sensors

MDPI

Article

Cluster Validity Index for Uncertain Data Based on a Probabilistic Distance Measure in Feature Space

Changwan Ko [1], Jaeseung Baek [2,3], Behnam Tavakkol [4] and Young-Seon Jeong [1,5,*]

[1] Department of Industrial Engineering, Chonnam National University, Gwangju 61186, Republic of Korea; kcw7536@gmail.com
[2] College of Business, Northern Michigan University, Marquette, MI 49855, USA; jbaek@nmu.edu
[3] Department of Industrial & Systems Engineering, Rutgers University, Piscataway, NJ 08854, USA
[4] School of Business, Stockton University, Galloway, NJ 08205, USA; behnam.tavakkol@stockton.edu
[5] Interdisciplinary Program of Arts and Design Technology, Chonnam National University, Gwangju 61186, Republic of Korea
* Correspondence: young.jeong@jnu.ac.kr; Tel.: +82-62-530-1790

Abstract: Cluster validity indices (CVIs) for evaluating the result of the optimal number of clusters are critical measures in clustering problems. Most CVIs are designed for typical data-type objects called certain data objects. Certain data objects only have a singular value and include no uncertainty, so they are assumed to be information-abundant in the real world. In this study, new CVIs for uncertain data, based on kernel probabilistic distance measures to calculate the distance between two distributions in feature space, are proposed for uncertain clusters with arbitrary shapes, sub-clusters, and noise in objects. By transforming original uncertain data into kernel spaces, the proposed CVI accurately measures the compactness and separability of a cluster for arbitrary cluster shapes and is robust to noise and outliers in a cluster. The proposed CVI was evaluated for diverse types of simulated and real-life uncertain objects, confirming that the proposed validity indexes in feature space outperform the pre-existing ones in the original space.

Keywords: uncertain data; cluster validity index; kernel probabilistic distance; feature space

Citation: Ko, C.; Baek, J.; Tavakkol, B.; Jeong, Y.-S. Cluster Validity Index for Uncertain Data Based on a Probabilistic Distance Measure in Feature Space. *Sensors* **2023**, *23*, 3708. https://doi.org/10.3390/s23073708

Academic Editors: Qiong Wang, Teng Huang and Yan Pang

Received: 11 March 2023
Revised: 29 March 2023
Accepted: 29 March 2023
Published: 3 April 2023

1. Introduction

The purpose of clustering is to partition objects into groups with criteria such that the similarity within the groups and the dissimilarity among different groups should be maximized [1,2]. Although clustering methods have been widely used in many applications, most clustering algorithms do not provide the optimal number of clusters. Partitional-based clustering algorithms such as K-means clustering [3] must preset the number of clusters [4]. As cluster information is rarely known in the real world, it is crucial to evaluate the clustering results depending on the different numbers of clusters. Although many clustering methods exist for diverse applications, such as pattern recognition [5], semiconductor manufacturing [6], and healthcare [7], they have been developed primarily for only certain data or fixed values. However, the embedded uncertainty of data is essential in many applications. For instance, a patient's blood pressure may not be consistent because of environmental conditions and instrument errors. Furthermore, measurement values are continuously changing because of the positions of instrumentation devices or workers' conditions. Aside from these examples, data randomness, missing data, delayed updates, and worker fatigue are other factors of data uncertainty [8,9].

Uncertain data are assumed to be prevalent information in the real world, e.g., measurement errors and environmental conditions. The uncertainty of uncertain data can be expressed by probability density functions (PDFs). Figure 1 illustrates two uncertain data, each distributed by a PDF. The standard method of converting uncertain data is to transform a summary statistic (e.g., mean or median) into certain data. However, these statistics

could lose extra information of uncertainty that is significant to capture the uncertainty information of uncertain objects.

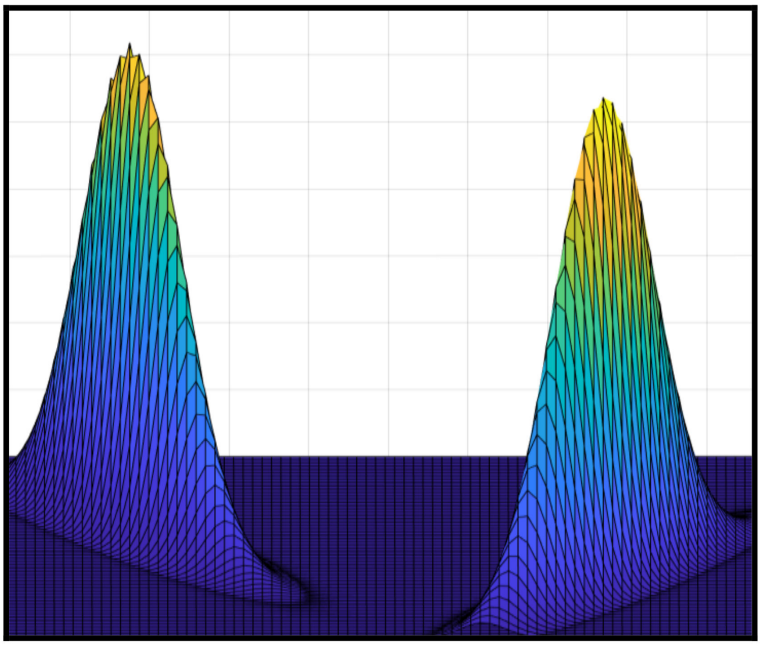


Figure 1. Two uncertain datasets, each expressed by a PDF.

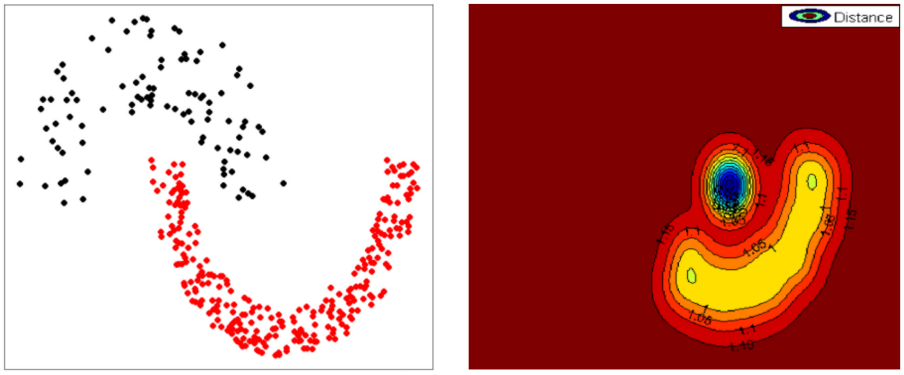
Cluster validity indices (CVIs), which are indicators for validating the quality of clustering algorithms, have been widely used to determine the correct number of clusters for the given data. As the CVIs only use input data information, they must be used according to the characteristics of the data. The two components of a CVI are compactness and separability measures. The former refers to an intra-cluster distance, and the latter represents an inter-cluster distance. Most CVIs indicate that a good partition produces a small compactness value and a high separability value. However, the existing CVIs are vulnerable to validating cluster results when the shapes of the clusters are not spherical clusters [10,11].

For certain data, several CVIs, such as the Dunn [12], Calinski–Harabasz [13], Davies–Bouldin [14], and Xie–Beni [15] indices, have been proposed based on combinations of compactness and separability measures. However, most of the existing CVIs have been developed for certain data. There have been few studies on uncertain data. Moreover, relatively new CVIs are also being designed to incorporate mathematical theories into pre-existing CVIs, such as the K-nearest neighbor algorithm, which is used to compute compactness and separation by taking into account shared/non-shared data pairs [10], and principal component analysis, which is used to capture the geometry of the clusters [16]; or to develop clustering algorithms to cluster more well-separated clusters [1].

To apply uncertain data to the existing CVIs' formulas, they should be changed to calculate distance measures of compactness and separability. In a study of uncertain CVIs, Tavakkol et al. [17] proposed CVIs for uncertain data to calculate the distance between two uncertain objects using probabilistic distance measures in the original space. However, it leads to sensitivity to arbitrary shapes of clusters, sub-clusters, and outliers because of the clusters shape that may cause inaccurate compactness and separability [11].

Consequently, this study proposes new uncertain CVIs for uncertain data objects based on kernel probabilistic distance measures in feature space. The proposed CVIs for uncertain objects are designed to adapt the kernel-based Bhattacharyya probabilistic distance in kernel spaces. In kernel space, the proposed CVIs produce accurate compactness and separability for the arbitrary shapes of clusters by transforming them into elliptical shapes in feature space. Figure 2 illustrates that the ambiguous shape of a dataset in the original

space is transformed into a relatively elliptical, circular shape in feature space; thus, the kernel transformation can improve performance in calculating accurate compactness and separability. Furthermore, the proposed approaches could be robust to noise and outliers in a cluster. The superior performance of the proposed CVIs was evaluated through diverse experiments, including simulated and real-life datasets.



(**a**) (**b**)

Figure 2. Visualization of kernel transformation: (**a**) asymmetry shape in original space; (**b**) transformed shape in feature space.

This paper is organized as follows. Section 2 reviews the previous studies on CVIs. New CVIs for uncertain data based on a kernel probabilistic distance measure are proposed in Section 3. After the extensive experiments are presented in Section 4, the conclusions and future studies are provided in Section 5.

2. Related Work

2.1. CVI for Certain Data

In the past few decades, many CVIs have been developed to determine the optimal number of clusters. Most CVIs focus on calculating compactness and separability measures. The combination of the two measures is composed of a ratio-type or summation-type index. This section presents several popular CVIs that have been evaluated in many applications.

The Dunn (DU) index [12]:

$$DU_K = \frac{min_{i,j=1,\cdots,K,\ i\neq j}\left\{min_{x\in C_i,\ y\in C_j} d(x,y)\right\}}{max_{i=1,\cdots,K}\left\{max_{x,y\in C_i} d(x,y)\right\}}. \quad (1)$$

Compactness and separability are computed using the maximum diameter among all clusters and the minimum pair-wise distance between objects in different clusters. The DU index is integrated by the ratio type of separability to compactness. Thus, the maximum value of the DU index is the optimal number of clusters (max. S/C).

Calinski–Harabasz (CH) index [13]:

$$CH_K = \frac{\sum_{i=1}^{K} n_i \cdot d(z_i \cdot z_{tot})^2}{K-1} \cdot \frac{n-K}{\sum_{1=1}^{K}\sum_{x\in C_i} d(x,z_i)^2} \quad (2)$$

The CH is composed of the ratio type of separability and compactness like the DU index. z_{tot} is the centroid of the entire dataset. Compactness and separability are computed using within- and between-cluster sums of squares. Thus, the maximum value for CH is the optimum partition (max. S/C).

The Davies–Bouldin (DB) index [14]:

$$DB_K = \frac{1}{K}\max_{i=1,\ \ldots,K,\ i\neq j}\left\{\left(\sqrt{\frac{1}{n_i}\sum_{x\in C_i} d(x,z_i)^2} + \sqrt{\frac{1}{n_j}\sum_{y\in C_j} d(y,z_j)^2}\right)/d(z_i,z_j)\right\} \quad (3)$$

where z_i and z_j are the centroids of each cluster. Compactness and separability are calculated using the sum of mean squares of individual clusters, unlike the DU index, which considers the compactness and separability of the total cluster. Compactness is the computed sum of the pair-wise distances between different clusters; separability is calculated differently for each cluster. The DB index is comprised of the ratio types of compactness and separability. Therefore, the minimum value of DB is the optimum partition (min. C/S).

The pre-existing CVIs are sensitive to sub-clusters, arbitrary shapes, and noise in clusters for the compactness measure [18]. This study overcomes those drawbacks by conducting a spatial transformation from the original space into feature space using a kernel function that correctly measures cluster compactness and separability.

2.2. CVI for Uncertain Data

Most CVIs have focused on certain data or fixed values [19]. Certain data do not have uncertainty caused by several factors and environments such as sensor measurement error, repeated measurements by workers, or equipment operating environments. Uncertain data objects come in two possible forms: (1) multiple points for each object and (2) a PDF for each object, either given or obtained by fitting the multiple points [20]. Several studies related to clustering uncertain data have been conducted. However, CVIs for uncertain data have rarely been used. The CVIs are crucial criteria for validating the results of clusters [21,22] to find the appropriate number of clusters. Therefore, the study of CVIs for uncertain data is necessary.

In this study, the proposed CVIs use kernel probabilistic distance measures to compute the distance between two uncertain data objects. There are many popular probabilistic distance measures, such as Bhattacharyya distance [23], Wasserstein distance, and Kullback–Leibler divergence [24]. This study uses the Bhattacharyya distance measure. The Bhattacharyya distance measure is one of the widely used probabilistic distance measures and has been generally used in diverse applications.

The Bhattacharyya distance between two probability distributions can be calculated in discrete and continuous cases. Let p and q be the continuous probability distributions over the same space. The definition of the Bhattacharyya distance for a continuous case in original space can be described as follows:

$$PD_{Bhatt}(p,q) = -\ln\left\{ \int_x \sqrt{p(x)q(x)}dx \right\} \tag{4}$$

There are closed-form solutions for many probabilistic distance measures, including the Bhattacharyya distance, for cases where uncertain data objects are modeled with multivariate normal distributions. As probabilistic distance measures can capture the distance between PDFs, they can also be used to capture the distance between uncertain data objects [25]. The Bhattacharyya distance is a special case of Chernoff distance with parameters $\alpha_1 = \alpha_2 = 1/2$, and the closed-from of Bhattacharyya distance for multivariate normal PDFs is defined in Equation (5):

$$PD_{Bhatt}(p,q) = \frac{1}{8}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)'(\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_q)^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q) + \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_q|}{2(|\boldsymbol{\Sigma}_p|+|\boldsymbol{\Sigma}_q|)^{\frac{1}{2}}}\right) \tag{5}$$

where $\boldsymbol{\mu}_p$ and $\boldsymbol{\mu}_q$ are means, and $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\Sigma}_q$ are covariance matrices of $P \sim MVN(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and $Q \sim MVN(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$.

This study models the Bhattacharyya distance between two uncertain data objects in kernel space. We can compute the probabilistic distance between two uncertain data objects in feature space using a kernel function.

3. Proposed CVIs for Uncertain Data

3.1. Kernel Probabilistic Distance Measure in Feature Space

Computing the probabilistic distance is a nontrivial problem. We can compute the Bhattacharyya distance in feature space by referring to several steps developed by Zhou and Chellappa [26]. In capturing the probabilistic distance, suppose that $\boldsymbol{x}_1 = \{x_{11}, x_{21}, \ldots, x_{N1}\}$ and $\boldsymbol{x}_2 = \{x_{12}, x_{22}, \ldots, x_{N2}\}$ are the given objects in original space $\mathbb{R}^d$ with a multivariate normal density function:

$$\mathrm{N}(\mathbf{x};\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \tag{6}$$

The radial basis function (RBF) kernel function displayed in Equation (7) can be used to transfer original data into feature space for calculating the distance between uncertain data objects $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The RBF kernel function is commonly used in various fields and algorithms because it outperforms other kernel functions [27,28].

$$K_{ij} = \exp\left(-\frac{1}{2\sigma^2} \|x_i - x_j\|^2 \right),\ i, j = 1, 2 \tag{7}$$

In kernel function $K(\boldsymbol{x}_1,\ \boldsymbol{x}_2)$, where $\boldsymbol{x}_1,\ \boldsymbol{x}_2 \in \mathbb{R}^d$, and the non-linear mapping function ϕ and kernel Gram matrix $\boldsymbol{K}$ are defined as $\boldsymbol{K} = \Phi^{\mathrm{T}}\Phi$, where $\Phi := \ \Phi_N = [\phi(x_1),\ \phi(x_2),\ \ldots,\ \phi(x_N)] \in \mathbb{R}^f$, and $f \gg d$ represents the data transformed to kernel space. The mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ in feature space are estimated as:

$$\boldsymbol{\mu} = N^{-1} \sum_{\mathrm{n}=1}^{\mathrm{N}} \phi(\mathrm{x}_n) = \boldsymbol{\Phi}, \tag{8}$$

$$\boldsymbol{\Sigma} = N^{-1} \sum_{n=1}^{N} (\phi_n - \mu)(\phi_n - \mu)^{\mathrm{T}} = \boldsymbol{\Phi} \boldsymbol{J} \boldsymbol{J}^T \boldsymbol{\Phi}^{\mathrm{T}}, \tag{9}$$

where $\boldsymbol{J} = \frac{1}{\sqrt{n}}(\boldsymbol{I}_N - s\vec{\mathbf{1}})$ with $s_{N\times 1} = \frac{1}{N}\vec{\mathbf{1}}^T$ and $\vec{\mathbf{1}} = [1, 1,\ \ldots,\ 1]$.

The covariance matrix $\boldsymbol{\Sigma}$ must be converted into approximation form because of its rank-deficient characteristic $f \gg d$. Therefore, we can use the approximation form as follows:

$$\boldsymbol{C} = \boldsymbol{\Phi} \boldsymbol{J} \boldsymbol{J}^T \boldsymbol{\Phi}^{\mathrm{T}} + \rho \boldsymbol{I}_f = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \rho \boldsymbol{I}_f = \boldsymbol{\Phi} \boldsymbol{A} \boldsymbol{\Phi}^T + \rho \boldsymbol{I}_f, \tag{10}$$

where $\mathbf{W} \doteq \ \boldsymbol{\Phi} \boldsymbol{J} \boldsymbol{Q}$, $\boldsymbol{A} \doteq \boldsymbol{J} \boldsymbol{Q} \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{J}^{\mathrm{T}}$, and ρ is a user parameter that should be pre-specified in advance.

Obtaining the matrix $\mathbf{Q}$ requires computing the top r eigenvalues matrix $\boldsymbol{\Lambda}_r$ and the top r eigenvectors matrix $\boldsymbol{V}_r$ of $\overline{\boldsymbol{K}} = \boldsymbol{J}^T \boldsymbol{K} \boldsymbol{J}$, where top r is a pre-specified parameter; thus, $r = 3$ is used. $\boldsymbol{Q}$ is an $N \times r$ matrix calculated as follows:

$$\boldsymbol{Q} \doteq \boldsymbol{V}_r \left(\boldsymbol{I}_r - \rho \boldsymbol{\Lambda}_r^{-1} \right)^{1/\,2}. \tag{11}$$

Define matrix $\boldsymbol{P}$ as:

$$\boldsymbol{P}_{(N_1+N_2)\times(r_1+r_2)} = \begin{bmatrix} \sqrt{\alpha_1} \boldsymbol{J}_1 \boldsymbol{Q}_1 & 0 \\ 0 & \sqrt{\alpha_2} \boldsymbol{J}_2 \boldsymbol{Q}_2 \end{bmatrix}. \tag{12}$$

The Bhattacharyya distance is a special case of Chernoff distance; it must be set to $\alpha_1 = \alpha_2 = 1/2$ for all experiments. The $\tau_i, i = 1, \ldots, r_1 + r_2$, are eigenvalues of a $\boldsymbol{L}_{ch}$ matrix, with dimensions of $(r_1 + r_2) \times (r_1 + r_2)$ given by

$$\boldsymbol{L}_{ch} = \mathbf{P}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{\Phi}_1^{\mathrm{T}} \\ \boldsymbol{\Phi}_2^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_1^T & \boldsymbol{\Phi}_2^T \end{bmatrix} \boldsymbol{P} = \boldsymbol{P}^T \begin{bmatrix} \boldsymbol{K}_{11} & \boldsymbol{K}_{12} \\ \boldsymbol{K}_{21} & \boldsymbol{K}_{22} \end{bmatrix} \boldsymbol{P}. \tag{13}$$

Scalar values ε_{11}, ε_{12}, ε_{22} are computed by Equation (14).

$$\varepsilon_{ij} = s_i^T \boldsymbol{K}_{ij} s_j - s_i^T [\boldsymbol{K}_{i1} \boldsymbol{K}_{i2}] \boldsymbol{B}_{ch} \begin{bmatrix} \boldsymbol{K}_{1j} \\ \boldsymbol{K}_{2j} \end{bmatrix} s_j \tag{14}$$

where $\boldsymbol{B}_{ch} = \boldsymbol{P}(\rho \boldsymbol{I}_{r_1+r_2} + \boldsymbol{L}_{ch})^{-1} \boldsymbol{P}^T$ with dimensions of $(N_1 + N_2) \times (N_1 + N_2)$.

The kernel-based probabilistic Bhattacharyya distance between two uncertain data objects $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in feature space is calculated as follows:

$$KPD_{Bhatt} = 0.5[\alpha_1 \alpha_2 \rho^{-1} (\varepsilon_{11} + \varepsilon_{22} - 2\varepsilon_{12}) + 0.5 \sum_{i=1}^{r_1+r_2} \log \frac{\rho + \tau_i}{\lambda_{i,1}} + \sum_{i=1}^{r_1+r_2} \log \frac{\rho + \tau_i}{\lambda_{i,2}}, \tag{15}$$

where $\lambda_{i,j}$, $i = 1, \ldots, r_j$ are the eigenvalues of C_j:

$$\lambda_{i,j} = \begin{cases} \lambda_{i,j}, & when\ i = 1, \ldots, r_j \\ \rho, & when\ i = r_j + 1, \ldots, r_1 + r_2 \end{cases} \tag{16}$$

3.2. New CVI for Uncertain Data

The uncertain data objects in the cluster are transformed into feature space to compute the compactness and separability in the feature space by applying a kernel function. The mapped uncertain data objects are used to compute the distance between different clusters for calculating compactness and separability, which are used to obtain the values of the proposed CVIs. The calculated value of the indices changes according to the number of clusters K, and the proposed uncertain feature space DU (UFSDU) and uncertain feature space CH (UFSCH) index, are defined in Equations (17) and (18), respectively:

UFSDU index:

$$UFSDU_K = \frac{min_{i,j=1,\cdots,K,\ i \neq j} \left\{ min_{x \in C_i,\ y \in C_j} KPD_{Bhatt}(x,y) \right\}}{max_{i=1,\cdots,K} \left\{ max_{x,y \in C_i} KPD_{Bhatt}(x,y) \right\}} \tag{17}$$

UFSCH index:

$$UFSCH_K = \frac{\sum_{i=1}^{K} n_i \cdot KPD_{Bhatt}(z_i \cdot z_{tot})^2}{K-1} \cdot \frac{n-K}{\sum_{i=1}^{k} \sum_{x \in C_i} KPD_{Bhatt}(x, z_i)^2} \tag{18}$$

These proposed CVI equations are similar to the DU and CH indices, except for the term $KPD_{Bhatt}(x,y)$, which is the computed distance between two uncertain data objects in feature space in Equation (15).

4. Experimental Results

In this study, we propose two CVIs that are calculated probabilistic distances between different uncertain data objects in feature space. The K-medoids clustering algorithm proposed by Jiang et al. [19] was used to compare the performances of the proposed CVIs in feature space. The K-medoids algorithm is one of the most useful algorithms in clustering problems, which uses probabilistic distance measures to capture the similarity between uncertain objects. It differs from the popular K-means clustering algorithm used for clustering data into groups in its robustness to outliers. The K-means method represents each cluster by the mean of all objects in this cluster, whereas the K-medoids method calculates the distance between every pair of all uncertain data objects and the medoid within a cluster [19]. Then, of all calculated distance values, uncertain data with the smallest distances are assigned as a new medoid for the cluster. We proceeded with the experiments by setting the value of K, which is the number of clusters and is used as the probabilistic distance measure. In this study, we varied the number of clusters (K) and the

Bhattacharyya distance measure to compute distances between different uncertain data objects in feature space.

4.1. Experimental Procedure for Uncertain Data

Experiments were performed with artificial and real-world datasets that may have sub-clusters and clusters with asymmetrical, arbitrary, and noisy shapes to evaluate the performances of the proposed CVIs. A normalization process was conducted for each feature of the datasets to reduce the scale gap between different features defined in Equation (19):

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{19}$$

where x_{min} and x_{max} are the minimum and maximum values of one feature of the dataset. We then simulated uncertain data objects from certain data objects by following the methodology used by [20].

The pre-existent DU and CH indexes were used to compute uncertain data objects in original space—uncertain original space, DU (UOSUD), and uncertain original space, CH (UOSCH)—to confirm the validity of the proposed CVIs. The overall experimental procedure is represented by Algorithm 1. The procedure used to compare the performances of the proposed CVIs with those of the previous CVIs was as follows: The inputs included the number of uncertain data objects N, the number of object features M, and the number of clusters K. We modeled the uncertain data with multivariate normal distributions. The means of the distributions were the original certain data. The covariances were estimated as follows:

$$f\left(S_i^k \middle| \Psi^k, df^k\right) = \frac{\left|\Psi^k\right|^{\frac{df^k}{2}}}{\frac{p \cdot df^k}{2} \Gamma_p\left(\frac{df^k}{2}\right)} \left|S_i^k\right|^{-\frac{df^k+p+1}{2}} e^{-\frac{1}{2} tr\left(\Psi^k \left(S_i^k\right)^{-1}\right)}, \quad i = 1, \ldots, n_k, \; k = 1, \ldots, K \tag{20}$$

where S_i^k represents the covariance matrices for objects in class k with the inverse Wishart PDF [29], as defined in Equation (20) [20]. Ψ^k is a positive definite scale matrix and df^k is the degree of freedom. p indicates the dimensions of S_i^k, $tr(\cdot)$ is the trace of a matrix, and Γ is the multivariate gamma function.

Algorithm 1: K-medoids for uncertain data using a probabilistic distance measure in feature space.

Input: $\boldsymbol{n}$: The number of objects in cluster k, $\boldsymbol{K}$: The number of clusters, iter = 0;
Randomly select the cluster medoids $\boldsymbol{C^{(0)}} = \{\boldsymbol{c_1^{(0)}}, \ldots, \boldsymbol{c_K^{(0)}}\}$ obtained from the initial clusters
Initialize
$\boldsymbol{CVIs} = \left\{\boldsymbol{cvi^{(1)}}, \ldots, \boldsymbol{cvi^{(K)}}\right\}$ obtained UOSDU, UOSCH, UFSDU, and UFSCH
Repeat
  for $k = 2$ to K
    $\boldsymbol{c_k^{(old)}} = \boldsymbol{c_k^{(0)}}; \boldsymbol{c_k^{(new)}} = 0$
    Compute the new medoids:
    while $c_k^{(old)} \neq c_k^{(new)}$
      $p = \underbrace{argmin}_{1 \leq i \leq n} \sum_{j=1}^{k} KPD_{Bhatt}(\boldsymbol{x_i}, \boldsymbol{c_{jk}})$, where j is an index of cluster medoid in $\boldsymbol{c_k}$
      $\boldsymbol{c_k^{(new)}} = \boldsymbol{x_p}$
    end
    Calculate the $\boldsymbol{cvi^{(k)}}$ using Equations (1), (2), (17), and (18).
  end
  iter = iter + 1
Until (iter = Maxiter)

Step 1: Set K initial clusters with uncertain objects randomly for a given dataset. Run a K-medoids clustering algorithm with different values for the K parameter ($2 \leq K \leq 10$).

Step 2: Obtain the medoids of each cluster for which the sum of the probabilistic distance between the objects is the smallest.

Step 3: Calculate CVIs for all the partitions. We calculated the compactness and separability in kernel space using an RBF kernel function with σ (bandwidth in the RBF kernel function). The optimal value was determined through a set of preliminary experiments by taking [0.1, 0.2, . . . , 4] in σ.

Step 4: We increased the reliability of experimental results by replicating the experiment 100 times for the same dataset with different trial seeds to obtain the initial medoids in Step 1 and used the average value of CVI for each cluster.

Step 5: Finally, we evaluated each CVI and the suggested number of clusters from a CVI; the actual numbers of clusters of a dataset were then compared.

4.2. Experiments with Artificial and Real-World Datasets

Experiments were conducted to evaluate the proposed CVIs in comparison to the pre-existent CVIs. These experiments used 10 datasets with sensitive characteristics containing arbitrariness, sub-clusters, asymmetry, and noise provided by the UCI (https://archive.ics.uci.edu/, accessed on 10 March 2023) [30] and Tomas Barton repositories (https://github.com/deric/clustering-benchmark, accessed on 10 March 2023), which have 122 artificial datasets with arbitrariness, sub-clusters, and asymmetric shapes in two or three features. The datasets from UCI repository, (e.g., D3, D4, D5, and D7) were collected in real environmental conditions; however, the other datasets were artificially created, which can be checked in Tomas Barton repositories.

The summary of datasets used for the experiments is presented in Table 1. Two-dimensional (2D) and three-dimensional (3D) dataset shapes are illustrated in Figure 3. The CVI values were computed by changing the number of clusters (K) in each dataset and then comparing the predicted labels of experiments to the actual labels in the datasets.

Table 1. Summary of datasets.

Dataset Index	Dataset Name	# of Obs.	# of Dim.	# of Clusters	Projection Shape
D1	A.K Jain's Toy	373	2	2	Asymmetry, Arbitrary shape
D2	Flame	240	2	2	Sub-cluster, Noise
D3	Iris	150	4	3	-
D4	Thyroid	215	5	2	-
D5	Wine	178	13	3	-
D6	Wisconsin	683	9	2	-
D7	Harberman	301	3	2	Random shape
D8	Chainlink	1000	3	2	Sub-cluster, Arbitrary shape
D9	Lsun	400	2	3	Asymmetry, Arbitrary shape
D10	Zelnik1	299	2	3	Sub-cluster

4.3. Performance Comparison of the Proposed CVIs

The experimental results are given in Tables 2–11. The actual number of clusters is below the name of the dataset. It is also noted with an asterisk (*) adjacent to the actual number of clusters along the top. Moreover, all the results of the datasets are presented in Table 12, indicating the performance of the proposed CVIs by a quantitative figure. Each cell in Table 12 represents the optimal number of clusters K determined by its CVI criteria.

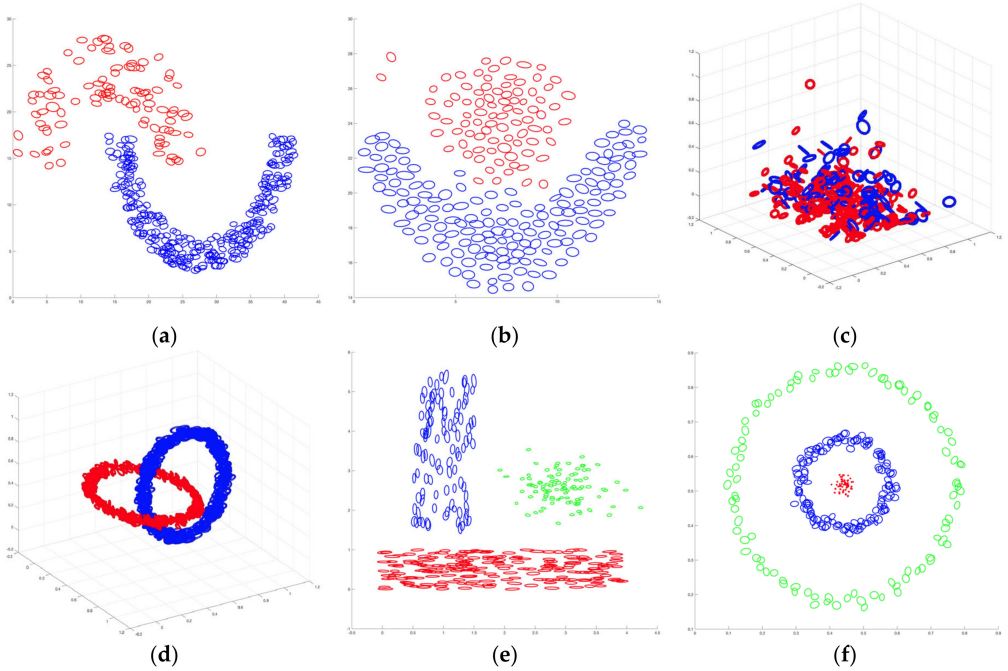


Figure 3. Shapes of 2D and 3D datasets: (**a**) D1 dataset; (**b**) D2 dataset; (**c**) D7 dataset; (**d**) D8 dataset; (**e**) D9 dataset; (**f**) D10 dataset.

Table 2. Performance results for D1.

	# of Clusters / CVI	2 *	3	4	5	6	7	8	9	10
D1 (2)	UOSDU	**0.00075**	0.00063	0.00049	0.00046	0.00043	0.00047	0.00044	0.00042	0.000410
	UOSCH	554.4796	537.8279	573.5387	**586.5310**	576.5872	562.0666	575.2021	566.6556	567.6008
	UFSDU	**0.011830**	0.00727	0.007410	0.006350	0.006920	0.006390	0.006740	0.00580	0.005630
	UFSCH	**256.0945**	204.767	167.9338	149.5915	138.3076	128.206	122.4676	117.0263	112.4593

Table 3. Performance results for D2.

	# of Clusters / CVI	2 *	3	4	5	6	7	8	9	10
D2 (2)	UOSDU	0.00578	0.00581	0.00583	**0.00533**	0.00494	0.00494	0.00452	0.00454	0.00448
	UOSCH	**218.9052**	188.6698	201.7685	195.0877	190.2412	190.7961	192.3785	187.7774	186.0032
	UFSDU	**0.01875**	0.01433	0.01619	0.01386	0.01284	0.01261	0.01263	0.0125	0.01271
	UFSCH	**246.7711**	190.3472	184.7522	163.52	150.3938	143.1108	138.9139	131.6189	127.3284

Table 4. Performance results for D3.

	# of Clusters / CVI	2	3 *	4	5	6	7	8	9	10
D3 (3)	UOSDU	**0.57393**	0.18691	0.06671	0.04599	0.03375	0.03045	0.02475	0.02443	0.02427
	UOSCH	**393.8149**	340.7616	288.9103	257.4766	227.8328	211.7321	193.9894	179.4227	172.1492
	UFSDU	**0.78121**	0.05291	0.0332	0.02818	0.0201	0.02217	0.02033	0.01676	0.01503
	UFSCH	97.24412	**100.9677**	83.47847	74.68629	65.08186	59.80128	55.42499	51.32508	48.54411

Table 5. Performance results for D4.

	# of Clusters / CVI	2 *	3	4	5	6	7	8	9	10
D4 (2)	UOSDU	**0.01059**	0.00702	0.00447	0.00389	0.00338	0.00285	0.0029	0.00264	0.00254
	UOSCH	**52.44662**	49.27229	45.29772	44.23136	46.29286	43.05835	40.0334	38.99379	36.43862
	UFSDU	**0.09045**	0.02678	0.02097	0.01941	0.0186	0.0166	0.01728	0.01604	0.01577
	UFSCH	**88.16833**	63.62494	54.54528	48.32164	43.33752	38.65073	35.53446	32.89777	30.6346

Table 6. Performance results for D5.

	# of Clusters / CVI	2	3 *	4	5	6	7	8	9	10
D5 (3)	UOSDU	**0.28546**	0.19218	0.16953	0.13451	0.13042	0.12188	0.1222	0.11775	0.11498
	UOSCH	**46.98845**	41.61822	34.08324	29.45127	26.66111	23.71564	21.97848	20.8878	19.0692
	UFSDU	0.1351	**0.13992**	0.12361	0.11102	0.1058	0.10544	0.10343	0.10402	0.10242
	UFSCH	**166.5115**	94.11775	70.17926	57.15066	48.48219	42.44803	38.19718	34.55733	31.1674

Table 7. Performance results for D6.

	# of Clusters / CVI	2 *	3	4	5	6	7	8	9	10
D6 (2)	UOSDU	**0.10223**	0.04719	0.02262	0.01209	0.00742	0.00342	0.0014	0.00109	0.00075
	UOSCH	**237.829**	186.8503	145.4631	119.3866	98.36381	89.72379	80.18472	70.83073	66.12163
	UFSDU	**0.22631**	0.10763	0.04928	0.03902	0.01416	0.01228	0.0084	0.00605	0.00391
	UFSCH	**349.3685**	261.4169	205.8692	171.2457	144.4285	124.5582	109.2653	97.50292	88.98401

Table 8. Performance results for D7.

	# of Clusters / CVI	2 *	3	4	5	6	7	8	9	10
D7 (2)	UOSDU	**0.00198**	0.0014	0.00112	0.00086	0.00078	0.00089	0.00069	0.00079	0.00076
	UOSCH	**128.8359**	117.8517	104.8203	97.56451	95.82686	92.17925	86.85381	84.98897	82.71107
	UFSDU	**0.13021**	0.02577	0.01681	0.01199	0.01108	0.01122	0.01132	0.01028	0.00945
	UFSCH	**319.3255**	171.7169	127.0638	104.5919	90.63319	80.94442	72.86994	67.51974	62.62473

Table 9. Performance results for D8.

	# of Clusters / CVI	2 *	3	4	5	6	7	8	9	10
D8 (2)	UOSDU	0.00019	0.00017	0.00017	0.00017	0.00017	0.00018	**0.00021**	0.00019	0.00017
	UOSCH	419.8882	371.9768	388.8548	430.2229	426.5956	430.8854	**449.3122**	438.7834	417.3569
	UFSDU	**0.00439**	0.00237	0.00204	0.00114	0.0013	0.00118	0.00149	0.00153	0.0014
	UFSCH	445.5408	**463.2664**	449.4758	439.8262	425.4487	411.5018	422.1565	428.8755	437.9047

Table 10. Performance results for D9.

	# of Clusters / CVI	2	3 *	4	5	6	7	8	9	10
D9 (3)	UOSDU	**0.01277**	0.00168	0.00087	0.00081	0.00069	0.00062	0.0006	0.00063	0.00054
	UOSCH	316.7407	**406.3877**	395.188	401.578	380.968	363.1193	365.4242	349.9761	351.8199
	UFSDU	0.01439	**0.02006**	0.01697	0.01119	0.00658	0.00574	0.00485	0.00472	0.00416
	UFSCH	190.3465	**205.1745**	189.6315	175.6124	164.8462	154.2108	149.907	141.5702	133.6363

Table 11. Performance results for D10.

	# of Clusters / CVI	2	3 *	4	5	6	7	8	9	10
D10 (3)	UOSDU	0.030644	**0.049296**	0.048849	0.048798	0.046752	0.044594	0.042478	0.037749	0.041905
	UOSCH	**235.4205**	161.3342	142.117	135.4194	127.012	126.4954	125.6673	123.9964	132.4379
	UFSDU	**0.00368**	0.00123	0.00123	0.00115	0.00103	0.00087	0.00073	0.00077	0.00056
	UFSCH	102.6013	**106.5976**	99.7133	98.79822	97.68495	95.67929	95.82844	96.62246	102.6371

Table 12. Difference between the actual and estimated numbers of clusters in lower-dimensional datasets.

Dataset	Dim	# of Clusters	UOSDU	UOSCH	UFSDU	UFSCH
D1	2	2	⊙	5	⊙	⊙
D2	2	2	4	⊙	⊙	⊙
D3	4	3	2	2	2	⊙
D4	5	2	⊙	⊙	⊙	⊙
D5	13	3	2	2	⊙	2
D6	9	2	⊙	⊙	⊙	⊙
D7	3	2	⊙	⊙	⊙	⊙
D8	3	2	8	8	⊙	3
D9	2	3	2	⊙	⊙	⊙
D10	2	3	⊙	2	2	⊙
# of successes in estimating the optimal number of clusters			5	5	8	8

The bold values with gray-shaded backgrounds indicate the optimal cluster K decided by each CVI. As presented in Table 2, three of the CVIs succeeded in estimating the number of clusters as two in D1. UOSCH failed. The proposed UFSDU and UFSCH also successfully predicted the number of clusters in D2. In contrast, UOSDU failed to estimate the number of clusters in D2.

Although the proposed UFSDU index and the pre-existent CVIs failed to predict the number of clusters in D3, UFSCH was successful. All CVIs correctly predicted the number of clusters for some datasets; see Tables 5, 7 and 8. In contrast, the proposed UFSDU index is the only CVI that correctly predicted the actual number of clusters in D5, as presented in Table 6. Furthermore, the UFSDU index predicted the actual number of clusters of D8. D8's shape (Figure 3) is classified distinctly into two classes when viewed visually. However, it is challenging to calculate the compactness and separability of a cluster in the original space. Nevertheless, the UFSDU index was successful in such predictions; the UFSCH forecasted the number of clusters as three, which is close to the

actual number of clusters, two. The kernel transformation facilitates computation to obtain greater compactness and separability in the feature space than the original space, leading to high-performance clustering.

The UOSCH index and the new CVIs predicted the number of clusters to be three in D9, and the UOSDU and UFSCH indexes successfully estimated the number of clusters in D10. Table 12 presents a summary of the results of the 10 datasets above, whereas the symbol of a circled dot (⊙) indicates that the CVI accurately predicted the actual number of clusters. As presented in Table 12, the pre-existent CVIs precisely estimated the number of clusters for five experimental datasets, whereas the newly proposed CVIs accurately predicted the number of clusters for eight datasets—three more than the pre-existent CVIs.

5. Conclusions

In this study, we proposed novel cluster validity indices (CVIs) for uncertain data objects in feature space. Unlike conventional CVIs in original space, the proposed CVIs are used for uncertain data objects with arbitrariness, sub-clusters, and noisy shapes of clusters that are hard to evaluate, by transforming the uncertain data from the original space to the feature space, which is performed by the kernel function. The proposed CVIs measure the compactness and separability of each cluster in kernel space, which transforms the original data into a higher-dimensional space, leading to less sensitivity to the arbitrary shapes of clusters and more robustness to noise and outliers. We compared the performances of the proposed CVIs with those of pre-existent CVIs that only consider for the original space. The Bhattacharyya distance measure, one of the most widely used for calculating distance, was used to perform experiments with several artificial and real-life datasets to capture the distances between probability density functions. Numerical examples, including a real-life case study and artificial datasets, confirmed that our proposed CVIs are robust to arbitrary cluster shapes, especially sub-clusters, and are promising alternatives for evaluating the fitness of clustering results that can find the optimal number of clusters, K. The proposed CVIs outperform the pre-existent CVIs because of the application of kernel functions to uncertain data, transforming them from the original space to the feature space. As for practical significance, the proposed CVIs could be utilized in diverse applications. For example, Kim et al. proposed new a multivariate kernel density estimator for uncertain data classification for mixed defect patterns on DRAM wafer maps [31]. The proposed CVI method could be applied for evaluating the number of defect patterns on wafer maps. However, there are some limitations to the proposed CVIs. The uncertain data are assumed to have multivariate normal distributions in advance to compute the distances between different uncertain data objects. The uncertainty of the uncertain data may have a variety of probability functions (normal distribution, exponential distribution, etc.), and some cannot be strictly modeled by PDFs. This might be overcome through methods for generating random variables and support-measure data description, which is a non-parametric machine learning method that does not require an assumption of a prior distribution to be made in advance.

Future research should consider the compactness measure in kernel space in advanced machine learning algorithms, such as support vector data descriptions or Bayesian frameworks of Bayesian support vector data descriptions. The concepts of our CVIs can also be applied to other clustering algorithms.

Author Contributions: Conceptualization, Y.-S.J.; data curation, C.K.; formal analysis, Y.-S.J.; investigation, B.T. and Y.-S.J.; methodology, C.K. and Y.-S.J.; resources, B.T.; software, B.T.; supervision, Y.-S.J.; validation, J.B.; visualization, J.B.; writing—original draft, C.K.; writing—review and editing, J.B., B.T. and Y.-S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by LG Yonam Foundation (of Republic of Korea) and by National Research Foundation of Republic of Korea Grant (No. NRF-2021S1A5A8060639, NRF-2022R1F1A1063174).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The real-world datasets used in this study are available at: https://archive.ics.uci.edu/ml/index.php accessed on 10 March 2023; the artificial datasets that contain data sensitive to shapes are available at: https://github.com/deric/clustering-benchmark/tree/master/ accessed on 10 March 2023.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Bhatt	Bhattacharyya distance measure
C/S	Separability/Compactness
CH	Calinski–Harabasz
CVIs	Cluster validity indices
DB	Davies–Bouldin
DU	Dunn
KPD	Kernel-based probabilistic distance
PD	Probabilistic distance
PDF	Probability density function
RBF	Radial basis function
S/C	Compactness/Separability
UFSCH	Uncertain feature space CH
UFSDU	Uncertain feature space DU
UOSCH	Uncertain feature space CH
UOSDU	Uncertain feature space DU

References

1. Abdalameer, A.K.; Alswaitti, M.; Alsudani, A.A.; Isa, N.A. A new validity clustering index-based on finding new centroid positions using the mean of clustered data to determine the optimum number of clusters. *Expert Syst. Appl.* **2022**, *191*, 116329. [CrossRef]
2. Irani, J.; Pise, N.; Phatak, M. Clustering techniques and the similarity measures used in clustering: A survey. *Int. J. Comput. Appl. Technol.* **2016**, *134*, 9–14. [CrossRef]
3. MacQueen, J.B. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 27 December 1965–7 January 1966; The Regents of the University of California: Santa Barbara, CA, USA, 1967; pp. 281–297.
4. Li, M.J.; Ng, M.K.; Cheung, Y.-m.; Huang, J.Z. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 1519–1534. [CrossRef]
5. Mahesh Kumar, K.; Rama Mohan Reddy, A. A fast DBSCAN clustering algorithm by accelerating neighbor searching using groups method. *Pattern Recognit.* **2016**, *58*, 39–48. [CrossRef]
6. Chien, C.-F.; Wang, W.-C.; Cheng, J.-C. Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Syst. Appl.* **2007**, *33*, 192–198. [CrossRef]
7. El-shafeiy, E.; Sallam, K.M.; Chakrabortty, R.K.; Abohany, A.A. A clustering based swarm intelligence optimization technique for the internet of medical things. *Expert Syst. Appl.* **2021**, *173*, 114648. [CrossRef]
8. Aggarwal, C.C.; Yu, P.S. A survey of uncertain data algorithms and applications. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 609–623. [CrossRef]
9. Shou, L.; Zhang, X.; Chen, G.; Gao, Y.; Chen, K. Mud: Mapping-based query processing for high-dimensional uncertain data. *Inf. Sci.* **2012**, *198*, 147–168. [CrossRef]
10. Duan, X.; Ma, Y.; Zhou, Y.; Huang, H.; Wang, B. A novel cluster validity index based on augmented non-shared nearest neighbors. *Expert Syst. Appl.* **2023**, *223*, 119784. [CrossRef]
11. Lee, S.-H.; Jeong, Y.-S.; Kim, J.-Y.; Jeong, M.K. A new clustering validity index for arbitrary shape of Clusters. *Pattern Recognit. Lett.* **2018**, *112*, 263–269. [CrossRef]
12. Dunn, J.C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104. [CrossRef]
13. Calinski, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **1974**, *3*, 1–27. [CrossRef]
14. Davies, D.L.; Bouldin, D.W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [CrossRef]
15. Xie, X.L.; Beni, G. A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 841–847. [CrossRef]
16. Rojas-Thomas, J.C.; Santos, M.; Mora, M. New internal index for clustering validation based on graphs. *Expert Syst. Appl.* **2017**, *86*, 334–349. [CrossRef]

17. Tavakkol, B.; Jeong, M.K.; Albin, S.L. Validity indices for clusters of uncertain data objects. *Ann. Oper. Res.* **2018**, *303*, 321–357. [CrossRef]
18. Wang, J.-S.; Chiang, J.-C. A cluster validity measure with a hybrid parameter search method for the support vector clustering algorithm. *Pattern Recognit.* **2008**, *41*, 506–520. [CrossRef]
19. Jiang, B.; Pei, J.; Tao, Y.; Lin, X. Clustering uncertain data based on probability distribution similarity. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 751–763. [CrossRef]
20. Tavakkol, B.; Jeong, M.K.; Albin, S.L. Object-to-group probabilistic distance measure for uncertain data classification. *IEEE Trans. Knowl. Data Eng.* **2017**, *230*, 143–151. [CrossRef]
21. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **2013**, *46*, 243–256. [CrossRef]
22. Rezaee, B. A cluster validity index for Fuzzy Clustering. *Fuzzy Sets Syst.* **2010**, *161*, 3014–3025. [CrossRef]
23. Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Sankhya Indian J. Stat.* **1946**, *7*, 401–406.
24. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
25. Tavakkol, B.; Son, Y. Fuzzy kernel K-medoids clustering algorithm for uncertain data objects. *Pattern Anal. Appl.* **2021**, *24*, 1287–1302. [CrossRef]
26. Zhou, S.K.; Chellappa, R. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 917–929. [CrossRef] [PubMed]
27. Patle, A.; Chouhan, D.S. SVM kernel functions for classification. In Proceedings of the 2013 International Conference on Advances in Technology and Engineering (ICATE), Mumbai, India, 23–25 January 2013.
28. Tbarki, K.; Ben Said, S.; Ksantini, R.; Lachiri, Z. RBF kernel based SVM Classification for landmine detection and discrimination. In Proceedings of the 2016 International Image Processing, Applications and Systems (IPAS), Sfax, Tunisia, 5–7 November 2016.
29. Nydick, S.W. The wishart and inverse wishart distributions. *Electron. J. Stat.* **2012**, *6*, 1–19.
30. UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu/ (accessed on 28 March 2023).
31. Kim, B.; Jeong, Y.-S.; Jeong, M.K. New multivariate kernel density estimator for uncertain data classification. *Ann. Oper. Res.* **2020**, *303*, 413–431. [CrossRef]