R. Nock, S. I. Amari, "On clustering histograms with k-means by using mixed -divergences," *Entropy*, vol. 16, 2014.

[24] R. Nock, F. Nielsen, S.-I. Amari, "On conformal divergences and their population minimizers," *IEEE Transactions on Information Theory*, vol. 62, 2016.

[25] M. D. Gupta, S. Srinivasa, J. Madhukara, M. Antony, "Kl divergence based agglomerative clustering for automated vitiligo grading," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2700–2709.

[26] A. Notsu, O. Komori, S. Eguchi, "Spontaneous clustering via minimum gamma-divergence," *Neural Computation*, vol. 26, 2014.

[27] K. K. Sharma, A. Seal, "Clustering analysis using an adaptive fused distance," *Engineering Applications of Artificial Intelli- gence*, vol. 96, p. 103928, 2020.

[28] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, 2010.

[29] L. Legrand, E. Grivel, "Jeffrey's divergence between moving-average models that are real or complex, noise-free or disturbed by additive white noises," *Signal Processing*, vol. 131, 2017.

[30] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, 2011.

[31] D. Dheeru, E. Karra Taniskidou, "Uci machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[32] U. Maulik, S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, 2002.

[33] N. X. Vinh, J. Epps, J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, no. Oct, 2010.

[34] L. Hubert, P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, 1985.

[35] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[36] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probabil- ity*, vol. 1, 1967, pp. 281–297, Oakland, CA, USA.

[37] S. Chakraborty, S. Das, "k- means clustering with a new divergence-based distance metric: Convergence and performance analysis," *Pattern Recognition Letters*, vol. 100, pp. 67–73, 2017.

[38] L. Hubert, P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

Ayan Seal

Ayan Seal received the PhD degree in engineering from Jadavpur University, West Bengal, India, in 2014. He is currently an Assistant Professor with the Computer Science and Engineering Department, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Madhya Pradesh, 482005, India. He is also associated with Center for Basic and Applied Research, Faculty of Informatics and Management, University, of Hradec Kralove, Rokitanskeho 62, Hradec Kralove, 50003, Czech Republic, He has visited the Universidad Politecnica de Madrid, Spain as a visiting research scholar. He is the recipient of several awards. Recently, he has received Sir Visvesvaraya Young Faculty Research Fellowship from Media Lab Asia, Ministry of Electronics and Information Technology, Government of India. He has authored or co-authored of several journals, conferences and book chapters in the area of biometric and medical image processing. His current research interests include image processing and pattern recognition.

Aditya Karlekar

Aditya Karlekar received the B.Tech degree in engineering from Hitkarini College of Engineering and Technology, Jabalpur, 482005. He is currently an intern with the Computer Science and Engineering Department, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Madhya Pradesh, India. His current research interests include image processing and pattern recognition.

Ondrej Krejcar

Ondrej Krejcar is a full professor in systems engineering and informatics at the University of Hradec Kralove, Czech Republic. In 2008 he received his Ph.D. title in technical cybernetics at Technical University of Ostrava, Czech Republic. He is currently a vice-rector for science and creative activities of the University of Hradec Kralove from June 2020. At present, he is also a director of the Center for Basic and Applied Research at the University of Hradec Kralove. In years 2016-2020 he was vice-dean for science and research at Faculty of Informatics and Management, UHK. His h-index is 19, with more than 1250 citations received in the Web of Science. In 2018, he was the 14th top peer reviewer in Multidisciplinary in the World according to Publons and a Top Reviewer in the Global Peer Review Awards 2019 by Publons. Currently, he is on the editorial board of the MDPI Sensors IF journal (Q1/Q2 at JCR), and several other ESCI indexed journals. He is a Vice-leader and Management Committee member at WG4 at project COST CA17136, since 2018. He has also been a Management Committee member substitute at project COST CA16226 since 2017. Since 2019, he has been Chairman of the Program Committee of the KAPPA Program, Technological Agency of the Czech Republic as a regulator of the EEA/Norwegian Financial Mechanism in the Czech Republic (2019-2024). Since 2020, he has been Chairman of the Panel 1 (Computer, Physical and Chemical Sciences) of the ZETA Program, Technological Agency of the Czech Republic. Since 2014 until 2019, he has been Deputy Chairman of the Panel 7 (Processing Industry, Robotics, and Electrical Engineering) of the Epsilon Program, Technological Agency of the Czech Republic. At the University of Hradec Kralove, he is a guarantee of the doctoral study program in Applied Informatics, where he is focusing on lecturing on Smart Approaches to the Development of Information Systems and Applications in Ubiquitous Computing Environments. His research interests include Control Systems, Smart Sensors, Ubiquitous Computing, Manufacturing, Wireless Technology, Portable Devices, biomedicine, image segmentation and recognition, biometrics, technical cybernetics, and ubiquitous computing. His second area of interest is in Biomedicine (image analysis), as well as Biotelemetric System Architecture (portable device architecture, wireless biosensors), development of applications for mobile devices with use of remote or embedded biomedical sensors.

Enrique Herrera-Viedma

Enrique Herrera-Viedma received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Granada, Spain, in 1993 and 1996, respectively. He is a Professor of computer science and the Vice-President for Research and Knowledge Transfer with University of Granada, Granada, Spain. His h-index is 81 with more than 23 000 citations received in Web of Science and 97 in Google Scholar with more than 37000 cites received. He has been identified as one of the world's most influential researchers by the Shanghai Center and Thomson Reuters/Clarivate Analytics in both computer science and engineering in the years 2014, 2015, 2016, 2017, 2018, 2019 and 2020. His current research interests include group decision making, consensus models, linguistic modeling, aggregation of information, information retrieval, bibliometric, digital libraries, web quality evaluation, recommender systems, and social media. Dr. Herrera-Viedma is Vice President for Publications in System Man & Cybernetic Society and an Associate Editor in several journals such as IEEE Transactions on Fuzzy Systems, IEEE Transactions on Systems, Man, and Cybernetics: Systems, IEEE Transactions on Intelligent Transport System, Information Sciences, Applied Soft Computing, Soft Computing, Fuzzy Optimization and Decision Making, and Knowledge-Based Systems.