IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 13, 2021, accepted May 22, 2021, date of publication May 26, 2021, date of current version June 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3083969

Clustering Uncertain Data Objects Using Jeffreys-Divergence and Maximum Bipartite Matching Based Similarity Measure

KRISHNA KUMAR SHARMA[1,2], AYAN SEAL[1,3], (Senior Member, IEEE), ANIS YAZIDI[4,5,6], (Senior Member, IEEE), ALI SELAMAT[3,7], (Member, IEEE), AND ONDREJ KREJCAR[3,7]

[1]Department of Computer Science and Engineering, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Jabalpur 482005, India
[2]Department of Computer Science and Informatics, University of Kota, Kota 324005, India
[3]Center for Basic and Applied Research, Faculty of Informatics and Management, University of Hradec Kralove, 50003 Hradec Kralove, Czech Republic
[4]Department of Computer Science, Oslo Metropolitan University, 460167 Oslo, Norway
[5]Malaysia Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur 54100, Malaysia
[6]Department of Computer Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway
[7]Department of Plastic and Reconstructive Surgery, Oslo University Hospital, 0424 Oslo, Norway

Corresponding author: Ayan Seal (ayanseal30@ieee.org)

This work is partially supported by the Ministry of Education, Youth and Sports of Czech Republic (project ERDF no. CZ.02.1.01/0.0/0.0/18_069/0010054); project ''Smart Solutions in Ubiquitous Computing Environments'', Grant Agency of Excellence, University of Hradec Kralove, Faculty of Informatics and Management, Czech Republic (under ID: UHK-FIM-GE-2204/2021); project at Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, Malaysia Research University Network (MRUN) Vot 4L876 and the Fundamental Research Grant Scheme (FRGS) Vot5F073 supported by the Ministry of Education Malaysia for the completion of the research.

ABSTRACT In recent years, uncertain data clustering has become the subject of active research in many fields, for example, pattern recognition, and machine learning. Nowadays, researchers have committed themselves to substitute the traditional distance or similarity measures with new metrics in the existing centralized clustering algorithms in order to tackle uncertainty in data. However, in order to perform uncertain data clustering, representation plays an imperative role. In this paper, a Monte-Carlo integration is adopted and modified to express uncertain data in a probabilistic form. Then three similarity measures are used to determine the closeness between two probability distributions including one novel measure. These similarity measures are derived from the notion of Kullback-Leibler divergence and Jeffreys divergence. Finally, density-based spatial clustering of applications with noise and *k*-medoids algorithms are modified and implemented on one synthetic database and three real-world uncertain databases. The obtained outcomes confirm that the proposed clustering technique defeats some of the existing algorithms.

INDEX TERMS Uncertain data clustering, probability density estimation, bipartite matching.

I. INTRODUCTION

In data mining, data uncertainty entails some deviation of the data from the ground truth due to small perturbations often known as noise or uncertainty. In the era of big data, uncertainty is one of the inherent characteristics of data. Nowadays, data is growing constantly in volume as people are becoming more connected than ever before through the internet. Uncertain data is found in abundance today in web applications, IoT sensor networks [1], [2], within enterprises [3], [4]. Uncertain data manifest both in structured and unstructured sources due to outdated sensors, inaccurate measurement, or sampling errors. For example, uncertainty is observed frequently in weather and climate prediction. Small and random perturbations to the atmospheric state variables viz., pressure, temperature, winds, and humidity readings captured by various sensors due to aging of the sensors or atmosphere itself is non-linear which in turn results in forecast divergence from the actual reality.

In recent years, uncertain data clustering has emerged as an indispensable mining task for pattern recognition and statistical analysis [5], [6] because uncertainty exists in almost

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

all real-world applications nowadays. A clustering algorithm can help to avoid risks and to take decisions precautiously. Previously various extensions of the conventional unsupervised machine learning algorithms designed for certain data were used to cluster uncertain data. For example, uncertain versions of k-means, Uk-means [7], [8], k-medoids, Uk-medoids [9], Uk-center clustering [10], and Uk-centroid clustering [11], [12]. However, most of these algorithms fail to cluster uncertain data properly because instead of considering probability distribution function (PDF), these algorithms consider a single point regarding an object like a certain database. Normally, an uncertain data object can be modeled by a PDF. Cormode *et al.* represented uncertain data with the help of point PDF and then found a closeness between two distributions by applying Euclidean distance [10]. Gullo *et al.* expressed uncertain data objects as PDFs which showed that the likelihood of a data object becoming visible at every location in a multi-dimensional region. These clustering algorithms depend upon geometric distance metric, hence, the difference between uncertain data with various distributions which are heavily overlapped are difficult to identify. So, the previous works that extended conventional unsupervised machine learning algorithms to group uncertain data are restricted to using geometric distance as similarity measures, and cannot acquire the actual distance between two uncertain data objects with different PDFs [13]. In [14], Jiang *et al.* represented uncertain data objects by applying kernel density estimation (KDE) [15] in discrete and continuous domains. Further, Kullback-Leibler (KL)-divergence was employed to measure the closeness between uncertain data objects, and combine it into partitioning clustering algorithm, density-based spatial clustering of applications with noise (DBSCAN), and k-medoids to group uncertain data objects [14], [16], [17]. Some other popular modeling schemes are kernel density estimation [14], [15], fuzzy-logic [18], evidence-oriented using Dempster-Shafer (DS) [19], [20], a stochastic method using Monte Carlo simulation (MCS) [21]–[23] etc. In our previous study [24], the Monte Carlo integration (MCI) based probabilistic approach was applied to model the uncertain data and it was proved that the MCI based modeling technique provides better clustering accuracy over some of the existing modeling techniques [18], [19], [21], [23]. The matching of multi-views/features is a common problem in uncertain data/object clustering. In general, the expected distance was considered in most of the previous methods for uncertain data clustering [7], [8], [8] along with Uk-means [7], Uk-medoids [9], fast DBSCAN [17]. However, some other distance measures such as geometric distance [16], uncertain distance [9], maximum distance density and weighted intersection [25] were also adopted in the literature. Clustering approaches with the above said distance methods can work on uncertain data to some extent. Moreover, the above-mentioned distance methods depend on the geometric positions of uncertain data. However, these algorithms did not appraise the probability distributions to represent uncertain data. So, these clustering approaches cannot distinguish the dissimilarity between uncertain data with various distributions that are extremely overlapped in locations. Furthermore, divergence-based similarity measures were used to handle the shortcomings of the geometric distances. Indeed, researchers have classically adopted divergence-based similarity measures, for example, KL-divergence [14], Jeffrey (J)-divergence [24] for clustering uncertain data. However, both the above discussed divergence-based similarity measures are non-linear and do not comply with the metric property. Moreover, some studies were based on the DS evidence theory which entails a new Belief Jensen–Shannon divergence to find the variance and conflict level between pieces of evidence for multi-sensor data fusion. These ideas emanate from the concept of evidence and belief entropy developed by F. Xiao [26]. Furthermore, Xiao improved the previous method [26] by introducing a reinforced belief divergence method to compute the difference between basic belief assignments in the DS evidence theory [27]. However, general evidence theory combines evidence from various sources and reaches a degree of belief by considering into account all the viable evidence, thus requires a large computation. Apart from probability theory to handle uncertain data, a data-driven structure can be an interesting approach to cope with uncertainty in the data along with better similarity measures because machine learning algorithms based on data-driven structures are more robust to outliers and uncertainties. Recently, Cavaliere *et al.* presented a layered geometrical structure that helps in analyzing the relationships among the data to achieve better clustering [28]. In [29], Kang *et al.* provided a graph-based learning framework that captures global and local data features to achieve robust clustering.

This paper presents a method to show the matching problem between two PDFs using a directed acyclic graph, where every node denotes a sample point of a PDF and arcs represent the similarity between two nodes. Here, the similarity is measured by applying symmetric J-divergence. Matching is required to compute the overall measure of closeness between two PDFs. This matching problem can also be named as the largest isomorphic subgraph problem. Then the majority voting scheme is considered to compute the final closeness between two distributions for such multi-views data. The similarity measure proposed in this work has been integrated with conventional k-medoids and DBSCAN clustering algorithms. Each of the experiments is performed on a weather database from the CPC of the National Centers for Environmental Prediction (NCEP), Japanese vowels, activities of daily living (ADL), and synthetic data [14] to authenticate our mentioned approach over state-of-the-art approaches.

The remaining work is structured as given below: The proposed measure of closeness and its use in clustering are presented in section II. In section III, we report the results that are obtained by various clustering algorithms including the proposed two approaches. Finally, conclusions are drawn in section IV.

II. PROPOSED METHODS

This section presents the proposed similarity measure between two PDFs, which is based on the concepts of directed acyclic graph and J-divergence. This section also describes clustering algorithms, which consider the proposed similarity measure. Later, the multivariate version of the proposed similarity measure is addressed along with majority of voting decision rules. Lastly, updated k-medoids and modified DBSCAN approaches are discussed for uncertain data clustering.

A. METHODS FOR ESTIMATING SIMILARITY BETWEEN TWO PDFs

An uncertain data object is expressed as a PDF using MCI, MCS, DS, and KDE modeling schemes using either univariate or multivariate random variable(s). In the field of statistics, the term univariate alludes to a probability distribution of an uncertain data object having one variable. Two divergences namely, KL-divergence, J-divergence, and the proposed similarity measure are considered for this work in the modified DBSCAN and k-medoids clustering algorithms.

1) ESTIMATION OF DIVERGENCE FOR UNIVARIATE DISTRIBUTION

Definition 1: In a continuous domain, ζ, Eq. 1 is applied to compute the closeness between two PDFs $\mathbf{P}$ *and* $\mathbf{Q}$ *over the same variable x, a measure called KL-divergence, where,* $\mathbf{P}$ *denotes a posterior distribution of data while* $\mathbf{Q}$ *represents prior distribution of* $\mathbf{P}$.

$$KL(\mathbf{P}||\mathbf{Q}) = \int_{x\in\zeta} \mathbf{P}(x)\log(\frac{\mathbf{P}(x)}{\mathbf{Q}(x)})dx \tag{1}$$

The difference between the two PDFs $\mathbf{P} = \{p_1, \ldots, p_s\}$ and $\mathbf{Q} = \{q_1, \ldots, q_s\}$ in discrete domain can be calculated by Eq. 2 [30].

$$KL(\mathbf{P}||\mathbf{Q}) = \sum_{x\in\zeta} \mathbf{P}(x)\log(\frac{\mathbf{P}(x)}{\mathbf{Q}(x)}), \tag{2}$$

where $\mathbf{P}(x) > 0$ and $\mathbf{Q}(x) > 0$ for any $x \in \zeta$. Equation 2 can also be expressed as Eq. 3.

$$KL(\mathbf{P}||\mathbf{Q}) = \sum_{p_i\in\mathbf{P}} \mathbf{P}(p_i)\log(\frac{\mathbf{P}(p_i)}{\mathbf{Q}(p_i)}) \tag{3}$$

Sometimes, it may happen that the value of $\mathbf{P}(x)$ is equal to zero then Eq. 4 is used to smoothing.

$$\mathbf{P}'(x) = \frac{\mathbf{P}(x)+\delta}{1+\delta|\zeta|}, \tag{4}$$

in which $|\zeta|$ is the number of sample data points of discrete domain, ζ and a constant δ used to smooth probability function, $0 < \delta < 1$. Furthermore, the aggregation of $\mathbf{P}'(x)$ in the whole domain mentioned previously is 1. In the rest of the article, instead of using $\mathbf{P}'(x)$, $\mathbf{P}(x)$ is used to represent normalized density. J-divergence reduces the type-I and type-II errors [31], [32] of KL-divergence by producing large J-divergence, which is expressed by Eq. 5.

$$J(\mathbf{P}||\mathbf{Q}) = KL(\mathbf{P}||\mathbf{Q}) + KL(\mathbf{Q}||\mathbf{P}) \tag{5}$$

Therefore, J-divergence could be sufficient to measure the complexity as well as it serves to distinguish the given hypothesis. So, a similarity measure is proposed with the help of J-divergence and the maximum bipartite matching algorithm in this work.

2) A SIMILARITY MEASURE AND ITS PROPERTIES

A bipartite graph, $G = (V, E)$, is a graph where, every vertex belongs to one of the two disjoint sets namely, $\mathbf{V}_1$ or $\mathbf{V}_2$, and each edge, $(\mathbf{V}_1^u - \mathbf{V}_2^v)$, joins a vertex in $\mathbf{V}_1$ to a vertex in $\mathbf{V}_2$. A maximum bipartite matching problem is defined by finding out the largest subset of edges in a bipartite graph in such a way that no two selected edges share a common vertex [33]. For this study, $\mathbf{V}_1$ and $\mathbf{V}_2$ represent two PDFs namely, $\mathbf{P}$ and $\mathbf{Q}$ where each data point of $\mathbf{P}$ and $\mathbf{Q}$ denotes a vertex. A non-negative weight is allotted to each edge of the bipartite graph formed by $\mathbf{P}$ and $\mathbf{Q}$ using J-divergence, $J(p_u, q_v) = (q_v - p_u)\log\frac{q_v}{p_u}$, between two data points $p_u \in \mathbf{P}$ and $q_v \in \mathbf{Q}$ which, represents the highest degree of similarity between $p_u \in \mathbf{P}$ and $q_v \in \mathbf{Q}$. The reason behind the use of the maximum bipartite matching algorithm is finding out the closest data points of two PDFs [14], [34], [35]. Here, a greedy-approach is adopted to obtain maximum bipartite matching. The size of the maximum matching is the total number of data points, s, of either $\mathbf{P}$ or $\mathbf{Q}$ as both having the same number of data points. After matching, the order of the data points of $\mathbf{Q}$ is changed. However, the data points of $\mathbf{P}$ are still unchanged. The new data points of $\mathbf{P}$ and $\mathbf{Q}$ are stored in $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively, which is shown in Fig. 1. So, the proposed similarity measure, d_m, of the two PDFs namely, $\mathbf{P}$ and $\mathbf{Q}$ could be obtained by Eq. 6 where, r denotes the index of $\mathbf{X}_1$ and $\mathbf{X}_2$. The pseudo-code of finding a maximum matching of two PDFs is illustrated in algorithm 1.

$$d_m(\mathbf{P}, \mathbf{Q}) = \sum_{r=1}^{s}(\mathbf{X}_2(r) - \mathbf{X}_1(r))\log\frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)} \tag{6}$$

Few attributes of the mentioned similarity measure are discussed as follows:

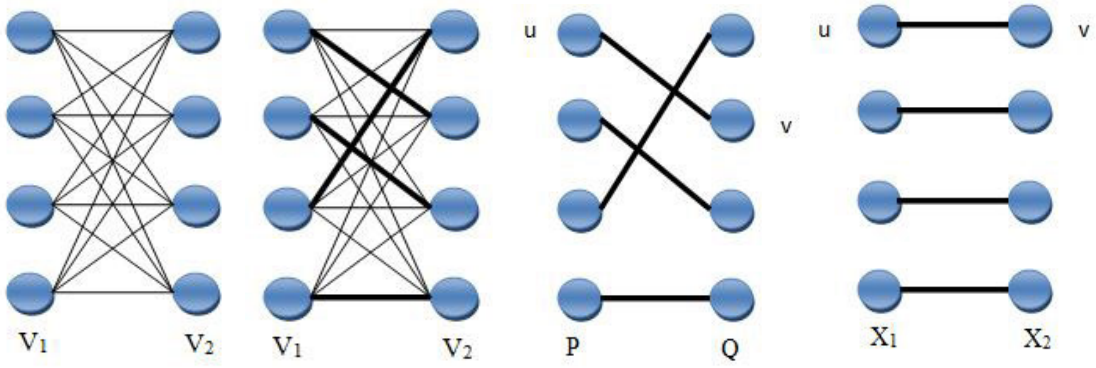
Proposition 2: $d_m(\mathbf{P}||\mathbf{Q}) \geq 0$ *and* $d_m(\mathbf{P}||\mathbf{Q}) = 0$ *iff* $\mathbf{P} = \mathbf{Q}$

Proof: According to Eq. 6 divergence based on maximum bipartite matching can be estimated between two PDF, $\mathbf{P}$ and $\mathbf{Q}$ as Eq. 7.

$$.d_m(\mathbf{P}||\mathbf{Q}) = \sum_{r=1}^{s}(\mathbf{X}_2(r) - \mathbf{X}_1(r))\log\frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)}, \tag{7}$$

where $\forall r \quad (\mathbf{X}_2(r) - \mathbf{X}_1(r))\log\frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)} \geq 0$ and $d_m(\mathbf{P}||\mathbf{Q}) = 0$ iff $\forall r \quad \mathbf{X}_2(r) = \mathbf{X}_1(r)$. So, if $\mathbf{P} = \mathbf{Q}$ then $d_m(\mathbf{P}||\mathbf{Q}) = 0$. □

Proposition 3: $d_m(\mathbf{P}||\mathbf{Q}) = d_m(\mathbf{Q}||\mathbf{P})$



(a) A bipartite Graph (b) (Maximum) Matching (c) (Maximum) Matching before ordering (d) (Maximum) Matching after ordering

FIGURE 1. Illustration of maximum bipartite matching.

Algorithm 1 A Pseudo-Code to Estimate Divergence Between Two PDF Using a Similarity Measure Based on Maximum Bipartite Matching.

INPUT: **P**, **Q** ▷ two PDFs
OUTPUT: $d_m(\mathbf{P}||\mathbf{Q})$ ▷ resultant similarity measure
for $u = 1$ to s **do**
  for $v = 1$ to s **do**
    edges[u, v]=$(q_v - p_u) \log \frac{q_v}{p_u}$
  end for
end for
Sorting of edges in ascending order using any linear sorting algorithm
$r = 0$ ▷ index for adding maximum optimal match
for $u = 1$ to s **do**
  for $v = 1$ to s **do**
    if end points of edges are not marked visited **then**
      $\mathbf{X}_1(r) = p_u$ ▷ the u^{th} data point in **P** and p_u is marked as visited
      $\mathbf{X}_2(r) = q_v$ ▷ the v^{th} data point in **Q** and q_v is marked as visited
      $r = r + 1$
    end if
  end for
end for
$d_m(\mathbf{P}||\mathbf{Q}) = \sum_{r=1}^{s}(\mathbf{X}_2(r) - \mathbf{X}_1(r)) \log \frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)}$

Proof: According to Eq. 7, $d_m(\mathbf{P}||\mathbf{Q}) = \sum_{r=1}^{s}(\mathbf{X}_2(r) - \mathbf{X}_1(r)) \log \frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)} = \sum_{r=1}^{s}(\mathbf{X}_1(r) - \mathbf{X}_2(r)) \log \frac{\mathbf{X}_1(r)}{\mathbf{X}_2(r)} = d_m(\mathbf{Q}||\mathbf{P})$ □

Theorem 4: f-divergence: The similarity measure proposed in this work.

Proof: The proposed similarity measure between PDFs $\mathbf{P} \in [0, 1]^s$ and $\mathbf{Q} \in [0, 1]^s$ can be given according to Eq. 7 as Eq. 8,

$$d_m(\mathbf{P}||\mathbf{Q}) = \sum_{r=1}^{s}(\mathbf{X}_2(r) - \mathbf{X}_1(r)) \log \frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)} \tag{8}$$

Putting $z_r = \frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)}$ in Eq. 8 $d_m(\mathbf{P}||\mathbf{Q}) = \sum_{r=1}^{s}(z_r \times \mathbf{X}_1(r) - \mathbf{X}_1(r)) \log \frac{z_r \times \mathbf{X}_1(r)}{\mathbf{X}_1(r)} \implies d_m(\mathbf{P}||\mathbf{Q}) = \sum_{r=1}^{s}((z_r - 1) \times \mathbf{X}_1(r)) \log(z_r) = \sum_{r=1}^{s} \mathbf{X}_1(r)\psi(z_r) = \sum_{r=1}^{s} \mathbf{X}_1(r)\psi(\frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)})$ $d_m(\mathbf{P}||\mathbf{Q})$ can be expressed as $\sum_{r=1}^{s} \mathbf{X}_1(r)\psi(\frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)})$. Hence, the proposed similarity measure is a f-divergence. □

Theorem 5: Bregman divergence: Not equivalent to the similarity measure proposed in this work.

Proof: If the similarity measure proposed is equivalent to a Bregman divergence, $d_m(\mathbf{P}||\mathbf{Q})$ would be strictly convex in **P**. So, our objective is to prove that $d_m(\mathbf{P}||\mathbf{Q})$ is not convex in **P**, which is as follows according to Eq. 7. $d_m(\mathbf{P}||\mathbf{Q}) = \sum_{l=1}^{s}(\mathbf{X}_2(r) - \mathbf{X}_1(r)) \log \frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)}$ Hence, $\frac{\delta d_m}{\delta \mathbf{X}_1(r)} = \frac{\mathbf{X}_2(r)}{\mathbf{X}_1(r)} - \log(\mathbf{X}_1(r)) - 1 - \log(\mathbf{X}_2(r))$ If $r \neq j$ $\frac{\delta^2 d_m}{\delta \mathbf{X}_1(j)\delta \mathbf{X}_1(r)} = 0$, and $\frac{\delta^2 d_m}{\delta^2 \mathbf{X}_1(r)} = -\frac{1}{\mathbf{X}_1^2(r)} - \frac{1}{\mathbf{X}_1(r)}$ Now, taking $\mathbf{X}_1(r)$ and $\mathbf{X}_2(r)$ for $\forall l$, we get $\frac{\delta^2 d_m}{\delta^2 \mathbf{X}_1(r)} < 0$. Hence, d_m divergence is not convex in **P**. □

3) ESTIMATION OF DIVERGENCE FOR MULTIVARIATE DISTRIBUTION

Generally, univariate or one feature of an object is not sufficient to separate itself from others. So, there is a need to use multivariate or multiple features. We can say in the area of statistics, the univariate PDF is a generalized form of the multivariate PDF in a larger dimension. We have discussed and proposed two approaches in this section, to calculate a matching within two multivariate distributions by employing d_m similarity measure. However, two other similarity measures based on KL-divergence and J-divergence are used for comparison purposes. A conventional clustering algorithm is utilized to substitute the similarity measure step in the proposed two approaches. The primary assumption of these

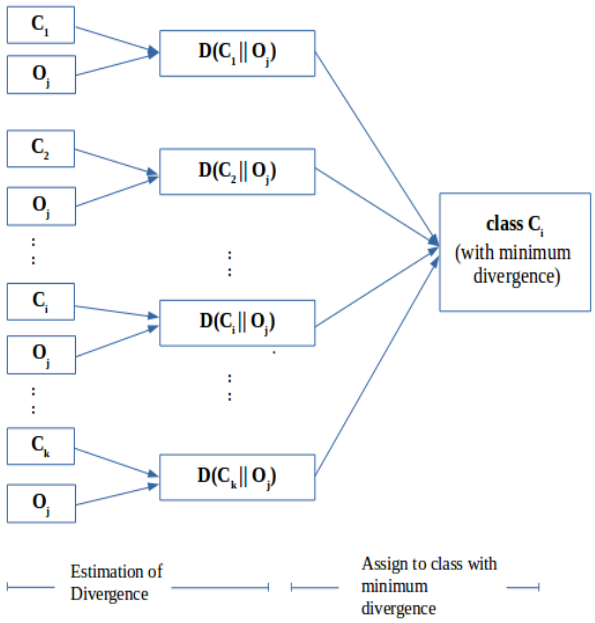


FIGURE 2. Divergence for multivariate case using product method.

approaches is that the corresponding characteristics of the distributions are exclusive to each other. One of the traditional clustering algorithms called k-medoids is adopted and is modified in this work where k is the total number of clusters having d-dimensional features. So, the difference between two PDFs of two uncertain data objects would be computed using Eqs. 9, 10, and 11. Fig. 2 illustrates how similarity is estimated between two PDFs of two uncertain data objects and a cluster representative using Eqs. 6, 9, and 10.

$$KL(C_i||O_j) = \sum_{x_i \in C_i} (C_{i,1}(x_{i,1}) \wedge \cdots \wedge C_{i,d}(x_{i,d})) \log\left(\frac{(C_{i,1}(x_{i,1}) \wedge \cdots \wedge C_{i,d}(x_{i,d}))}{(O_{j,1}(x_{i,1}) \wedge \cdots \wedge O_{j,d}(x_{i,d}))}\right) \quad (9)$$

$$J(C_i||O_j) = KL(C_i||O_j) + KL(O_j||C_i) \quad (10)$$

$$d_m(C_i||O_j) = \sum_{l=1}^{d} d_m(C_{i,l}||O_{j,l}), \quad (11)$$

where i, j and l represent i^{th}-cluster, j^{th}-uncertain data object and l^{th} feature respectively. Moreover, $x_{i,\_}$ indicates a specific data point of each distributions.

The closeness within the distributions of an uncertain data object and a cluster representative is measured by applying a majority voting rule [36]–[38]. The majority voting scheme executes in two steps: generation and consensus of votes. The closeness within the l^{th} distribution of an uncertain data object and k-cluster representatives are estimated by applying Eq. 12 and is saved in the array of dimension k named $dist_l$ in the first step.

$$dist_l = [D(C_{1,l}||O_{j,l}), D(C_{2,l}||O_{j,l}), \ldots, D(C_{i,l}||O_{j,l}), \ldots, D(C_{k,l}||O_{j,l})], \quad (12)$$

in which i and j form the indicatives of the i^{th} cluster and j^{th} uncertain data object respectively. Furthermore, a vote v_l of the l^{th}-distribution for the j^{th} uncertain data object is calculated by applying Eq. 13.

$$[va, v_l] = min(dist_l), \quad (13)$$

where v_l is the index of va, which is the first smallest value in $dist_l$. Similarly, $V = \{v_1, v_2, \ldots, v_d\}$, vector would be created using Eqs. 12 and 13. The vector V consists of d components and d is the number of features of an uncertain data object. Then O_j, uncertain data object, will be clustered to i^{th}-class if the highest votes are given using Eq. 14 in the consensus stage. Fig. 3 shows the estimation of similarity between two PDFs for a multivariate case using the majority voting decision rule.

$$C_i = \max_i \{v_1, v_2, \ldots, v_d\} \quad (14)$$

If two and more classes have obtained the same number of votes then this tie is resolved by random allocation to O_j and the associated random number is generated by the Mersenne Twister algorithm [39].

B. UPDATED k-MEDOIDS CLUSTERING ALGORITHM AND ANALYSES

It is clear from the literature that k-medoids is one of the popularly used unsupervised algorithms for clustering uncertain data objects [14]. So, the conventional k-medoids algorithm is modified with the application of the similarity measure proposed in this study.

Let $O = \{O_1, \ldots, O_\Gamma\}$ be a set, which consists of PDFs of 'Γ' uncertain data objects in $[0, 1]^{sd}$. The k-medoids clustering algorithm aims to divide 'Γ' uncertain data objects in 'k' groups. Mathematically, the definition of a clustering algorithm is as follows:

$$\begin{aligned} \chi : minimize\, h(M, C) &= \sum_{i=1}^{\Gamma}\sum_{j=1}^{k} m_{ij} D(O_i||C_j) \\ where, \sum_{j=1}^{k} m_{ij} &= 1,\ m_{ij} = 0\, or\ 1, \\ \forall i \in \{1, \ldots, \Gamma\}, \forall j &\in \{1, \ldots, k\} \\ C = \{C_1, \ldots, C_k\}\, C_j &\in [0, 1]^s\, \forall j \in \{1, \ldots, k\} \\ D(O_i||C_j)\ &is\ a\ similarity\ measure \\ between\ two\ &uncertain\ objects \end{aligned} \quad (15)$$

The solution of χ can be obtained iteratively using algorithm 2. The modified algorithm has two steps namely, the building step and swapping step [14], [40].

$$Y = \{M \in [0, 1]^{\Gamma k} : \sum_{j=1}^{k} m_{ij} = 1,\ m_{ij} \geq 0 \quad \forall i \in \{1, \ldots, \Gamma\}, \forall j \in \{1, \ldots, k\}\} \quad (16)$$

The utmost point of Y obeys the condition of Eq. 15. The χ' is known as the reduced problem of χ, which is redefined as follows:

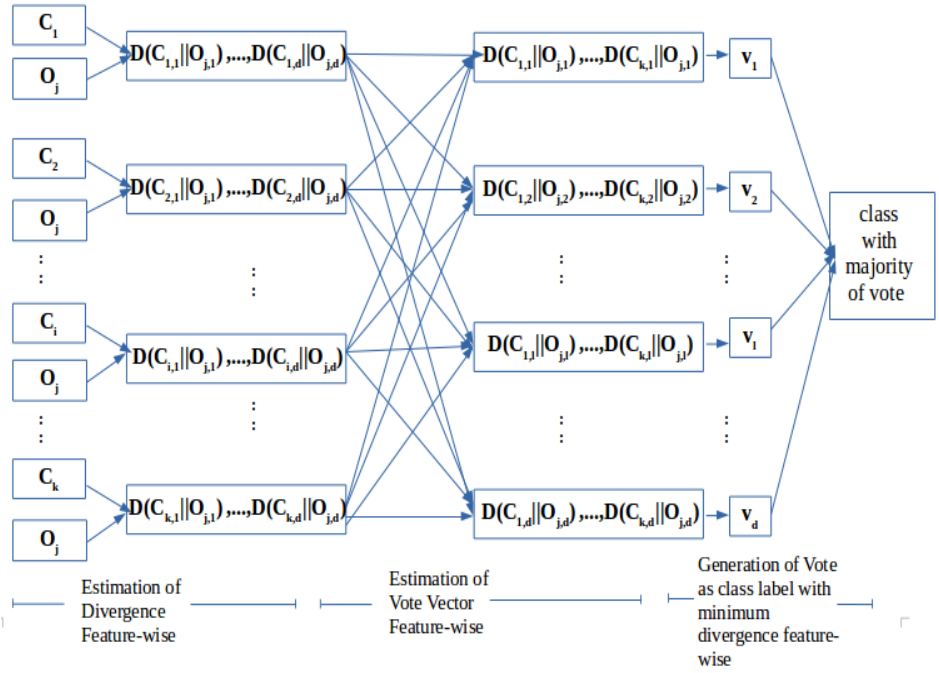


FIGURE 3. Majority voting decision rule integrates with divergences to compute the closeness between two PDFs.

Definition 6: The problem χ can be written as χ': χ': minimize H(M), where H(.) is a concave function and $M \in Y$

Remark: if $H(.)$ is a concave function then there exists extreme point solution for the problem χ' lies in Y with defined constraints in Eq. 16 similar to problem χ. If χ problem is formulated using $m_{ij} \in [0, 1]$ then solution of χ will satisfy $m_{ij} = 0$ or 1. In other words, if $H(M, C)$ is a concave in M for any fixed constraints C then the solution of problem χ will satisfy $m_{ij} = 0$ or 1 and in case of convex function estimation of class label will be difficult [41].

Earlier result of problem χ has an optimal solution. Now, it's time to characterize the partial optimal solution of problem χ.

Definition 7: A point (M^*, C^*) will be a partial optimal solution to problem χ if $h(M^*, C^*) \leq h(M, C^*)$, $\forall M \in Y$ $h(M^*, C^*) \leq h(M^*, C)$, $\forall C \in [0, 1]^{sd}$ This partial optimal solution is further solved with the help of two sub-problems in each successive step of algorithm 2:

- *Problem χ_1: minimization of $h(M, \hat{C})$ subject to $M \in Y$ and a fixed value of $\hat{C} \in [0, 1]^{sd}$*
- *Problem χ_2: minimization of $h(\hat{M}, C)$ subject to a fixed value of $\hat{M} \in Y$ and $C \in [0, 1]^{sd}$*

χ_1 and χ_2 are solved iteratively. The solution of problem χ_1 is implicit for a specific O_i, $m_{iv} = 1$ if $D(O_i, C_v) \leq D(O_i, C_j)$ $j, v = 1, 2, \ldots, k$ However, the solution of problem χ_2 is not as easy as χ_1. The procedure for the solution of χ_2 is described in next theorem.

Lemma 8: The problem χ_2 for any fixed $M_0 \in Y$ has a solution if J-divergence based similarity measure d_m is used in k-medoids clustering.

Proof: Let us prove that if $\hat{C}$ is a solution of χ_2 for a given $M_0 \in Y$ then we have $h(M, C) = \sum_{i=1}^{\Gamma} \sum_{j=1}^{k} m_{ij} d_m(O_i, C_j) \implies h(M, C) = \sum_{i=1}^{\Gamma} \sum_{j=1}^{k} m_{ij} \sum_{r=1}^{s} ((C_j(r) - O_i(r)) \log \frac{C_j(r)}{O_i(r)})$ If $\hat{C}$ is a solution of χ_2 for any given $M_0 \in Y$ then $\hat{C}$ must satisfy the following $\frac{\delta h}{\delta C_j(r)}\Big|_{C_j(r)=C_j\hat{(r)}} = 0$ Now, for any fixed $M \in Y$ $\frac{\delta h}{\delta C_j(r)}\Big|_{C_j(r)=C_j\hat{(r)}} = \sum_{i=1}^{\Gamma} m_{ij}(1 + \log(C_j(r)) - \frac{O_i(r)}{C_j(r)} - \log(O_i(r))) \frac{\delta h}{\delta C_j(r)}\Big|_{C_j(r)=C_j\hat{(r)}} = 0$ implies that

$$\sum_{i=1}^{\Gamma} m_{ij} \frac{1}{\frac{1}{1 + \log(C_j\hat{(r)})}} = \sum_{i=1}^{\Gamma} m_{ij} \frac{1}{\frac{C_j\hat{(r)}}{O_i(r) + C_j\hat{(r)} \log(O_i(r))}} \tag{17}$$

Now, if we take $min_{1 \leq i \leq \Gamma} O_i(r) = O_i\hat{(r)}$ and apply in Eq. 17 then

$$\sum_{i=1}^{\Gamma} m_{ij} \frac{1}{1 + \log(C_j\hat{(r)})} \leq \sum_{i=1}^{\Gamma} m_{ij} \frac{C_j\hat{(r)}}{\hat{O}_i(r) + C_j\hat{(r)} \log(O_i\hat{(r)})}$$

$\implies min_{1 \leq i \leq \Gamma} O_i(r) \leq C_j\hat{(r)}$ Similarly, $max_{1 \leq i \leq \Gamma} O_i(r) \geq C_j\hat{(r)}$

Hence, $max_{1 \leq i \leq n} O_i(r) \geq C_j\hat{(r)} \geq min_{1 \leq i \leq \Gamma} O_i(r)$ □

Lemma 9: The problem χ_2 for any fixed $M_0 \in Y$ has a unique solution if the proposed closeness measure, d_m, is employed in clustering.

Algorithm 2 Uncertain Data Objects Clustering Using Updated k-Medoids Algorithm

INPUT: O, k ▷ Γ-uncertain data objects and number of clusters

OUTPUT: Clusters $\mathbb{C} = \{\mathbb{C}_1, \ldots, \mathbb{C}_k\}$ ▷ **Building Phase** k representatives are selected randomly from the set of uncertain data objects O. $C_1, C_2, \ldots, C_k$

for all $O_i \in O$ **do**

O_i is assigned to closest cluster represented C_j and corresponding $m_{ij} = 1$ else 0 ▷ Using multivariate product and majority of voting methods

end for

$TD = \sum_{i=1}^{\Gamma} \sum_{j=1}^{k} m_{ij} D(O_i||C_j)$ ▷ total divergence after assignment ▷ **Swapping Phase**

repeat

Select a random uncertain data object $\mathbf{P} \in O\backslash\{C_1, \ldots, C_k\}$

if P $\in$ any cluster $\mathbb{C}_i$ **then**

C_i will be swapped by **P**

end if

for all $O_i \in O$ **do**

O_i is assigned to closest cluster represented C_j and corresponding $m_{ij} = 1$ else 0 ▷ Using multivariate product and majority of voting methods

end for

$TD1 = \sum_{i=1}^{\Gamma} \sum_{j=1}^{k} m_{ij} D(O_i||C_j)$ ▷ total divergence after assignment

DEC=TD1-TD ▷ the total decrease in divergence after swapping

$iteration \leftarrow iteration + 1$ ▷ number of iteration in algorithm

$$P_{swap}(DEC) = \begin{cases} 1 & \text{if } DEC > 0 \\ e^{DEC \times \log(iteration)} & \text{if } DEC \leq 0, \end{cases}$$

until $P_{swap}(DEC) >$ random number

Proof: Assume for a fixed $M_0 \in Y$ there are two solutions namely, C_j^1 and C_j^2 for problem χ such that $C_j^1 \neq C_j^2$, where, $C_j^1(r) \neq C_j^2(r)$ i.e. $\exists j \in \{1, \ldots, k\}$ and $r \in \{1, \ldots, s\}$. Assume, $C_j^1(r) \leq C_j^2(r)$ and apply in Eq. 17 as derived in lemma 8. Now, two solutions can be expressed according to Eq. 18.

$$\sum_{i=1}^{\Gamma} m_{ij}(1 + \log(C_j^1(r)) - \frac{O_i(r)}{C_j^1(r)}) = \sum_{i=1}^{\Gamma} m_{ij} \log(O_i(r))$$

and

$$\sum_{i=1}^{\Gamma} m_{ij}(1 + \log(C_j^2(l)) - \frac{O_i(r)}{C_j^2(r)}) = \sum_{i=1}^{\Gamma} m_{ij} \log(O_i(r)) \tag{18}$$

But, $\sum_{i=1}^{\Gamma} m_{ij}(1 + \log(C_j^1(r)) - \frac{O_i(r)}{C_j^1(r)}) \geq \sum_{i=1}^{\Gamma} m_{ij}(1 + \log(C_j^2(r)) - \frac{O_i(r)}{C_j^2(r)})$ since $C_j^1(r) \leq C_j^2(r)$ Therefore, Eq. 18 is expressed as Eq. 19

$$\sum_{i=1}^{\Gamma} m_{ij} \log(O_i(r)) \geq \sum_{i=1}^{\Gamma} m_{ij} \log(O_i(r)), \tag{19}$$

Eq. 19 is a contradiction. □

Now, the convergence of k-medoids clustering algorithm would be discussed in theorem 2.7.

Theorem 10: The modified k-medoids clustering algorithm converges to an optimal solution of problem χ in finite steps.

Proof: The k-medoids clustering algorithm converges to optimal solutions for uncertain data objects in finite steps [14]. Moreover, the solution of problem χ is also obtained in finite iterations [41]. In this study, the proposed measure of closeness combines with the k-medoids clustering algorithm to partition uncertain data objects and the objective function is defined to minimize $h(M, C)$. Lemmas 8 and 9 ensure that the problem χ has an optimal solution in finite steps. □

The conventional DBSCAN clustering algorithm is also updated in this work by applying the proposed similarity measure. The modified algorithm is as follows:

C. MODIFIED DBSCAN CLUSTERING ALGORITHM

In [17], Kriegel *et al.* first used DBSCAN in order to partition uncertain data objects. In the data region, clusters are formed as dense spaces, which are segregated by sparse spaces or spaces of lower object density [42], [43]. This proposed algorithm finds out the arbitrary shape clusters. The major concept fundamental to DBSCAN is determining core uncertain data objects, reachable uncertain data objects, density reachable uncertain data objects, and outliers [17].

Definition 11: A core uncertain data object **P** *consists of at least μ neighborhood uncertain data objects inside the radius ϵ in the set O as presented in Eq. 20.*

$$|\{\mathbf{Q} \in O \mid D(\mathbf{P}||\mathbf{Q}) \leq \epsilon\}| \geq \mu \tag{20}$$

Definition 12: **P** *and* **Q** *are both uncertain data objects and can be called direct density reachable from each other, if* **P** *is a core uncertain data object and* **Q** *is in ϵ-distance from uncertain data object* **P** *with condition to* $\mathbf{Q} \in O\backslash\{\mathbf{P}\}$.

Definition 13: **P** *and* **Q** *are both uncertain data objects and can be called reachable from each other, if w.r.t. ϵ and μ if a path of uncertain data objects $Q_1, \ldots, Q_i, Q_{i+1}, \ldots, Q_n$, with $Q_1 = \mathbf{P}$, and $Q_n = \mathbf{Q}$ such that Q_i can reach directly to uncertain data object Q_{i+1} w.r.t. ϵ and μ for $\forall i$, $1 \leq i \leq n$.*

Definition 14: Outlier uncertain data objects are not reachable from any other objects.

If **P** is any core uncertain data object and it creates a cluster along with the other non-core or core uncertain data objects or both provided they are approachable from **P**. Every cluster contains a minimum of one core uncertain

Algorithm 3 The Optimal Value of ϵ in Database Can Be Computed by Applying the Following Pseudocode

Require: O ▷ n-uncertain data objects
Require: μ ▷ density threshold ▷ M is a 2D matrix of size $n \times n$ and all the entries are o, initially
Ensure: ϵ
1: **for** $i = 1$ to n **do**
2:
3: **for** $j = 1$ to $n \wedge i \neq j$ **do**
4: $M[i, j] \leftarrow D(O_i||O_j)$
5: **end for**
6: find the first μ-smallest values from M[i,1:n] and store in W[i, 1:μ]
7: L[i]=max(W[i,1:μ])
8: **end for**
9: Sort $L[1 : n]$ in ascending order and mark these values to find the crucial variation in the curve and that corresponds to ϵ.

data object and may also contain non-core uncertain data objects. The initial stage of the modified DBSCAN is as follows: first, a cluster is formed using each core uncertain data object. Then, two clusters are integrated if any core uncertain data object would be density-reachable to some other core uncertain data object of another cluster. Moreover, all the non-core members are assigned to the nearest core objects, where distance is measured by applying multivariate divergence techniques. It proceeds until there is nil alteration in the membership of uncertain data objects of clusters.

DBSCAN relies on three input parameters: density threshold, μ, radius, ϵ. Initially, these parameters were chosen arbitrarily by most of the researchers. However, optimum values were obtained at the end of their experiments. In this study, an optimum ϵ value is computed by introducing a heuristic scheme. The proposed heuristic scheme is presented in algorithm 3, which starts with the computation of similarity measure between every set of the PDF of uncertain data objects which is then followed by determining the greatest valuation within the μ-minimum values related to each uncertain data object and save in an array. Eventually, one needs to arrange the resultant array in ascending order and mark the values to obtain the decisive fluctuation appearing in the curve and then store it as ϵ which will be one input to DBSCAN together with μ [44].

Now, algorithm 4 presents a updated DBSCAN algorithm to cluster uncertain data objects based on their distributions [14], [43].

III. EXPERIMENTAL RESULTS AND DISCUSSION

Each of the experiments is performed in Spyder 3.3.3 Python development environment in virtue of 64-bits Python 3.7.0 compiler on a laptop Intel(R) Core(TM) i5 CPU@1.80GHz and 8-GB RAM running on macOS Mojave version 10.14.5. Here, the I/O cost is not reported.

Algorithm 4 Updated DBSCAN Clustering Algorithm for Uncertain Data Objects

Require: O ▷ n-uncertain data objects
Require: k ▷ number of clusters
Require: μ ▷ Density threshold
Require: ϵ ▷ Radius obtained from Algorithm 1
Ensure: Clusters $\mathbb{C} = \{\mathbb{C}_1, \ldots, \mathbb{C}_k\}$
for each uncertain data object $\mathbf{P} \in O$ **do** ▷ Identify the core uncertain data objects
if $|\{\mathbf{Q} \in O | D(\mathbf{P}||\mathbf{Q}) \leq \epsilon\}| \geq \mu$ **then**
$S \leftarrow S \cup \mathbf{P}$
end if
end for
for each core uncertain data object $\mathbf{P} \in S$ **do** ▷ Join neighboring objects
if $\{\mathbf{Q} \in S | D(\mathbf{P}||\mathbf{Q}) \leq \epsilon \wedge label(\mathbf{Q}) \neq undefined\}$ **then**
Join such $\mathbf{Q}$ neighboring uncertain core objects into cluster C_i and label them
end if
end for
for each uncertain data object $\mathbf{P} \in O$ **do**
if $label(\mathbf{P}) \neq undefined$ **then**
label($\mathbf{P}$) $\leftarrow \min_{C_i \in \mathbb{C}} \{D(C_i||\mathbf{P}) \leq \epsilon\}$ ▷ Assign objects using multivariate product method and majority of voting method
else
label($\mathbf{P}$) $\leftarrow$ Outliers ▷ Assign Outliers
end if
end for

A. DATABASE DESCRIPTION

1) WEATHER DATA

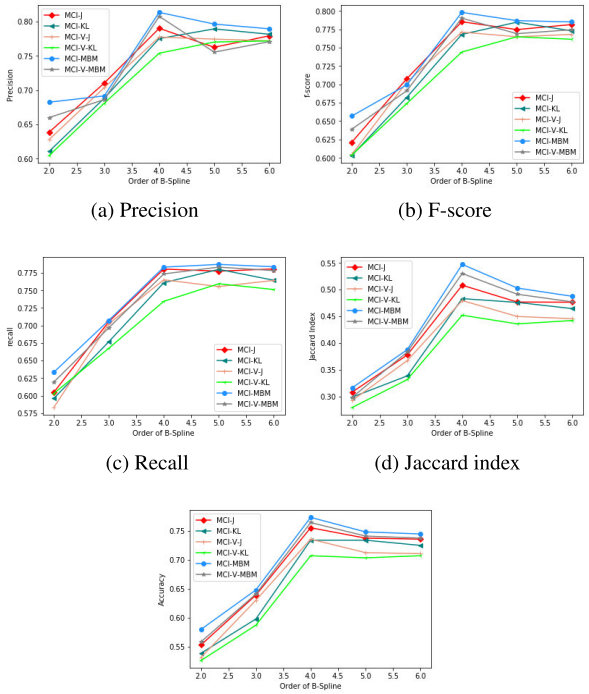
Accumulation of weather data is done from 2500 various stations by the National Center for Atmospheric Research data archive in 2008. Each station recorded daily weather data in 2008. These weather data are downloaded from http://rda.ucar.edu/datasets/ds512.0/index.html#! for this study [45]. Three features namely, average humidity, the average degree of temperature, and precipitation would be noted for every entry. Each of the stations is classified according to the Koppen-Geiger climate classification based on the type of weather. Five classes: polar climate, tropical climate, temperate climate, dry climate, and continental climate [46].

2) JAPANESE VOWELS

We have downloaded Japanese vowels from the UCI data repository (http: //arc hive.ics.uci.edu/ml/). This database consists of the utterance of two Japanese vowels spoken by 9 males, which is represented using 640 time-series data. In other words, the number of classes and the number of speakers are the same. Each time series data consists of 7 to 29 whereas, every entry has 12 linear predictive cepstrum coefficients. In this study, every time series

TABLE 1. Comparison of validation indexes for different approaches by the updated DBSCAN clustering algorithm on weather data.

Index	KDE-MBM	KDE-V-MBM	KDE-J	KDE-KL	KDE-V-J	KDE-V-KL	MCS-MBM	MCS-V-MBM	MCS-J	MCS-KL	MCS-V-J	MCS-V-KL	DS-MBM	DS-V-MBM	DS-J	DS-KL	DS-V-J	DS-V-KL	MCI-KL	MCI-V-J	MCI-V-KL	MCI-J	MCI-MBM	MCI-V-MBM
Accuracy	0.73703	0.73524	0.68694	0.68515	0.64043	0.67084	0.64401	0.63864	0.60823	0.59571	0.61717	0.58497	0.62969	0.60107	0.61717	0.60644	0.59392	0.60286	0.73524	0.72987	0.70125	0.76029	**0.81932**	0.76208
Precision	0.79657	0.79024	0.75441	0.7342	0.72507	0.72204	0.71421	0.70803	0.6845	0.67408	0.69077	0.6495	0.70385	0.66091	0.69345	0.66343	0.67435	0.66106	0.76321	0.77452	0.74958	0.79506	**0.82648**	0.80927
Recall	0.77739	0.75766	0.74461	0.72106	0.70844	0.71149	0.69801	0.69338	0.66589	0.65919	0.67819	0.64369	0.70579	0.67899	0.6744	0.6586	0.66576	0.68084	0.75171	0.76129	0.7306	0.78689	**0.83432**	0.77813
F-Score	0.78686	0.77361	0.74948	0.72757	0.71666	0.71672	0.70602	0.70063	0.67507	0.66655	0.68442	0.64658	0.70482	0.66983	0.68379	0.66101	0.67003	0.67081	0.75742	0.76785	0.73997	0.79096	**0.83038**	0.79339
Jaccard index	0.47674	0.47909	0.43588	0.43278	0.38059	0.42127	0.38097	0.37353	0.36314	0.34879	0.36591	0.3444	0.36731	0.35174	0.36908	0.37554	0.34364	0.35385	0.47338	0.47145	0.44559	0.51473	**0.62852**	0.51547

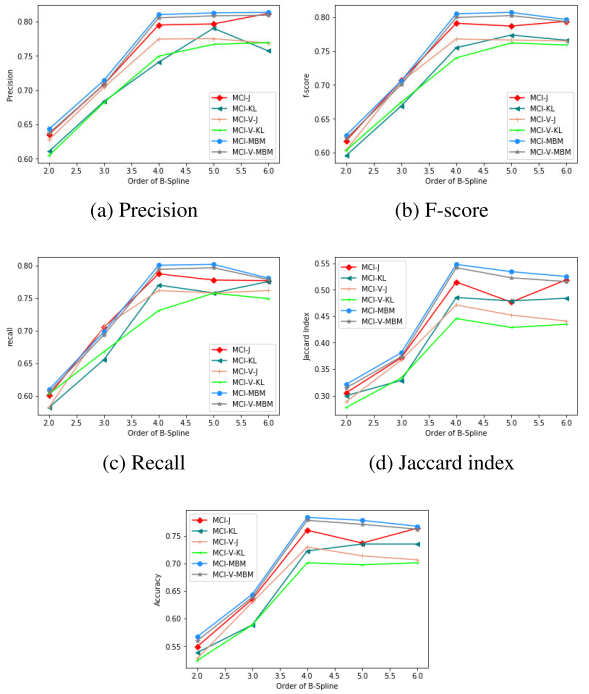


(a) Precision (b) F-score

(c) Recall (d) Jaccard index

(e) Accuracy

FIGURE 4. The performance measures in terms of validation indexes obtained by executing the updated *k*-medoids clustering algorithm on weather data.



(a) Precision (b) F-score

(c) Recall (d) Jaccard index

(e) Accuracy

FIGURE 5. The performance measures in terms of validation indexes obtained by executing the updated DBSCAN clustering algorithm on weather data.

data and each record are represented as an uncertain data object and a sample point of the uncertain data object respectively.

3) ACTIVITIES OF DAILY LIVING

The ADL database is again obtained from the UCI data repository. It contains 262778 entries, which are measured from 503 accelerometers. Every entry of the database contains 3 information about the accelerometer namely, acceleration along with the x-axis, y-axis, and z-axis. It is further divided into five categories based on daily living activities: getting up, climbing stairs, drinking, pouring water, and walking. An uncertain data object is represented by each accelerometer and each record obtained from an accelerometer is considered a sample point of the uncertain data object.

4) SYNTHETIC DATA

The synthetic database is synthesized using five distributions: Gaussian, Inverse Gaussian, Logistic, Inverse logistic, and Uniform in d-dimensional space. Five hundred different PDFs are generated for each type of distribution within [0, 1] by changing variance and sample size, s, as $0.05i$ and within 50 and 250 respectively, where i represents the i^{th} uncertain data object of a specific distribution.

B. EVALUATION METRICS

Accuracy is one of the well-known employed metrics in machine learning and it refers to the closeness of predicted value to the corresponding actual value. Intuitively, higher accuracy represents a better and more effective machine learning algorithm. However, accuracy alone can be misleading when the databases are imbalanced. Here, accuracy is considered with other metrics such as Jaccard index, precision, recall, and f-score [47], [48] for assessing the results generated by the modified clustering algorithms. These metrics also help to compare the performance of the updated clustering algorithms with existing approaches. Non-parametric statistical hypothesis test called Wilcoxon's Rank-Sum test is also performed with 5% significance level to determine whether two dependent samples are selected from the same data or not [49], [50].

TABLE 2. Comparison of validation indexes for different approaches by the updated *k*-medoids clustering algorithm on weather data.

Index	KDE-MBM	KDE-V-MBM	KDE-J	KDE-KL	KDE-V-J	KDE-V-KL	MCS-MBM	MCS-V-MBM	MCS-J	MCS-KL	MCS-V-J	MCS-V-KL	DS-MBM	DS-V-MBM	DS-J	DS-KL	DS-V-J	DS-V-KL	MCI-KL	MCI-V-J	MCI-V-KL	MCI-J	MCI-MBM	MCI-V-MBM
Accuracy	0.72987	0.72093	0.69589	0.68157	0.6458	0.67084	0.66189	0.63864	0.62075	0.58497	0.6136	0.57782	0.65474	0.64401	0.63685	0.59928	0.63506	0.60107	0.73345	0.73524	0.70662	0.75492	**0.79069**	0.76386
Precision	0.75625	0.74905	0.75441	0.7342	0.72507	0.72204	0.70686	0.68352	0.69591	0.65248	0.68721	0.64102	0.72726	0.68349	0.70158	0.66539	0.70635	0.66091	0.77535	0.77748	0.75397	0.79031	**0.8209**	0.80776
Recall	0.79151	0.79029	0.74461	0.72106	0.70844	0.71149	0.72106	0.70536	0.66918	0.64469	0.67576	0.63868	0.69876	0.69596	0.68244	0.65889	0.69529	0.67899	0.76093	0.76509	0.73442	0.78037	**0.79985**	0.77352
F-Score	0.77348	0.76912	0.74948	0.72757	0.71666	0.71672	0.71389	0.69427	0.68228	0.64856	0.68144	0.63985	0.71272	0.68967	0.69188	0.66212	0.70078	0.66983	0.76808	0.77124	0.74407	0.78531	**0.81024**	0.79027
Jaccard index	0.47982	0.46622	0.43588	0.43278	0.38059	0.42127	0.39962	0.37635	0.37217	0.35474	0.36234	0.33956	0.40088	0.38732	0.3898	0.36436	0.3831	0.35174	0.48317	0.47959	0.45209	0.5082	**0.57706**	0.53019

TABLE 3. Comparison of validation indexes for different approaches by the updated DBSCAN clustering algorithm on japanese vowel database.

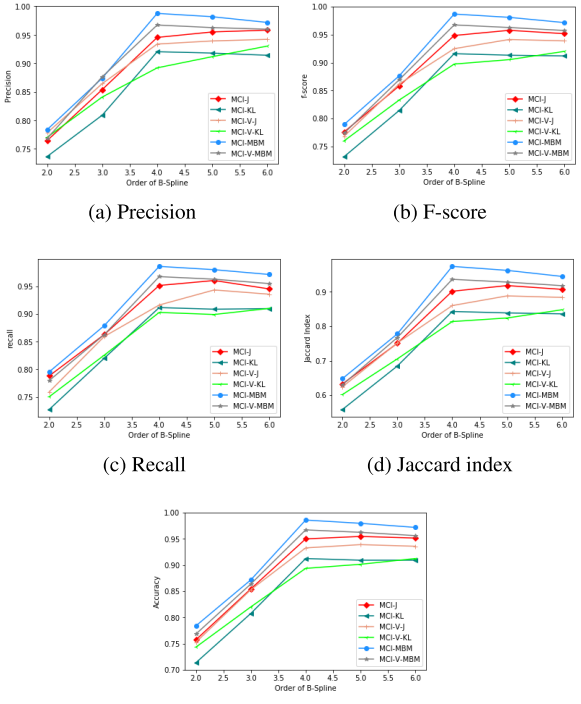
Index	KDE-MBM	KDE-V-MBM	KDE-J	KDE-KL	KDE-V-J	KDE-V-KL	MCS-MBM	MCS-V-MBM	MCS-J	MCS-KL	MCS-V-J	MCS-V-KL	DS-MBM	DS-V-MBM	DS-J	DS-KL	DS-V-J	DS-V-KL	MCI-KL	MCI-V-J	MCI-V-KL	MCI-J	MCI-MBM	MCI-V-MBM
Accuracy	0.91875	0.91094	0.90156	0.87812	0.8875	0.87969	0.84375	0.82969	0.82031	0.81406	0.825	0.80156	0.85938	0.85469	0.84062	0.80312	0.82188	0.81094	0.91135	0.92812	0.90781	0.94219	**0.97344**	0.95
Precision	0.92202	0.92366	0.90241	0.87922	0.88814	0.88044	0.85412	0.8313	0.8214	0.81522	0.83109	0.809	0.86281	0.86303	0.84503	0.81858	0.82674	0.82743	0.91168	0.92844	0.90863	0.94254	**0.97469**	0.95829
Recall	0.92888	0.91144	0.9045	0.88107	0.89209	0.88315	0.84372	0.84726	0.82286	0.81761	0.83452	0.82024	0.87071	0.85712	0.84658	0.81333	0.82995	0.81824	0.91116	0.92883	0.90743	0.9425	**0.97735**	0.95069
F-Score	0.92544	0.91751	0.90345	0.88015	0.89011	0.88179	0.84889	0.8392	0.82213	0.81641	0.8328	0.81458	0.86674	0.86007	0.8458	0.81594	0.82834	0.82281	0.91156	0.92863	0.90803	0.94252	**0.97602**	0.95448
Jaccard index	0.86295	0.84563	0.82223	0.7865	0.8019	0.78919	0.73497	0.7171	0.69608	0.6897	0.7141	0.69521	0.76459	0.75177	0.73565	0.68731	0.70768	0.69837	0.83745	0.86656	0.8314	0.89073	**0.95262**	0.91208

TABLE 4. Comparison of validation indexes for different approaches by the updated *k*-medoids clustering algorithm on japanese vowel database.

Index	KDE-MBM	KDE-V-MBM	KDE-J	KDE-KL	KDE-V-J	KDE-V-KL	MCS-MBM	MCS-V-MBM	MCS-J	MCS-KL	MCS-V-J	MCS-V-KL	DS-MBM	DS-V-MBM	DS-J	DS-KL	DS-V-J	DS-V-KL	MCI-KL	MCI-V-J	MCI-V-KL	MCI-J	MCI-MBM	MCI-V-MBM
Accuracy	0.92813	0.92188	0.91094	0.88594	0.89375	0.88318	0.84688	0.84063	0.83438	0.82031	0.825	0.8125	0.86094	0.85156	0.84844	0.81094	0.82969	0.81406	0.9125	0.93906	0.90156	0.95469	**0.98594**	0.96719
Precision	0.93881	0.92569	0.91174	0.88677	0.89424	0.88313	0.86134	0.84306	0.83472	0.82062	0.83109	0.81926	0.86931	0.86015	0.8514	0.82496	0.83514	0.83058	0.91314	0.93949	0.90224	0.95493	**0.9875**	0.96708
Recall	0.92869	0.92818	0.91267	0.88798	0.89619	0.88694	0.84623	0.8527	0.83613	0.824	0.83452	0.82807	0.86335	0.85378	0.85173	0.82066	0.83554	0.82127	0.91334	0.93926	0.90237	0.95498	**0.98627**	0.96775
F-Score	0.93372	0.92694	0.91221	0.88738	0.89521	0.88503	0.85372	0.84786	0.83542	0.8224	0.8328	0.82364	0.86632	0.85695	0.85157	0.8228	0.83534	0.8259	0.91324	0.93643	0.90231	0.95496	**0.98689**	0.96742
Jaccard index	0.87509	0.86066	0.83768	0.79793	0.81059	0.79432	0.74119	0.73177	0.71627	0.69788	0.7141	0.70624	0.76342	0.7483	0.74453	0.69843	0.71749	0.70222	0.83431	0.8857	0.82171	0.91371	**0.97377**	0.93622

TABLE 5. Comparison of validation indexes for different approaches by the updated DBSCAN clustering algorithm on ADL database.

Index	KDE-MBM	KDE-V-MBM	KDE-J	KDE-KL	KDE-V-J	KDE-V-KL	MCS-MBM	MCS-V-MBM	MCS-J	MCS-KL	MCS-V-J	MCS-V-KL	DS-MBM	DS-V-MBM	DS-J	DS-KL	DS-V-J	DS-V-KL	MCI-KL	MCI-V-J	MCI-V-KL	MCI-J	MCI-MBM	MCI-V-MBM
Accuracy	0.696	0.686	0.68588	0.67197	0.67793	0.66004	0.608	0.594	0.59046	0.57256	0.58449	0.57455	0.656	0.6506	0.64016	0.63022	0.63419	0.61829	0.6839	0.72167	0.67793	0.7336	**0.78**	0.76
Precision	0.702	0.69797	0.69961	0.6777	0.68215	0.66466	0.61679	0.60791	0.59524	0.57833	0.58896	0.57697	0.69212	0.65039	0.64879	0.63676	0.63977	0.62577	0.68652	0.72244	0.67863	0.73378	**0.80724**	0.76184
Recall	0.72682	0.68904	0.68969	0.67899	0.68936	0.66583	0.62714	0.62648	0.59502	0.57785	0.59125	0.58088	0.65878	0.66967	0.64663	0.64248	0.64338	0.62549	0.68782	0.7256	0.68253	0.73864	**0.78919**	0.78521
F-Score	0.71419	0.69348	0.69462	0.67834	0.68574	0.66524	0.62192	0.61706	0.59513	0.57809	0.5901	0.57892	0.67503	0.65989	0.64771	0.63961	0.64157	0.62563	0.68717	0.72401	0.68057	0.7362	**0.79811**	0.77335
Jaccard index	0.54004	0.52561	0.52409	0.50854	0.51453	0.4942	0.43914	0.42366	0.41877	0.40154	0.41332	0.40354	0.49649	0.48728	0.47157	0.46246	0.46606	0.45002	0.5247	0.56821	0.51462	0.58268	**0.64657**	0.62505

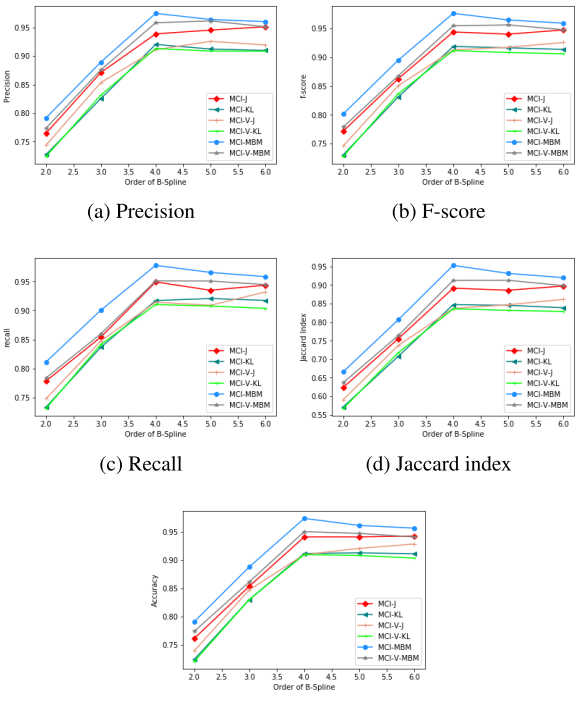


(a) Precision (b) F-score

(c) Recall (d) Jaccard index

(e) Accuracy

FIGURE 6. The performance measures in terms of validation indexes obtained by executing the updated *k*-medoids clustering algorithm on Japanese Vowels.



(a) Precision (b) F-score

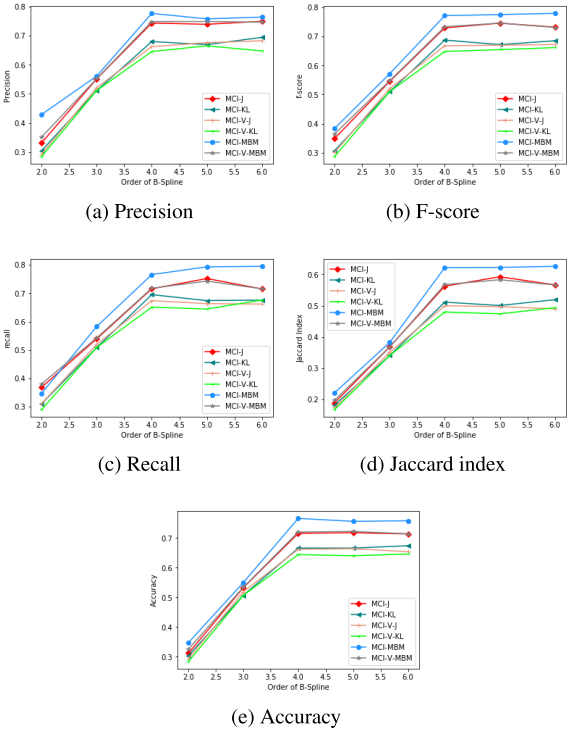
(c) Recall (d) Jaccard index

(e) Accuracy

FIGURE 7. The performance measures in terms of validation indexes obtained by executing the updated DBSCAN clustering algorithm on Japanese Vowels.

C. RESULTS AND COMPARISONS

A b-spline function integrates with MCI for modeling uncertain data objects [24]. The performance of MCI can be evaluated using a clustering algorithm. In the experiment, two clustering algorithms: modified *k*-medoids and DBSCAN are considered. However, the performance relies upon the order of a b-spline function and the correct order is obtained experimentally. Initially, the order is altered from 2 to 6 with

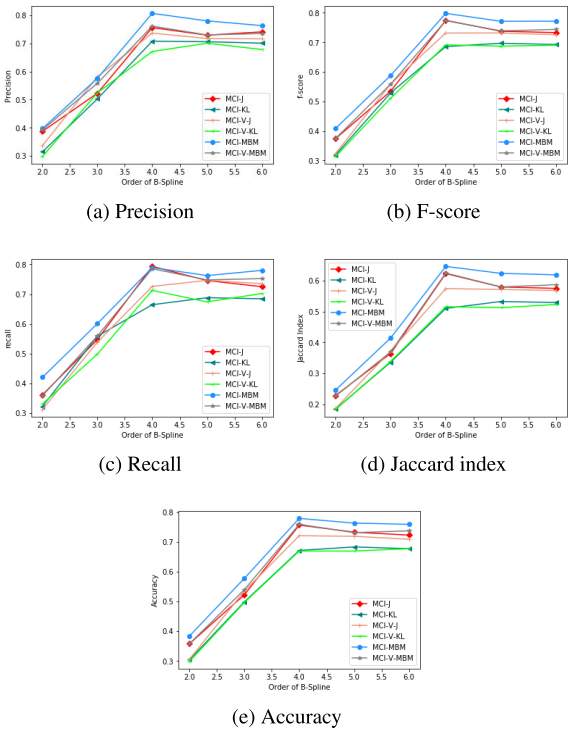


(a) Precision (b) F-score

(c) Recall (d) Jaccard index

(e) Accuracy

FIGURE 8. The performance measures in terms of validation indexes obtained by executing the updated k-medoids clustering algorithm on ADL.



(a) Precision (b) F-score

(c) Recall (d) Jaccard index

(e) Accuracy

FIGURE 9. The performance measures in terms of validation indexes obtained by executing the updated DBSCAN clustering algorithm on ADL.

an increment of size 1. This experiment is performed on three above-mentioned real databases. The performance of MCI is evaluated with the proposed similarity measure (MBM), which is labeled as MCI-MBM. Similarly, MCI with J-divergence is denoted as MCI-J. MCI with KL-divergence is marked by MCI-KL. MCI with MBM and majority voting scheme is tagged by MCI-V-MBM. MCI-J and the majority voting scheme is called MCI-V-J. MCI-KL and majority voting scheme is known as MCI-V-KL. Figs. 4 and 5 show the results of MCI-MBM, MCI-J, MCI-KL, MCI-V-MBM, MCI-V-J, and MCI-V-KL by executing the updated DBSCAN and k-medoids clustering algorithms on the weather data respectively using above discussed five validity indexes. It is clear from Figs. 4 and 5 that the MCI-MBM method obtained the highest value for accuracy, precision, recall, f-score, and Jaccard index with the b-spline function of order 4. These results also present the sensitivity of curves to variable order. These figures also prove the effectiveness of the proposed similarity measure i.e. MBM over J-divergence and KL-divergence.

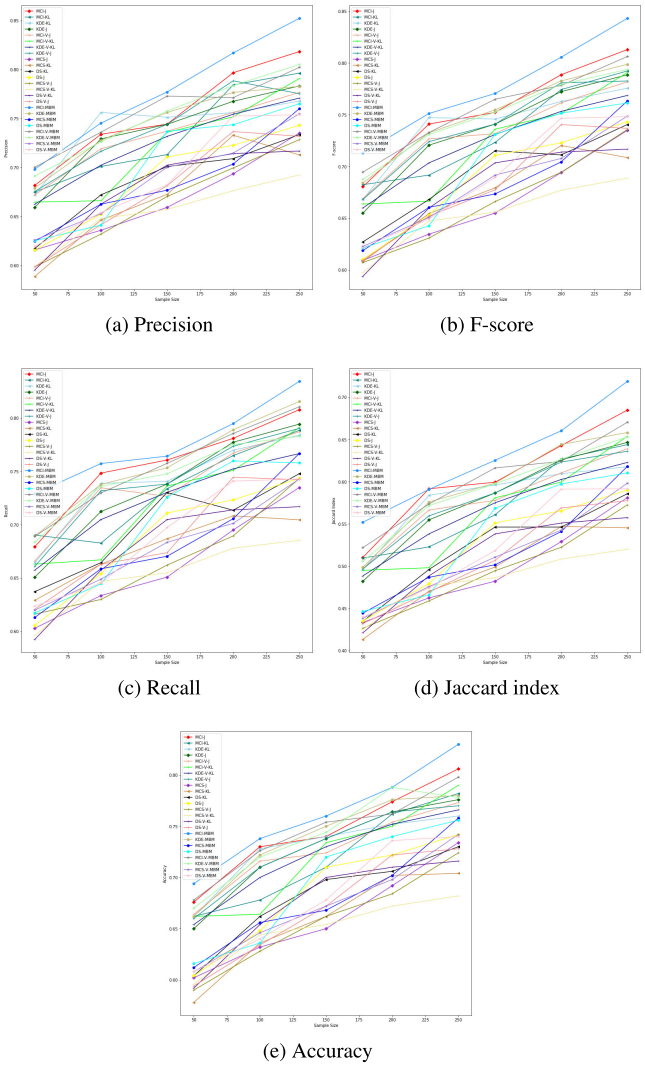
Similarly, Figs. 6 and 7 show the results obtained by the modified k-medoids and DBSCAN clustering algorithms respectively on the Japanese vowels database. These two figures also illustrate the effectiveness of the MCI-MBM method over existing approaches. Moreover, all the methods demonstrate similar trends as Figs. 4 and 5 with the b-spline function of order 4.

Figs. 8 and 9 show the results obtained by the updated k-medoids and DBSCAN clustering algorithms respectively on the ADL database. These two figures also support the previous conclusion. In other words, the order of the b-spline function is 4 which helps to achieve the highest accuracy, precision, recall, f-score, and Jaccard index.

In the second experiment, a comparative analysis of MCI-MBM with three existing approaches is discussed, where three uncertain data objects are modeled using KDE [14], [15], DS [19], [20], and MCS schemes [21]–[23]. However, the updated DBSCAN and k-medoids algorithms are used to divide uncertain data objects. This experiment is conducted on three real databases mentioned in Section 3.1. The KDE with MBM is marked by KDE-MBM. Similarly, KDE with J-divergence and KL-divergence are tagged by KDE-J and KDE-KL respectively. The KDE with MBM and majority voting technique is labeled by KDE-V-MBM. The KDE with J-divergence, KL-divergence, and majority voting approach are called KDE-V-J and KDE-V-KL. Similarly, we call DS-MBM: DS with MBM, DS-J: DS with J-divergence, DS-KL: DS with KL-divergence, DS-V-MBM: DS with MBM and majority voting technique, DS-V-J: DS with J-divergence and majority voting scheme, DS-V-KL: DS with KL-divergence and majority voting approach, MCS-MBM: MCS with MBM, MCS-J: MCS with J-divergence, MCS-KL: MCS with KL-divergence, MCS-V-MBM: MCS with MBM and majority voting technique, MCS-V-J: MCS with J-divergence and majority voting scheme, and MCS-V-KL: MCS with KL-divergence and majority voting approach. These methods are merged to the updated k-medoids and DBSCAN

TABLE 6. Comparison of validation indexes for different approaches by the updated *k*-medoids clustering algorithm on ADL database.

Index	KDE-MBM	KDE-V-MBM	KDE-J	KDE-KL	KDE-V-J	KDE-V-KL	MCS-MBM	MCS-V-MBM	MCS-J	MCS-KL	MCS-V-J	MCS-V-KL	DS-MBM	DS-V-MBM	DS-J	DS-KL	DS-V-J	DS-V-KL	MCI-KL	MCI-V-J	MCI-V-KL	MCI-J	MCI-MBM	MCI-V-MBM
Accuracy	0.682	0.67665	0.67396	0.66203	0.66402	0.6501	0.63	0.614	0.6004	0.57853	0.59642	0.5666	0.664	0.658	0.6501	0.63419	0.64215	0.63022	0.67396	0.66402	0.64414	0.71769	**0.766**	0.72
Precision	0.68985	0.69278	0.68793	0.66695	0.66929	0.65407	0.65379	0.62399	0.60511	0.58458	0.60124	0.56935	0.67786	0.68866	0.6594	0.64046	0.65438	0.63613	0.67489	0.66783	0.64811	0.71877	**0.77637**	0.74821
Recall	0.69209	0.67942	0.67753	0.66702	0.66954	0.65692	0.64634	0.63894	0.60503	0.58404	0.6037	0.57266	0.66406	0.66339	0.65736	0.64825	0.64962	0.64001	0.67897	0.66756	0.65304	0.72107	**0.76579**	0.71778
F-Score	0.69096	0.68604	0.68269	0.66699	0.66941	0.65549	0.65004	0.63138	0.60507	0.58431	0.60247	0.571	0.67089	0.67579	0.65838	0.64433	0.65199	0.63806	0.67693	0.6677	0.64531	0.71992	**0.77104**	0.73268
Jaccard index	0.52579	0.51969	0.51073	0.497	0.49918	0.48218	0.46381	0.45328	0.42893	0.40729	0.42497	0.39612	0.50203	0.49494	0.48142	0.46677	0.47512	0.463	0.51037	0.50175	0.48067	0.56329	**0.62221**	0.5675

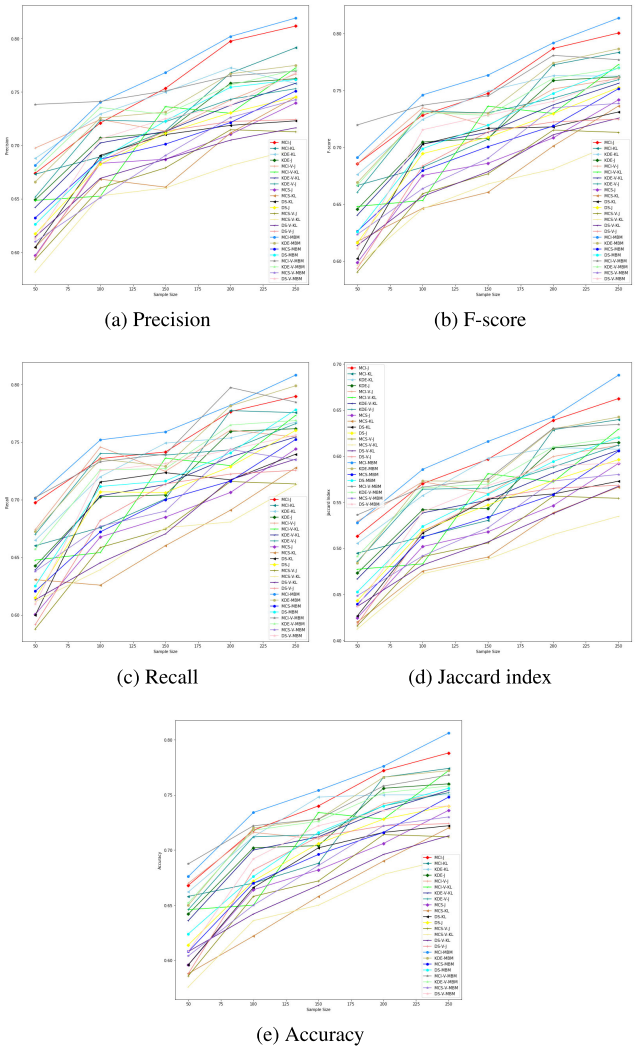


(a) Precision (b) F-score

(c) Recall (d) Jaccard index

(e) Accuracy

FIGURE 10. The performance measures in terms of validation indexes obtained by executing the updated *k*-medoids clustering algorithm on synthetic data.



(a) Precision (b) F-score

(c) Recall (d) Jaccard index

(e) Accuracy

FIGURE 11. The performance measures in terms of validation indexes obtained by executing the updated DBSCAN clustering algorithm on synthetic data.

algorithms to cluster uncertain data objects. All the obtained results are noted in Tables 1 to 6. Table 1 displays the outcomes achieved by different approaches using the modified DBSCAN on weather data. The obtained results using *k*-medoids on weather data are reported in Table 2. Tables 3 and 4 illustrate the outcomes achieved by DBSCAN and *k*-medoids algorithms on the Japanese vowels database. The obtained outcomes on the ADL database using DBSCAN and *k*-medoids algorithms are reported in Tables 5 and 6 respectively. It is observed from Tables 1 to 6 that the proposed method i.e. MCI-MBM outperforms existing approaches.

In the final experiment, all the above-said methods are applied to the synthesized database by changing the sample size range 50-250 with an increment of 50 samples. Fig. 10 illustrates all the obtained outcomes using different approaches in the case of the updated *k*-medoids. Similarly, Fig. 11 shows the outcomes of the different approaches by the modified DBSCAN clustering algorithm. So, it may be inferred from all the results that the demonstration of the proposed methods surpasses other existing techniques.

Then non-parametric Wilcoxon's Rank-Sum is also performed to compare the proposed technique over existing

TABLE 7. Comparison of p-values from precision for wilcoxon Rank-Sum test of MCI-MBM by applying modified DBSCAN clustering algorithms.

Dataset	MCI-V-MBM	MCI-J	MCI-KL	MCI-V-J	MCI-V-KL	KDE-MBM	KDE-V-MBM	KDE-J	KDE-KL	KDE-V-J	KDE-V-KL	MCS-MBM	MCS-V-MBM	MCS-J	MCS-KL	MCS-V-J	MCS-V-KL	DS-MBM	DS-V-MBM	DS-J	DS-KL	DS-V-J	DS-V-KL
Weather Data	0.0156	0.0101	0.0181	0.0637	0.0101	0.0101	0.0101	0.0661	0.7929	0.1278	0.0181	0.0101	0.0101	0.0101	0.0101	0.0240	0.0101	0.0453	0.0101	0.0101	0.0101	0.0101	0.0101
Japanese Vowels	0.125	0.0101	0.0520	0.0793	0.0101	0.0637	0.0101	0.4948	0.0101	0.1250	0.0101	0.7132	0.0101	0.0520	0.0101	0.0101	0.0156	0.0101	0.0101	0.0101	0.0453	0.0101	0.0181
ADL	0.0313	0.0453	0.0313	0.0875	0.0101	0.0564	0.0831	0.6365	0.0313	0.1278	0.0313	0.0156	0.0101	0.0101	0.0406	0.0101	0.0156	0.0101	0.0156	0.0101	0.0101	0.0372	0.0101
Synthesized Data	0.01562	0.0101	0.0661	0.8748	0.0101	0.0313	0.024	0.2272	0.1893	0.0101	0.0713	0.6365	0.0406	0.0101	0.0101	0.0136	0.0101	0.0156	0.0101	0.0101	0.0101	0.0101	0.0156

TABLE 8. Comparison of p-values from precision for wilcoxon Rank-Sum test of MCI-MBM by applying modified *k*-medoids clustering algorithms.

Dataset	MCI-V-MBM	MCI-J	MCI-KL	MCI-V-J	MCI-V-KL	KDE-MBM	KDE-V-MBM	KDE-J	KDE-KL	KDE-V-J	KDE-V-KL	MCS-MBM	MCS-V-MBM	MCS-J	MCS-KL	MCS-V-J	MCS-V-KL	DS-MBM	DS-V-MBM	DS-J	DS-KL	DS-V-J	DS-V-KL
Weather Data	0.0313	0.0227	0.0313	0.2272	0.0101	0.0661	0.7929	0.0156	0.0101	0.0156	0.0101	0.0240	0.0101	0.0101	0.0101	0.2272	0.0101	0.0101	0.0156	0.0101	0.024	0.0101	0.0101
Japanese Vowels	0.0156	0.0793	0.0101	0.0713	0.0101	0.4948	0.0101	0.0637	0.0101	0.1250	0.0101	0.0313	0.0101	0.2272	0.0101	0.0661	0.0101	0.0101	0.0453	0.0101	0.1250	0.0101	0.027
ADL	0.0101	0.0161	0.0313	0.0156	0.0101	0.6365	0.0313	0.0564	0.0431	0.0661	0.0661	0.0156	0.0101	0.0101	0.1036	0.0313	0.0101	0.0101	0.1250	0.0101	0.0240	0.0101	0.0101
Synthesized Data	0.0161	0.0875	0.0101	0.0637	0.0406	0.02272	0.0189	0.0313	0.0240	0.0101	0.0406	0.0136	0.0101	0.0240	0.0101	0.0101	0.0240	0.0101	0.024	0.0101	0.0240	0.0101	0.0156

techniques based on the p-values achieved by the accuracy. Tables 7 and 8 report the obtained p-values. Most of the obtained p-values support eliminating the null hypothesis at a 5% level. In other words, available significant evidence based on data states the superiority of the proposed method as compared to that of state-of-the-art methods in this work. Moreover, Tables 7 and 8 also show that the statistical experimental results for accuracy validation index are not significant in some cases, where p-values are higher than 0.05. However, Tables 1 to 6 also show the superiority of the proposed method based on the values of accuracy, precision, recall, f-score, and Jaccard index.

IV. CONCLUSION

In this study, uncertain data objects clustering is addressed based on their distributions. Three measures of closeness: KL-divergence, J-divergence, as well as a new devised measure are combined with *k*-medoids and DBSCAN clustering algorithms. Some of the important properties of the proposed similarity measure are discussed. The b-spline function is one of the components of MCI meaning that the performance of MCI depends on the order of the b-spline function. Generally, determining the correct order of the b-spline function is a difficult task. Thus, we conduct an empirical analysis to get the value of the order. Three existing modeling schemes: KDE, DS, and MCS are considered to compare with MCI. All the experiments are performed on three real databases: Japanese vowels, weather, and ADL, and one synthetic. It is clear from the experimental results that MCI performs well when the order of the b-spline function is 4 as compared to other orders. Moreover, the MCI-MBM method is superior to existing approaches. As a future work, we would like to develop an algorithm for finding out the optimum order of the b-spline function. We would also like to merge the proposed similarity measure with conventional clustering algorithms.

Malaysia for the completion of the research.

REFERENCES

[1] K. K. Sharma and A. Seal, "Spectral embedded generalized mean based k-nearest neighbors clustering with S-distance," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114326.

[2] K. K. Sharma, A. Seal, E. Herrera-Viedma, and O. Krejcar, "An enhanced spectral clustering algorithm with S-distance," *Symmetry*, vol. 13, no. 4, p. 596, Apr. 2021.

[3] K. Zheng, Z. Huang, A. Zhou, and X. Zhou, "Discovering the most influential sites over uncertain data: A rank-based approach," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 12, pp. 2156–2169, Dec. 2012.

[4] A. Seal, A. Karlekar, O. Krejcar, and E. Herrera-Viedma, "Performance and convergence analysis of modified C-means using Jeffreys-divergence for clustering," *Int. J. Interact. Multimedia Artif. Intell.*, pp. 1–9, 2021, doi: 10.9781/ijimai.2021.04.009.

[5] L. Wang, B. Yang, Y. Chen, X. Zhang, and J. Orchard, "Improving neural-network classifiers using nearest neighbor partitioning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2255–2267, Oct. 2017.

[6] K. K. Sharma and A. Seal, "Clustering analysis using an adaptive fused distance," *Eng. Appl. Artif. Intell.*, vol. 96, Nov. 2020, Art. no. 103928.

[7] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: An example in clustering location data," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Singapore: Springer, 2006, pp. 199–204.

[8] S. D. Lee, B. Kao, and R. Cheng, "Reducing UK-means to K-means," in *Proc. 7th IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Oct. 2007, pp. 483–488.

[9] F. Gullo, G. Ponti, and A. Tagarelli, "Clustering uncertain data via k-medoids," in *Proc. Int. Conf. Scalable Uncertainty Manage.* Naples, Italy: Springer, 2008, pp. 229–242.

[10] G. Cormode and A. McGregor, "Approximation algorithms for clustering uncertain data," in *Proc. 27th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (PODS)*, 2008, pp. 191–200.

[11] F. Gullo, G. Ponti, and A. Tagarelli, "Minimizing the variance of cluster mixture models for clustering uncertain objects," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 839–844.

[12] F. Gullo and A. Tagarelli, "Uncertain centroid based partitional clustering of uncertain data," *Proc. VLDB Endowment*, vol. 5, no. 7, pp. 610–621, Mar. 2012.

[13] B. Jiang, J. Pei, Y. Tao, and X. Lin, "Clustering uncertain data based on probability distribution similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 751–763, Apr. 2011.

[14] B. Jiang, J. Pei, Y. Tao, and X. Lin, "Clustering uncertain data based on probability distribution similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 751–763, Apr. 2013.

[15] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, vol. 26. Boca Raton, FL, USA: CRC Press, 1986.

[16] H. Kriegel and M. Pfeifle, "Hierarchical density-based clustering of uncertain data," in *Proc. 5th IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2005, pp. 1–4.

[17] H.-P. Kriegel and M. Pfeifle, "Density-based clustering of uncertain data," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: ACM, 2005, pp. 672–677.

[18] J. Galindo, *Fuzzy Databases: Modeling, Design and Implementation: Modeling, Design and Implementation*. Hershey, PA, USA: IGI Global, 2005.

[19] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," in *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Berlin, Germany: Springer, 2008, pp. 57–72.

[20] S. K. Lee, "Imprecise and uncertain information in databases: An evidential approach," in *Proc. 8th Int. Conf. Data Eng.*, 1992, pp. 614–621.

[21] J. Giraud, M. Lindsay, V. Ogarko, M. Jessell, R. Martin, and E. Pakyuz-Charrier, "Integration of geoscientific uncertainty into geophysical inversion by means of local gradient regularization," *Solid Earth*, vol. 10, no. 1, pp. 193–210, Jan. 2019.
[22] E. Pakyuz-Charrier, J. Giraud, V. Ogarko, M. Lindsay, and M. Jessell, "Drillhole uncertainty propagation for three-dimensional geological modeling using Monte Carlo," *Tectonophysics*, vols. 747–748, pp. 16–39, Nov. 2018.
[23] E. Pakyuz-Charrier, M. Lindsay, V. Ogarko, J. Giraud, and M. Jessell, "Monte Carlo simulation for uncertainty estimation on structural data in implicit 3-D geological modeling, a guide for disturbance distribution selection and parameterization," *Solid Earth*, vol. 9, no. 2, pp. 385–402, Apr. 2018.
[24] K. K. Sharma and A. Seal, "Modeling uncertain data using Monte Carlo integration method for clustering," *Expert Syst. Appl.*, vol. 137, pp. 100–116, Dec. 2019.
[25] K.-T. Liao and C.-M. Liu, "An effective clustering mechanism for uncertain data mining using centroid boundary in UKmeans," in *Proc. Int. Comput. Symp. (ICS)*, Dec. 2016, pp. 300–305.
[26] F. Xiao, "Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy," *Inf. Fusion*, vol. 46, pp. 23–32, Mar. 2019.
[27] F. Xiao, "A new divergence measure for belief functions in D–S evidence theory for multisensor data fusion," *Inf. Sci.*, vol. 514, pp. 462–483, Apr. 2020.
[28] D. Cavaliere, S. Senatore, and V. Loia, "Context-aware profiling of concepts from a semantic topological space," *Knowl.-Based Syst.*, vol. 130, pp. 102–115, Aug. 2017.
[29] Z. Kang, C. Peng, Q. Cheng, X. Liu, X. Peng, Z. Xu, and L. Tian, "Structured graph learning for clustering and semi-supervised classification," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107627.
[30] F. Perez-Cruz, "Kullback-Leibler divergence estimation of continuous distributions," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1666–1670.
[31] K. K. Sharma and A. Seal, "Multi-view spectral clustering for uncertain objects," *Inf. Sci.*, vol. 547, pp. 723–745, Feb. 2021.
[32] K. K. Sharma and A. Seal, "Outlier-robust multi-view clustering for uncertain data," *Knowl.-Based Syst.*, vol. 211, Jan. 2021, Art. no. 106567.
[33] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Chelmsford, MA, USA: Courier, 1998.
[34] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive bayes for text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2508–2521, Sep. 2016.
[35] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, no. 1, pp. 99–109, 1943.
[36] A. T. Albaham and N. Salim, "Adapting voting techniques for online forum thread retrieval," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl.* Springer, 2012, pp. 439–448.
[37] A. S. Belenky and D. C. King, "A mathematical model for estimating the potential margin of state undecided voters for a candidate in a US federal election," *Math. Comput. Model.*, vol. 45, nos. 5–6, pp. 585–593, Mar. 2007.
[38] T. K. Paul and H. Iba, "Prediction of cancer class with majority voting genetic programming classifier using gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 6, no. 2, pp. 353–367, Apr. 2009.
[39] M. Matsumoto and T. Nishimura, "Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator," *ACM Trans. Model. Comput. Simul.*, vol. 8, no. 1, pp. 3–30, Jan. 1998.
[40] L. Kaufman and P. Rousseeuw, *Clustering by Means of Medoids. Statistical Data Analysis Based on the L1 Norm*, Y. Dodge, Ed. 1987, pp. 405–416.
[41] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 1, pp. 81–87, Jan. 1984.
[42] A. Karami and R. Johansson, "Choosing DBSCAN parameters automatically using differential evolution," *Int. J. Comput. Appl.*, vol. 91, no. 7, pp. 1–11, Apr. 2014.
[43] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, p. 19, Aug. 2017.
[44] N. Rahmah and I. S. Sitanggang, "Determination of optimal epsilon (eps) value on DBSCAN algorithm to clustering data on peatland hotspots in sumatra," in *Proc. IOP Conf., Earth Environ. Sci.*, vol. 31, 2016, Art. no. 012012.
[45] *National Weather Service*, CPC Global Summary of Day/Month Observations, 1979-Continuing, NOAA U.S. Dept. Commerce Climate Predict. Center, Nat. Centers Environ. Predict., College Park, MD, USA, 1987.
[46] M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel, "World map of the Köppen-Geiger climate classification updated," *Meteorol. Zeitschrift*, vol. 15, no. 3, pp. 259–263, 2006.
[47] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[48] A. Karlekar and A. Seal, "SoyNet: Soybean leaf diseases classification," *Comput. Electron. Agricult.*, vol. 172, May 2020, Art. no. 105342.
[49] A. Seal, A. Karlekar, O. Krejcar, and C. Gonzalo-Martin, "Fuzzy c-means clustering using Jeffreys-divergence based similarity measure," *Appl. Soft Comput.*, vol. 88, Mar. 2020, Art. no. 106016.
[50] A. Karlekar, A. Seal, O. Krejcar, and C. Gonzalo-Martin, "Fuzzy K-means using non-linear S-distance," *IEEE Access*, vol. 7, pp. 55121–55131, 2019.

KRISHNA KUMAR SHARMA received the M.Tech. degree in information technology from IIIT Allahabad, India, in 2011. He is currently pursuing the Ph.D. degree with the Computer Science and Engineering Department, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, India. He is currently an Assistant Professor with the Computer Science and Informatics Department, University of Kota, Kota, India. His current research interest includes pattern recognition.

AYAN SEAL (Senior Member, IEEE) received the Ph.D. degree in engineering from Jadavpur University, 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, India. He has visited the Universidad Politecnica de Madrid, Spain, as a Visiting Research Scholar. He has authored or coauthored 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. He was a recipient of several awards. He has received Sir Visvesvaraya Young Faculty Research Fellowship from Media Laboratory Asia, Ministry of Electronics and Information Technology, Government of India.

ANIS YAZIDI (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees from the University of Agder, Grimstad, Norway, in 2008 and 2012, respectively. He was a Researcher with Teknova AS, Grimstad. From 2014 to 2019, he was an Associate Professor with the Department of Computer Science, Oslo Metropolitan University, Oslo, Norway, where he is currently a Full Professor, leading with the Research Group in Applied Artificial Intelligence. He is also a Professor II with the Norwegian University of Science and Technology (NTNU), Trondheim, Norway, and a Senior Researcher at Oslo University Hospital, Oslo. His current research interests include machine learning, learning automata, stochastic optimization, and autonomous computing.

ALI SELAMAT (Member, IEEE) is currently a Full Professor with Universiti Teknologi Malaysia (UTM), Malaysia. He has been the Dean of the Malaysia Japan International Institute of Technology (MJIIT), UTM, since 2018. An academic institution established under the cooperation of the Japanese International Cooperation Agency (JICA) and the Ministry of Education Malaysia (MOE) to provide the Japanese style of education in Malaysia. He is also a Professor with the Software Engineering Department, Faculty of Computing, UTM. He has published more than 60 IF research articles. His H-index is 20, and his number of citations in WoS is over 800. His research interests include software engineering, software process improvement, software agents, web engineering, information retrievals, pattern recognition, genetic algorithms, neural networks, soft computing, computational collective intelligence, strategic management, key performance indicator, and knowledge management. He has been serving as the Chair of the Computer Society Malaysia, since 2018. He is on the Editorial Board of the *Knowledge-Based Systems* journal (Elsevier).

ONDREJ KREJCAR received the Ph.D. degree in technical cybernetics from the Technical University of Ostrava, Czech Republic. He is a Full Professor in systems engineering and informatics with the University of Hradec Kralove, Czech Republic. He has been a Vice-Rector of science and creative activities with the University of Hradec Kralove (UHK), since June 2020. He is currently the Director of the Center for Basic and Applied Research, UHK. From 2016 to 2020, he was the Vice-Dean of Science and Research with the Faculty of Informatics and Management, UHK. At UHK, he is a guarantee of the Ph.D. 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, biomedicine image analysis, as well as biotelemetric system architecture portable device architecture, wireless biosensors, and development of applications for mobile devices with use of remote or embedded biomedical sensors. 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, in 2019 by Publons. He has been the Vice-Leader and the 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 the Chairman of the Program Committee of the KAPPA Program and Technological Agency of the Czech Republic as a Regulator of the EEA/Norwegian Financial Mechanism, Czech Republic, from 2019 to 2024. Since 2020, he has been the Chairman of the Panel 1 in computer, physical, and chemical sciences of the ZETA Program, Technological Agency of the Czech Republic. From 2014 to 2019, he was the Deputy Chairman of the Panel 7 in processing industry, robotics, and electrical engineering of the Epsilon Program, Technological Agency of the Czech Republic. His H-index is 19, with more than 1250 citations received in the Web of Science. He is currently on the Editorial Board of the *Sensors* IF journal (MDPI) (Q1/Q2 at JCR) and several other ESCI indexed journals.