Annals of Operations Research (2021) 303:321–357
https://doi.org/10.1007/s10479-018-3043-4

S.I.: DATA MINING AND DECISION ANALYTICS

Validity indices for clusters of uncertain data objects

Behnam Tavakkol[1] · Myong K. Jeong[2] · Susan L. Albin[2]

Published online: 10 September 2018

Abstract

Clustering validity indices are the main tools for evaluating the quality of formed clusters and determining the correct number of clusters. They can be applied on the results of clustering algorithms to validate the performance of those algorithms. In this paper, two clustering validity indices named uncertain Silhouette and Order Statistic, are developed for uncertain data. To the best of our knowledge, there is not any clustering validity index in the literature that is designed for uncertain objects and can be used for validating the performance of uncertain clustering algorithms. Our proposed validity indices use probabilistic distance measures to capture the distance between uncertain objects. They outperform existing validity indices for certain data in validating clusters of uncertain data objects and are robust to outliers. The Order Statistic index in particular, a general form of uncertain Dunn validity index (also developed here), is well capable of handling instances where there is a single cluster that is relatively scattered (not compact) compared to other clusters, or there are two clusters that are close (not well-separated) compared to other clusters. The aforementioned instances can potentially result in the failure of existing clustering validity indices in detecting the correct number of clusters.

Keywords Clustering validity index · Uncertain data · Probabilistic distance measures · Data mining

1 Introduction

In traditional data mining problems, each data object is associated with only a single point value. This means that no uncertainty is assumed for each data object. This type of data mining problem is referred to as certain data mining. There is another type of data mining problem that is called uncertain data mining. In uncertain data mining problems, each data object is not a single point value anymore and a level of uncertainty is assumed for each data object.

Uncertain data objects come in two possible forms: (1) multiple points for each object, and (2) a probability density function (pdf) for each object, either given or obtained by fitting

✉ Myong K. Jeong
mjeong@rci.rutgers.edu

[1] Stockton University, Galloway, USA

[2] Rutgers University, Piscataway, USA

to the multiple points (Tavakkol et al. 2017). The most general application of uncertain data is where at each setting there are multiple repeated measurements instead of a single measurement. Other applications of uncertain data mining are found in sensor networks, moving object databases, and medical and biological databases (Qin et al. 2009). Images are also the type of objects that can be considered as uncertain data objects. Given an image with at least one object, each image object can be converted to a group of two-dimensional points. Then the whole image can be modeled with all the points from all the objects in the image, or with the probability density function that can be fitted to all the points.

Uncertain data objects can be reduced to certain data objects if only a representative statistic such as the mean of each object is considered (Tavakkol et al. 2017). However, uncertain data objects naturally carry extra information compared to certain data objects that would be discarded if they are converted to certain data objects. This shows the importance of designing data mining techniques that can handle uncertain data objects and capture their extra information.

Developing uncertain data mining techniques has been the topic of many researches. In Aggarwal and Philip (2009), a comprehensive review of the literature in four categories of uncertain data mining problems i.e. classification, clustering, outlier detection and frequent pattern mining, is provided.

Clustering, one of the main techniques in data mining, falls under the category of unsupervised techniques. Unsupervised techniques work with no class label information provided. Clustering is about organizing the objects in a data set into coherent and contrasted groups or as we call them clusters (Pakhira et al. 2004). The objective is to form clusters so that the objects in the same cluster are close to each other but are far from the objects in other clusters. In other words, the objective of clustering is to form clusters so that they are compact and also well-separated from each other. For more information about clustering algorithms, see Qin et al. (2017), Marinakis et al. (2011) and Duan et al. (2009).

For certain data, many popular clustering algorithms exist in the literature. One of the most well-known is K-means (Chiang et al. 2011; Hartigan and Wong 1979). With the number of clusters known a priori, the K-means algorithm optimizes either by minimizing the within-cluster spread (forming compact clusters), or by maximizing the between-cluster spread (forming separated clusters).

Uncertain data clustering algorithms have been the topic of a few research studies that appear in Aggarwal and Philip (2009), Chau et al. (2006), Lee et al. (2007), Gullo et al. (2008a, b, 2010, 2013, 2017), Kao et al. (2010), Yang and Zhang (2010) and Kriegel and Pfeifle (2005). A comprehensive survey of uncertain data algorithms which includes clustering algorithms as well is provided in Aggarwal and Philip (2009). In Chau et al. (2006), a K-means clustering algorithm for uncertain data objects is developed which uses the expected distance to capture the dissimilarity between two uncertain objects. It is shown in Lee et al. (2007) that the uncertain K-means algorithm of Chau et al. (2006) can be reduced to certain K-means algorithm. A hierarchical clustering algorithm for uncertain data is proposed in Gullo et al. (2008b, 2017). Clustering uncertain data using Voronoi diagrams and r-tree index is developed in Kao et al. (2010). Mixture model clustering of uncertain data objects is investigated in Gullo et al. (2010, 2013). In Gullo et al. (2008a) and Yang and Zhang (2010), K-medoids clustering algorithms for uncertain data objects using the expected distance as the distance between the two objects are proposed. In Jiang et al. (2013) and Kriegel and Pfeifle (2005), density-based clustering algorithms FDBSCAN and uncertain DBSCAN with probabilistic distance measures are developed. A K-medoids clustering algorithm that uses probabilistic distance measures for capturing the distance between uncertain objects is also developed in Jiang et al. (2013). In this paper, we use the uncertain K-medoids clustering algorithm with probabilistic distance measures to evaluate the performance of our proposed clustering validity indices.

There are two important questions that need to be addressed in any clustering problem (Fraley and Raftery 1998; Halkidi et al. 2001). One is about the actual number of clusters that are present in the data set. And another question is about the validity and goodness of the formed clusters. The answers to these two questions can be obtained by using clustering validity indices. Clustering validity indices are single numerical values that are obtained by incorporating both the compactness and separation of clusters (Pal and Biswas 1997). When the question is to find the best number of clusters, first, a clustering algorithm such as K-means should be used. The desirable number of generated clusters k, $k=1,\ldots,n$ can be set as an input for the clustering algorithm. After the clusters are formed, clustering validity indices use the formed clusters i.e. the output of the clustering algorithm and provide a value for each k, $k=1,\ldots,n$. Depending on the clustering validity index, the best number of clusters is detected as the one that produces the largest or smallest value of the index. Similar to the procedure used to find the correct number of clusters, clustering validity indices can be used to evaluate the goodness of clusters. For any fixed number of clusters, different clustering algorithms might produce different clusters. In these cases, also the best formed clusters can be detected as the ones that produce the largest or smallest index values.

There are many clustering validity indices for certain data objects such as Dunn (Dunn 1973), Davies–Bouldin (Davies and Bouldin 1979), Xie–Beni (Xie and Beni 1991), Silhouette (Rousseeuw 1987), Caliński–Harabasz (Caliński and Harabasz 1974), and Pakhira–Bandyopadhyay–Maulik (Pakhira et al. 2005). The first four indices, i.e. Dunn, Davies–Bouldin, Xie–Beni, and Silhouette, are of the most well-known and widely used ones in the literature, and therefore are used for evaluation purposes in this paper. To the best of our knowledge, there is not any clustering validity index in the literature that is designed for uncertain objects modeled with pdf or multiple points and can be used for validating the performance of uncertain clustering algorithms.

In this paper, we propose two uncertain clustering validity indices for uncertain data objects: uncertain Silhouette and Order Statistic (OS) index. The proposed indices are both superior to existing certain clustering validity indices for validating clusters of uncertain data objects.

Our proposed clustering validity indices are designed to detect the best setting as the one where the formed clusters are as compact and separated as possible. The uncertain Silhouette index considers the exclusive contribution of every single object to compactness and separation of clusters in the data set. The index also uses scaled values (between -1 and 1) for every object's contribution, and hence is very robust to outliers. In this index, to capture the distance between uncertain data objects probabilistic distance measures are used.

The OS index considers the average of r smallest inter-cluster distances for separation, and the average of r largest intra-cluster distances for compactness, instead of using the exclusive contribution of every object. This can be potentially useful for cases where the key characteristics of clusters are determined by only a few objects and considering other unimportant objects might fade away the contribution of the key objects and weaken the performance of the index. The OS index is also the general case of the uncertain Dunn index which is developed in this paper as well. The advantage of the OS index over the uncertain Dunn is to detect the correct number of clusters in cases where there is either a very large dominant compactness value (a very spread cluster), or there is a small dominant separation value (two very close clusters). Those are the two types of problems for which uncertain Dunn index does not perform well. Like uncertain Silhouette, the OS index uses probabilistic distance measures to capture the distance between uncertain data objects and is robust to existence of outliers too.

Through several experiments, we evaluate the performance of our proposed clustering validity indices over the certain clustering validity indices. The experiments include synthetic data sets with different sizes and dimensions, a real weather data set, and two image data sets. We also show the ability of handling outliers with experiments with synthetic data sets.

The remaining sections of this paper are as follow. In Sect. 2, four of the most widely used clustering validity indices for certain data objects are explained in detail: Dunn; Davies–Bouldin; Silhouette; and Xie–Beni. In Sect. 3, probabilistic distance measures that are used for capturing the distance between two uncertain objects are introduced. The utilized uncertain K-medoids algorithm is also explained in this section. Our proposed uncertain clustering validity indices are explained in Sect. 4. In Sect. 5, experiments for evaluating the performance of the developed clustering validity indices on synthetic and real data are presented. Finally, the paper is concluded in Sect. 6.

2 Clustering validity indices for certain data objects

In this paper we only consider crisp clusters, i.e. clusters in which objects only belong to one cluster. For this reason, four clustering validity indices that are widely used for crisp certain data are explained in this section. These indices are used for benchmarking. The four indices are Dunn (Dunn 1973), Davies–Bouldin (Davies and Bouldin 1979), Xie–Beni (Xie and Beni 1991), and Silhouette (Rousseeuw 1987). Dunn, Davies–Bouldin, and Silhouette are indices that are derived based on crisp clusters. Xie–Beni though, is originally derived for fuzzy clusters, i.e. clusters in which objects can belong to more than one cluster. However, its reduced form can be used for crisp clusters. For further discussion on validity indices for crisp and fuzzy clusters, see Halkidi et al. (2001).

2.1 Dunn index

Dunn index is a clustering validity index for clusters of certain data objects. It considers the distance between the two least separated clusters as the separation of the K clusters. It also considers the compactness of the least compact cluster as the compactness of the K clusters. The index is defined in Eq. (1) for K clusters:

$$DU_K = \frac{\min\limits_{\substack{1 \le i, j \le K \\ j \ne i}} \left(dist\left(C_i, C_j\right)\right)}{\max\limits_{1 \le m \le K} \{diam(C_m)\}}, \tag{1}$$

where $dist\left(C_i, C_j\right)$ denotes the distance between two clusters C_i and C_j and is defined as the distance between the two closest objects of the two clusters and $diam(C_m)$ denotes the diameter of cluster C_m which is used for capturing the compactness of the cluster. The diameter of a cluster is defined as the distance between the two farthest objects in the cluster. Large values of the Dunn index indicate existence of compact and well-separated clusters.

2.2 Davies–Bouldin

Davies and Bouldin (1979) propose incorporating separation and compactness of all pairs of certain data clusters C_i and C_j, with $R_{ij}, i, j = 1, \ldots, K, \ i \neq j$, where

$$R_{ij} = \frac{\left(S_i + S_j\right)}{d_{ij}}, \tag{2}$$

and captures both the separation and compactness for the pair of clusters C_i and C_j. S_i and S_j are the components that capture the compactness of certain data clusters C_i and C_j, and d_{ij} captures the distance between the two clusters. The compactness of cluster C_i can be defined as the average Euclidean distance of objects in cluster C_i to the centroid of the cluster. The distance between clusters C_i and C_j is used to capture the separation of the two clusters. It can be defined as the distance between the centroids of clusters C_i and C_j: Davies–Bouldin uses $\max_{j=1,\ldots,K, i \neq j} R_{ij}$ to define R_i for cluster C_i and eventually returns the index value as $DB_K = \frac{1}{K} \sum_{i=1}^{K} R_i$. Small values of the Davies–Bouldin index may indicate more compact and well-separated clusters.

2.3 Silhouette

The Silhouette index captures separation and compactness for every single certain object. For K clusters the index is defined in Eq. (3) as follows:

$$SI_K = \frac{1}{n} \sum_{i=1}^{n} \frac{(b_i - a_i)}{\max(a_i, b_i)}. \tag{3}$$

In this index, separation and compactness are captured through two components. Compactness for object $\boldsymbol{x_i}$ is captured by component a_i that is defined as the average pairwise distance between object $\boldsymbol{x_i}$ and all objects in the same cluster as object $\boldsymbol{x_i}$.

Separation for object $\boldsymbol{x_i}$ is captured by component b_i that is considered as the separation between object $\boldsymbol{x_i}$ and the closest cluster to it. The separation between object $\boldsymbol{x_i}$ and cluster C_j is defined as the average pairwise distance between object $\boldsymbol{x_i}$ and all objects in cluster C_j.

Silhouette, for each object, computes a scaled value of the difference between separation and compactness and eventually, returns the average of the scaled differences over all objects. Higher values of the index imply large separation and also more compactness which are the desirable characteristics of clusters.

2.4 Xie–Beni

Xie–Beni index for crisp certain data is defined in Eq. (4). The index captures compactness by obtaining the mean of squared distances between data objects and their cluster centroids. Separation is captured with the minimum squared distance between cluster centroids.

$$XB_K = \frac{\sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_{l_i}} d(\boldsymbol{x}, \boldsymbol{z_i})^2}{n \cdot \min\limits_{\substack{i,j=1,\ldots,K \\ i \neq j}} d\left(\boldsymbol{z_i}, \boldsymbol{z_j}\right)^2} \tag{4}$$

Here z_i is the centroid of cluster C_i and C_{l_i} indicates the cluster that object $\boldsymbol{x_i}$ has been assigned to. For Xie–Beni, smaller values of the index indicate large separation and more compactness.

3 Probabilistic distance measures and an uncertain K-medoids clustering algorithm

3.1 Measuring the distance between two uncertain objects

In this paper, we utilize probabilistic distance measures (pdm) to capture the distance between two uncertain objects. There are numerous applications for pdms in many areas such as pattern recognition, communication theory, and statistics (Cover and Thomas 2012; Csiszar and Körner 2011; Zhou and Chellappa 2004). They are also used for estimating the bound on Bayes classification error, signal selection, and asymptotic analysis (Basseville 1989; Chernoff 1952; Devijver and Kittler 1982). Some of the most well-known probabilistic distance measures are: Variational, Chernoff, Generalized Matusita, Kullback–Leibler, Hellinger, and Bhattacharyya (Basseville 1989). Hellinger and Bhattacharyya are special cases of Generalized Matusita and Chernoff respectively. Any of these pdms can be used to capture the distance between two uncertain objects but in this paper, we use Bhattacharyya pdm (Bhattacharyya 1946), one of the most well-known measures. The definition of Bhattacharyya distance is shown in Eq. (5):

$$pd_B(\boldsymbol{X}, \boldsymbol{Y}) = -\ln\left(\int_{\boldsymbol{t}} \sqrt{p_{\boldsymbol{X}}(\boldsymbol{t})p_{\boldsymbol{Y}}(\boldsymbol{t})}d\boldsymbol{t}\right), \tag{5}$$

where $p_{\boldsymbol{X}}(\boldsymbol{t})$ and $p_{\boldsymbol{Y}}(\boldsymbol{t})$ denote the pdfs of uncertain objects $\boldsymbol{X}$ and $\boldsymbol{Y}$ and $\boldsymbol{t} \in R^{\boldsymbol{p}}$. If uncertain objects are given in form of multiple points, instead of pdfs, histograms can be built for each object. Equation (6) shows the definition of Bhattacharyya pdm when objects are given in form of multiple points (Cha 2007).

$$pd_B(\boldsymbol{X}, \boldsymbol{Y}) = -\ln\left(\sum_{i=1}^{b} \sqrt{p_{\boldsymbol{X}}^{(i)} p_{\boldsymbol{Y}}^{(i)}}\right), \tag{6}$$

where $p_{\boldsymbol{X}}^{(i)}$ and $p_{\boldsymbol{Y}}^{(i)}$ denote the frequency of points in the i-th bin for uncertain objects $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. In the equation, b denotes the number of bins.

One of the main advantages of using Bhattacharyya pdm is when uncertain objects follow multivariate normal distributions, Bhattacharyya yields an analytical solution for the pdm between the two objects as shown in Eq. (7):

$$pd_B(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{4}(\boldsymbol{\mu_X} - \boldsymbol{\mu_Y})'(\Sigma_X + \Sigma_Y)^{-1}(\boldsymbol{\mu_X} - \boldsymbol{\mu_Y}) + \frac{1}{2}\ln\left(\frac{|\Sigma_X + \Sigma_Y|}{2(|\Sigma_X||\Sigma_Y|)^{\frac{1}{2}}}\right), \tag{7}$$

where $\boldsymbol{X} \sim MVN(\boldsymbol{\mu_X}, \Sigma_X)$ and $\boldsymbol{Y} \sim MVN(\boldsymbol{\mu_Y}, \Sigma_Y)$. Here, $\boldsymbol{\mu_X}$ and $\boldsymbol{\mu_Y}$ are means, and Σ_X and Σ_Y are covariance matrices.

3.2 Uncertain K-medoids clustering algorithm

Different uncertain K-medoids clustering algorithms have been proposed in the literature. Uncertain K-medoids algorithms that use the expected distance to capture the dissimilarity between two uncertain objects are developed in Gullo et al. (2008a) and Yang and Zhang (2010). In Jiang et al. (2013), an uncertain K-medoids algorithm that uses pdms to capture the distance between uncertain objects, is proposed. In this paper, we use the latter algorithm and use Bhattacharyya as the pdm. The steps of the uncertain K-medoids algorithm are as follow:

Step 1 Pick K initial uncertain objects (medoids) randomly. Form clusters by assigning each object to the cluster for which the probabilistic distance between the object and the cluster medoid is smallest.

Step 2 Obtain the new medoids, $m_k,\ k = 1, \ldots, K$, as follow:

$$m_k = arg \min_{\boldsymbol{X}_i \in C_k} \sum_{\boldsymbol{X}_j \in C_k \setminus \{\boldsymbol{X}_i\}} pd_B\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) \tag{8}$$

where $pd_B\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)$ denotes the Bhattacharyya probabilistic distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$.

Step 3 Using the new medoids, re-assign each object to the cluster of its nearest medoid. Repeat Step 2 and Step 3 until there is no change in the clusters.

4 The proposed uncertain clustering validity indices

In this section we explain the reason uncertain data objects require their own clustering validity indices through an example. Figure 1a shows a two-dimensional example where there are two clusters of uncertain data objects. Objects in both clusters are in form of bivariate normal pdfs and are represented by ellipses. Each ellipse basically represents a contour of a bivariate normal pdf of an object. In Tavakkol (2018), the correlation for uncertain objects is defined as the correlation among the dimensions considering the object mean points only, plus the average object-correlation, where object-correlation is defined as the correlation among the dimensions within object. Based on that definition, in Fig. 1a, objects in one cluster (shown in red) have positive correlation among their two dimensions, while objects in the other cluster (shown in blue) have negative correlation. Applying uncertain K-medoids clustering algorithm with $K=2$ on the objects, the two clusters are detectable.

In order to find the correct number of clusters of this example (i.e. two), a clustering validity index is needed. If the clustering validity indices for certain data objects that only use the object means, are used, the results would not be desirable and one cluster would be preferred to two clusters. The reason can be seen in Fig. 1b, where only the object means are shown and it is impossible to distinguish between the red and blue clusters. Clustering validity indices designed for uncertain data objects should prefer two clusters over a single cluster in this example. We show in the experiments section that our developed uncertain clustering validity indices are well capable of capturing the structures of uncertain objects and distinguishing overlapping clusters such as the ones in Fig. 1.

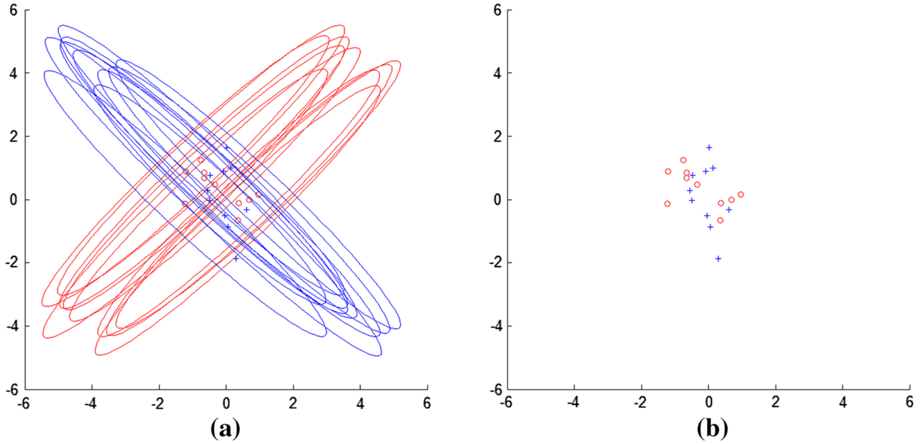


Fig. 1 Two clusters of uncertain data **a** each uncertain object shown with its whole pdf. **b** Each uncertain object from (**a**) shown with its mean only

4.1 Uncertain Silhouette

Our first proposed cluster validity index for uncertain data objects is called uncertain Silhouette index. The definition of the uncertain Silhouette is shown in Eq. (9).

$$USI_K = \frac{1}{n}\sum_{i=1}^{n}\frac{(ub_i - ua_i)}{\max(ua_i, ub_i)}, \tag{9}$$

where ua_i denotes the compactness and ub_i denotes the separation for uncertain object $\boldsymbol{X_i}$. The definitions of ua_i and ub_i are shown in Eqs. (10) and (11) respectively. As it can be seen from Eq. (10), similar to the case for certain data, compactness of an object $\boldsymbol{X_i}$ is defined as the average pairwise distance between the object $\boldsymbol{X_i}$ and all objects in the same cluster as object $\boldsymbol{X_i}$. The main difference between ua_i and a_i (compactness component of Silhouette index for certain data objects) is that in ua_i pdms are used to better capture the distance between uncertain objects, while in a_i distance measures for certain data objects such as Euclidean are used.

$$ua_i = \frac{1}{|C_{l_i}|}\sum_{\boldsymbol{Y}\in C_{l_i}} pd(\boldsymbol{X_i}, \boldsymbol{Y}) \tag{10}$$

As it can be seen from Eq. (11), similar to the case for the original Silhouette index, ub_i is considered as the separation between object $\boldsymbol{X_i}$ and the closest cluster to it C_j, $C_j \neq C_{l_i}$. The separation between object $\boldsymbol{X_i}$ and cluster C_j is defined as the average pairwise distance between object $\boldsymbol{X_i}$ and all objects in cluster C_j. Again, the main difference between ub_i and b_i is that in ub_i pdms are used to capture the distance between objects, while in b_i distance measures for certain data objects are used.

$$ub_i = \min_j\left(\frac{1}{|C_j|}\sum_{\substack{\boldsymbol{Y}\in C_j\\ C_j \neq C_{l_i}}} pd(\boldsymbol{X_i}, \boldsymbol{Y})\right) \tag{11}$$

In this, we use Bhattacharyya as the pdm for computing the uncertain Silhouette index. Same as the original Silhouette, the optimal setting is the one that produces the largest index value and possibly the one that has the most compact and well-separated clusters.

4.2 OS index

In this section we propose a new clustering validity index for uncertain data objects, named Order Statistic (OS). The OS index can be considered as a general form of uncertain Dunn index, which is also developed in this paper. The OS index is composed of two components for capturing separation and compactness of clusters. It considers the average of r ($r > 1$) smallest inter-cluster distances for separation, and also the average of r ($r > 1$) largest intra-cluster distances for compactness. This enables the index to correctly detect the correct number of clusters in cases where there is either a very scattered cluster, or two very close clusters. These cases are the ones for which uncertain Dunn index ($r = 1$) fails in detecting the correct clusters. In a data set with K formed clusters, the maximum possible number of intra-cluster distances is K. In such a cluster, the maximum possible number of inter-cluster distances is $\frac{K(K-1)}{2}$. Since in the proposed validity index, r is used as both the number of considered inter-cluster and intra-cluster distances, it should be smaller than or equal to both K and $\frac{K(K-1)}{2}$. As in clustering we consider cases with at least two clusters, the choice of r should work for $K \geq 2$. All these can be written as:

$$r \leq \min\left(K, \frac{K(K-1)}{2}\right), \quad \text{for } K \geq 2. \tag{12}$$

Since we would like to consider the most complete information by taking into account the highest possible number of inter-cluster and intra-cluster distances for each K, $r=K-1$ would be the best choice. This can be written as: $K-1 = \max_r\left[r \leq \min\left(K, \frac{K(K-1)}{2}\right)\right]$.

The OS index is shown in Eq. (13).

$$OS_K = \frac{\sum_{i=1}^{r} sp_{(i)} \Big/ r}{\sum_{j=K-r+1}^{K} cp_{(j)} \Big/ r} \tag{13}$$

where $sp_{(i)}, i = 1, \ldots, \frac{K(K-1)}{2}$ is the i-th smallest order statistic of inter-cluster distances. The first order statistic of inter-cluster distances is $sp_{(1)} = \min_{\substack{1 \leq C_i, C_j \leq K \\ C_i \neq C_j}} \left[dist\left(C_i, C_j\right)\right]$. Here, for $dist\left(C_i, C_j\right)$, which denotes the distance between clusters C_i and C_j, we propose the average of s smallest pairwise probabilistic distances between objects in cluster C_i and objects in cluster C_j:

$$dist\left(C_i, C_j\right) = \frac{\sum_{\substack{k=1 \\ X \in C_i, Y \in C_j}}^{s} \left\{pd_{(k)}(\boldsymbol{X}, \boldsymbol{Y})\right\}}{s}, \tag{14}$$

where $s \leq |C_i|.\left|C_j\right|$. Capturing the distance between two clusters in this fashion has the advantage of being more robust to the existence of outlier values.

$cp_{(j)}, j = 1, \ldots K$, is the j-th smallest order statistic of intra-cluster distances. The K-th order statistic of intra-cluster distances is $cp_{(K)} = \max_{1 \leq C_m \leq K}[diam(C_m)]$. Here, $diam(C_m)$ denotes the diameter of cluster C_m and basically captures the compactness of the cluster. For $diam(C_m)$, we propose the average of t largest pairwise probabilistic distances between objects in cluster C_m:

$$diam(C_m) = \frac{\sum_{\substack{k=1\\ X,Y\in C_m}}^{t} \left\{pd_{(|C_m|^2-k+1)}(\boldsymbol{X},\boldsymbol{Y})\right\}}{t}, \tag{15}$$

where $t \leq |C_m|^2$. Capturing the diameters of clusters in this fashion has the advantage of being more robust to the existence of outlier values as well.

In the experiment section, we try different settings for the parameters s and t of the OS index. Generally, if the objects in the cluster are more uniformly scattered, smaller values of s and t are recommended. If the clusters are less uniformly scattered, higher values of s and t are suggested. If we choose $r = s = t = 1$, the OS index will reduce to an index that we call uncertain Dunn. Uncertain Dunn index is defined in Eq. (16).

$$UDU_K = \frac{sp_{(1)}}{cp_{(1)}} = \frac{\min\limits_{\substack{1\leq i,j\leq K\\ j\neq i}} \left(dist\left(C_i, C_j\right)\right)}{\max_{1\leq m\leq K}\{diam(C_m)\}} \tag{16}$$

In this index, $dist\left(C_i, C_j\right)$ and $diam(C_m)$ are defined based on Eq. (17) and Eq. (18).

$$dist\left(C_i, C_j\right) = pd_{(1)}(\boldsymbol{X},\boldsymbol{Y}) = \min_{\boldsymbol{X}\in C_i,\boldsymbol{Y}\in C_j}\{pd(\boldsymbol{X},\boldsymbol{Y})\} \tag{17}$$

$$diam(C_m) = pd_{\left(|C_m|^2\right)}(\boldsymbol{X},\boldsymbol{Y}) = \max_{\boldsymbol{X},\boldsymbol{Y}\in C_m}\{pd(\boldsymbol{X},\boldsymbol{Y})\} \tag{18}$$

Large values of the index indicate existence of compact and well-separated clusters.

One of the drawbacks of this uncertain Dunn index is its sensitivity to outlier values. Existence of outliers can highly affect Eqs. (17) and (18), and therefore the whole index. Another drawback of the uncertain Dunn index is its poor performance in the presence of either dominant small separation or large compactness values.

In the experiments section we show the capability of the OS and Silhouette indices over uncertain Dunn in overcoming these drawbacks through several experiments.

5 Experiments

The effectiveness of our proposed uncertain clustering validity indices is demonstrated through experiments on data sets with different sizes and dimensions. We conducted experiments on three two-dimensional synthetic data sets named SD1, SD2, and SD3 which had three, five, and three clusters respectively. The generated number of objects for each cluster of each data set was 50 but we also considered cases where 100, 200, and 500 objects where generated for each cluster. We also conducted experiments on higher dimensions of each data set including cases with two, three, five, and ten dimensions.

In addition, three two-dimensional data sets with major outliers were included in the experiments to show the robustness of our proposed validity indices in detecting the correct number of clusters of uncertain data objects.

The experiments include real-world data as well. A weather data set including the daily weather information of 1522 weather stations around the world for the year 2011, was considered. We also conducted experiments on two sets of images that were considered as uncertain objects.

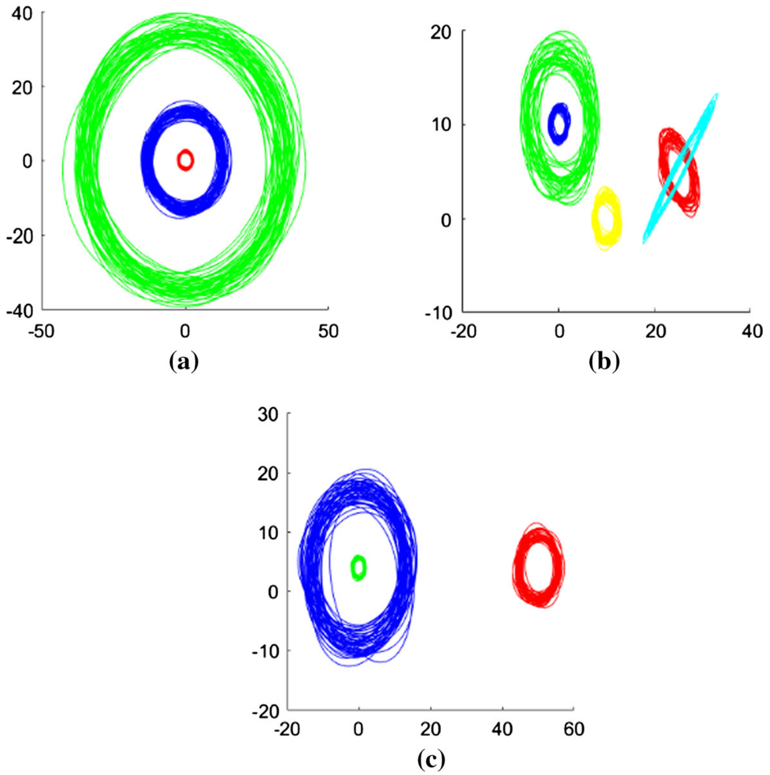


Fig. 2 Three two-dimensional synthetic data sets of uncertain data objects, **a** SD1, **b** SD2, and **c** SD3. The correct number of clusters for (**a**) to (**c**) are respectively 3, 5, and 3

Uncertain objects in each synthetic data set are modeled with multivariate normal distribution. To generate each uncertain object, first the mean point was generated based on a multivariate normal distribution and then the covariance matrix was generated based on an inverse Wishart distribution (Nydick 2012).

In this section, the performance of Dunn, Davies–Bouldin, Xie–Beni, Silhouette, uncertain Dunn, uncertain Silhouette, and the OS index with different parameters are compared.

5.1 Experiments with synthetic data sets

In this section we provide the experiments on the synthetic data sets. We considered three main two-dimensional data sets named SD1, SD2, and SD3. For each data set, different number of clusters was generated and each cluster contained 50 uncertain objects. Figure 2 shows the generated clusters of objects for each data set. Each ellipse in the figure represents an uncertain object modeled with a bivariate normal pdf. Different colors in the figure indicate different clusters. For SD1 and SD3, three, and for SD2, five clusters were generated. For SD1, the objects in all three clusters have overlapping mean points but they have different sizes of covariances. There are two sets of clusters with overlapping mean points in SD2 as well: one cluster shown in blue has smaller covariance than another one shown in green, and there are two clusters shown in red and cyan that are similar to the clusters shown in Fig. 1. For SD3, there are two overlapping clusters and a farther cluster.

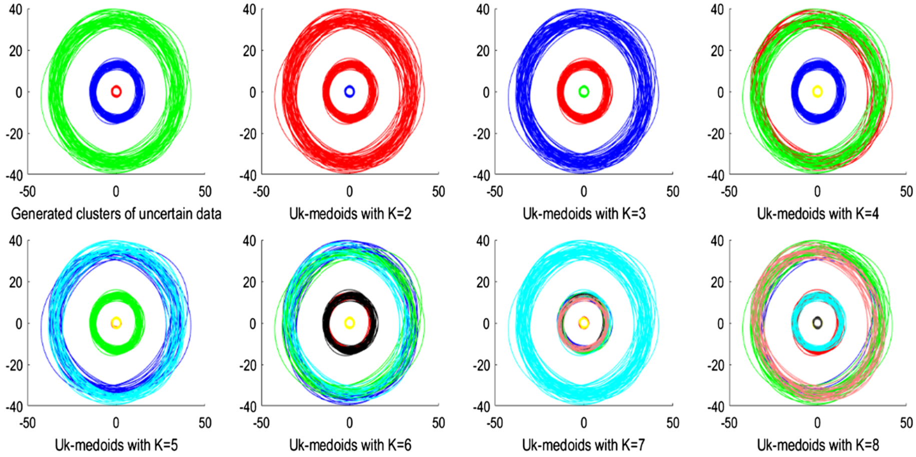


Fig. 3 The optimal formed clusters for *k,* $k = 2, 3, \ldots, 8$, after applying the uncertain K-medoids algorithm on the two-dimensional data set SD1

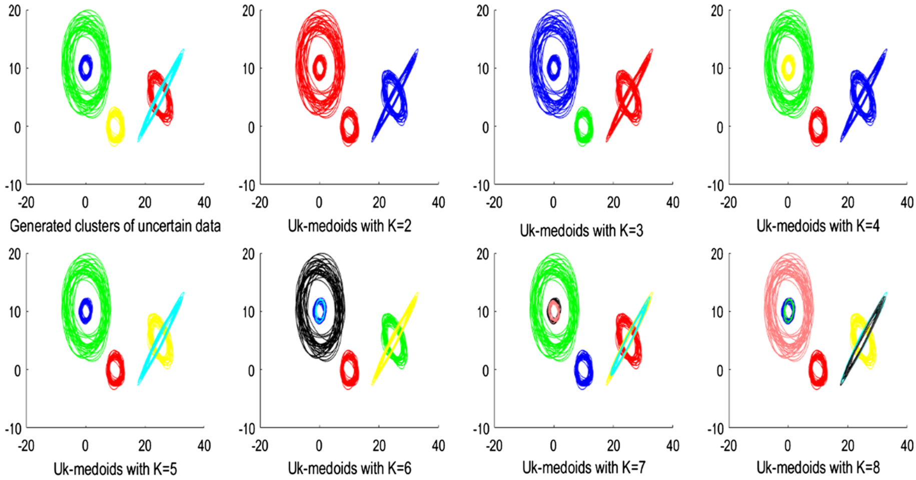


Fig. 4 The optimal formed clusters for *k,* $k = 2, 3, \ldots, 8$, after applying the uncertain K-medoids algorithm on the two-dimensional data set SD2

Figures 3, 4 and 5 show the optimal formed clusters for each *k,* $k = 2,3,\ldots,8$, after applying the uncertain K-medoids algorithm on SD1–SD3 respectively. Different colors in the figure indicate different clusters of objects again. As it can be also verified from the figures, the optimal number of clusters should be respectively 3, 5, 3 for SD1–SD3.

Tables 1, 2 and 3, contain the values of eight different indices: Dunn, Davies–Bouldin, Xie–Beni, Silhouette, uncertain Dunn (OS with $r = 1$, $s = t = 1$), uncertain Silhouette, OS with $s = t = 3$, and OS with $s = t = 5$ for SD1–SD3.

As it can be seen from Table 1, Dunn, Davies–Bouldin, Xie–Beni, and Silhouette, the four clustering validity indices for certain data objects that only use the mean of each object, fail in detecting the correct number of clusters for SD1, which is 3. However, it can be seen from the table that the developed clustering validity indices for uncertain data objects: uncertain Dunn,

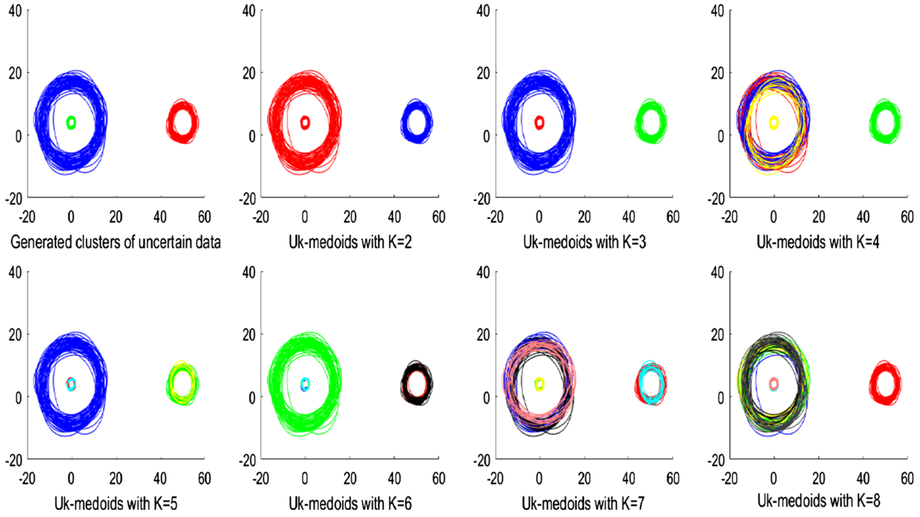


Fig. 5 The optimal formed clusters for *k, k*=2, 3,…, 8, after applying the uncertain K-medoids algorithm on the two-dimensional data set SD3

uncertain Silhouette, OS with, $s=t=3$, and OS with $s=t=5$ are all successful in detecting the correct number of clusters.

From Table 2, it can be seen that in addition to Dunn, Davies–Bouldin, Xie–Beni, and Silhouette, uncertain Dunn also fails in detecting the correct number of clusters for SD2, which is 5. The reason for that is large dominant compactness and small dominant separation values. Again, it can be seen from the table that uncertain Silhouette, OS with $s=t=3$, and OS with $s=t=5$ are all successful in detecting the correct number of clusters.

Same conclusions are valid for the results of Table 3. Again, Dunn, Davies–Bouldin, Xie–Beni, and Silhouette fail because of disability to capture the uncertain nature of the data objects, and uncertain Dunn also fails because of large dominant compactness and small dominant separation values. Uncertain Silhouette, OS with $s=t=3$, and OS with $s=t=5$, again successfully detect the correct number of clusters for SD3 which is 3.

As we mentioned earlier, for each data set: SD1, SD2, and SD3, we considered cases with higher number of objects and higher dimensions as well. Those include cases where 100, 200, and 500 were generated for each cluster in the data sets and cases where 3, 5, and 10 dimensions were considered.

Figure 6 shows the results of applying the different clustering validity indices on the SD1 data set with different number of objects. As it can be seen from the figure, although the existing clustering validity indices for certain data do not perform well in detecting three as the correct number of clusters, all the proposed clustering validity indices including uncertain Dunn show consistent behavior and successfully detect three clusters as the correct one for the SD1 data set with different number of objects. Figure 7 shows the results of applying the different clustering validity indices on the SD1 data set with different dimensions. Again, all the proposed validity indices perform well in detecting three as the correct number of clusters for the data sets with different dimensions while the existing validity indices show inconsistent behavior for different dimensions and mostly fail.

Figure 8 shows the results of applying the different clustering validity indices on the SD2 data set with different number of objects. As it can be seen from the figure, the existing

Table 1 Applying certain and uncertain clustering validity indices on the two-dimensional data set SD1

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=3$	OS $s=t=5$
2	**0.0029**	57.8018	19.0056	– 0.1014	1.8400	0.8933	1.7833	1.8400
3	0.0019	20.6739	13.8479	– 0.0697	**4.7321**	**0.9817**	**8.1492**	**9.2647**
4	0.0019	16.3579	25.6527	**– 0.0634**	0.1479	0.7700	3.4246	4.0633
5	0.0005	28.3027	199.6638	– 0.2966	0	0.4666	1.4602	1.7265
6	0.0019	10.1117	**5.3224**	– 0.1009	0.1062	0.5019	1.4644	1.7998
7	0.0010	**9.9790**	27.2011	– 0.5010	0.0214	0.4158	0.1059	0.1454
8	0.0005	21.1325	838.8978	– 0.4815	0	0.2677	0.2240	0.3109

Uncertain clustering validity indices are all successful in detecting the correct number of clusters while all the certain validity indices fail
The returned value by each index for the number of clusters that it detects as correct is shown in bold

Table 2 Applying certain and uncertain clustering validity indices on the two-dimensional data set SD2

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=3$	OS $s=t=5$
2	0.9075	0.3360	0.0006	0.8507	0.2569	0.8285	0.2483	0.2569
3	**4.6451**	**0.0593**	**0.0000**	**0.9975**	**16.2640**	0.9535	44.1609	47.3189
4	0.0092	0.7717	0.0171	0.7411	0.5816	0.9595	44.0676	48.0640
5	0.0092	3.8580	0.4700	0.3686	2.0522	**0.9614**	**57.3715**	**70.1076**
6	0.0092	3.7491	0.4940	0.5140	0.1609	0.8264	19.7402	23.6331
7	0.0018	4.9815	1.2939	0.1897	0	0.6832	1.9928	2.4392
8	0.0100	22.7409	67.6587	0.1742	0.1397	0.6926	2.0621	2.5426

Uncertain clustering validity indices uncertain Silhouette, OS with $s=t=3$, and OS with $s=t=5$ are all successful in detecting the correct number of clusters while all the certain validity indices fail

The returned value by each index for the number of clusters that it detects as correct is shown in bold

Table 3 Applying certain and uncertain clustering validity indices on the two-dimensional data set SD3

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=3$	OS $s=t=5$
2	**7.7244**	**0.0269**	**0.0000**	**0.9995**	**35.2860**	0.9710	34.3055	35.2860
3	0.0059	12.6706	2.1931	0.4564	4.9203	**0.9899**	**110.6491**	**136.6136**
4	0.0093	7.1799	1.0184	0.5371	0.0173	0.6208	0.0572	0.0710
5	0.0012	11.5813	5.2946	−0.0506	0	0.5502	1.7286	2.0808
6	0.0012	12.1038	19.2728	−0.4720	0	0.4898	1.4815	1.7834
7	0.0011	2.6580	2.7301	−0.0824	0.0759	0.2533	2.3242	2.9036
8	0.0025	6.2771	6.3955	0.2959	0.1178	0.5061	0.2339	0.4163

Uncertain clustering validity indices uncertain Silhouette, OS with $s=t=3$, and OS with $s=t=5$ are all successful in detecting the correct number of clusters while all the certain validity indices fail

The returned value by each index for the number of clusters that it detects as correct is shown in bold

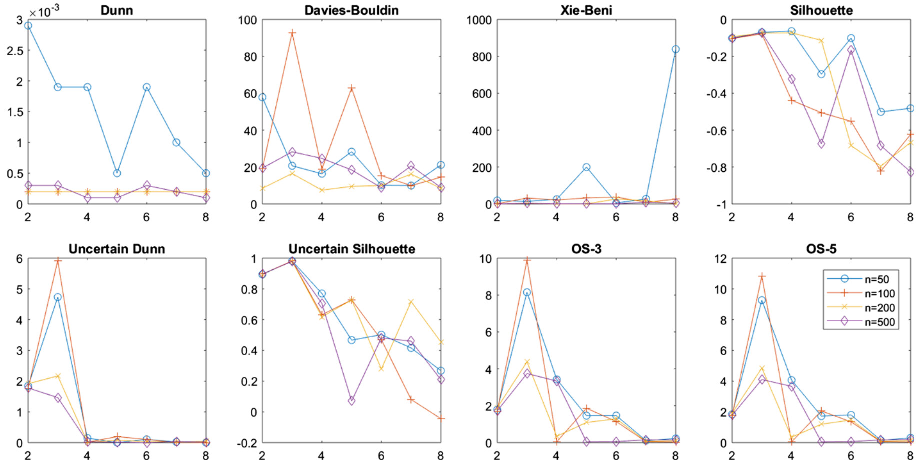


Fig. 6 Values of the studied indices with respect to *k, k*=2, 3,…, 8, for the SD1 data set with different number of objects

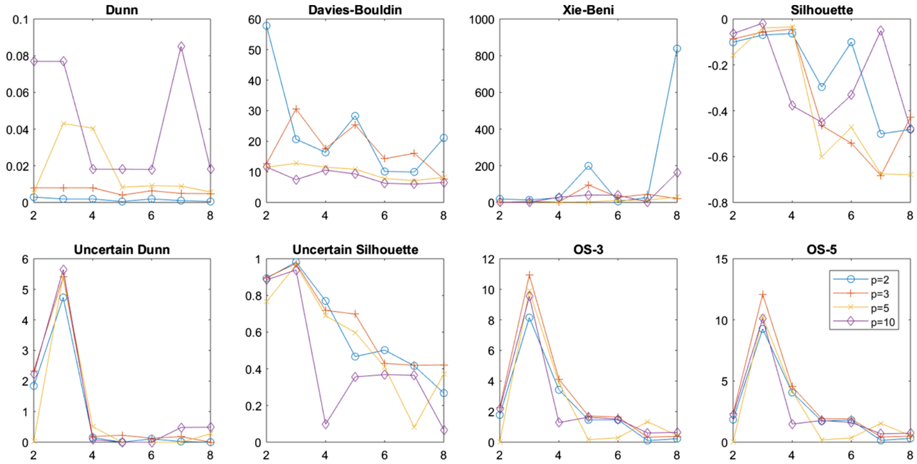


Fig. 7 Values of the studied indices with respect to *k, k*=2, 3,…, 8, for the SD1 data set with different dimensions

clustering validity indices for certain data along with the uncertain Dunn index do not perform well in detecting five as the correct number of cluster. Some of those indices wrongfully detect three as the correct number of clusters. However, all the proposed clustering validity indices show consistent behavior in detecting five as the correct number of clusters for the SD2 data set with different number of objects. Figure 9 shows the results of applying the different clustering validity indices on the SD2 data set with different dimensions. Again, all the proposed validity indices perform well in detecting five as the correct number of clusters for the SD2 data set with different dimensions while the existing validity indices and the uncertain Dunn index show inconsistent behavior and fail.

Figure 10 shows the results of applying the different clustering validity indices on the SD3 data set with different number of objects. As it can be seen from the figure, the existing clustering validity indices for certain data and the uncertain Dunn index do not detect three

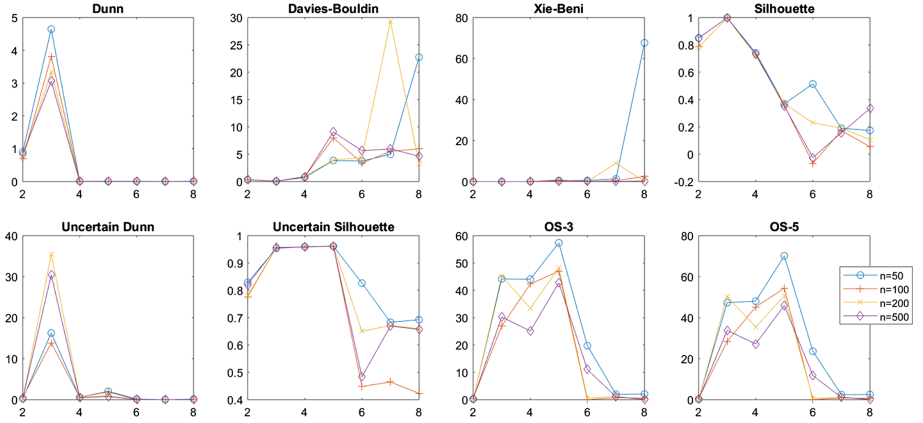


Fig. 8 Values of the studied indices with respect to *k*, *k* = 2, 3,…, 8, for the SD2 data set with different number of objects

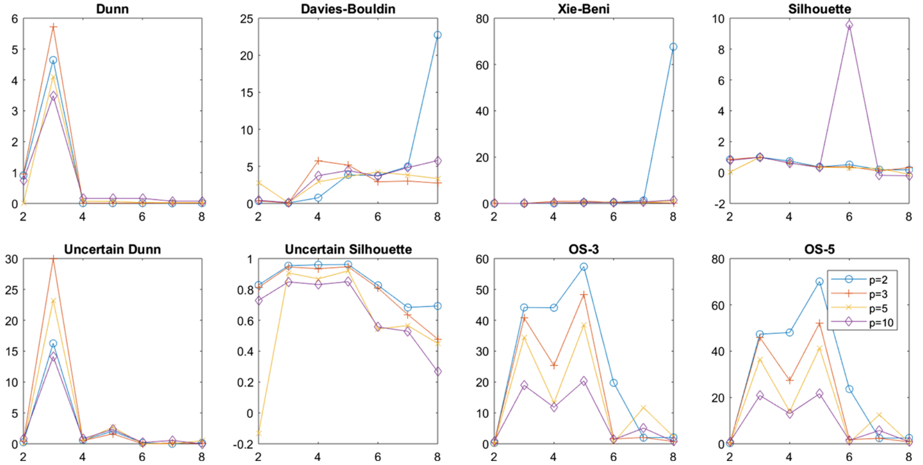


Fig. 9 Values of the studied indices with respect to *k*, *k* = 2, 3,…, 8, for the SD2 data set with different dimensions

as the correct number of clusters but the proposed clustering validity indices show consistent behavior and successfully detect three clusters as the correct one for the SD3 data set with different number of objects. Figure 11 shows the results of applying the different clustering validity indices on the SD3 data set with different dimensions. Again, all the proposed validity indices correctly detect three as the correct number of clusters for the data sets with different dimensions while the existing validity indices and the uncertain Dunn index show inconsistent behavior for different dimensions and mostly fail.

In addition to the three studied data sets, we conducted experiments on three more two-dimensional synthetic data sets named SD4, SD5, and SD6. These data sets can be seen in Fig. 12. As it can be seen from the figures, each data set contains a major outlier. The outliers of SD4 and SD5 are dashed green ellipses and the one for SD6 is a dashed blue ellipse.

Figures 13, 14 and 15 show the optimal formed clusters for each *k*, *k* = 2, 3,…,8, after applying the uncertain K-medoids algorithm on SD4–SD6 respectively. Again different colors

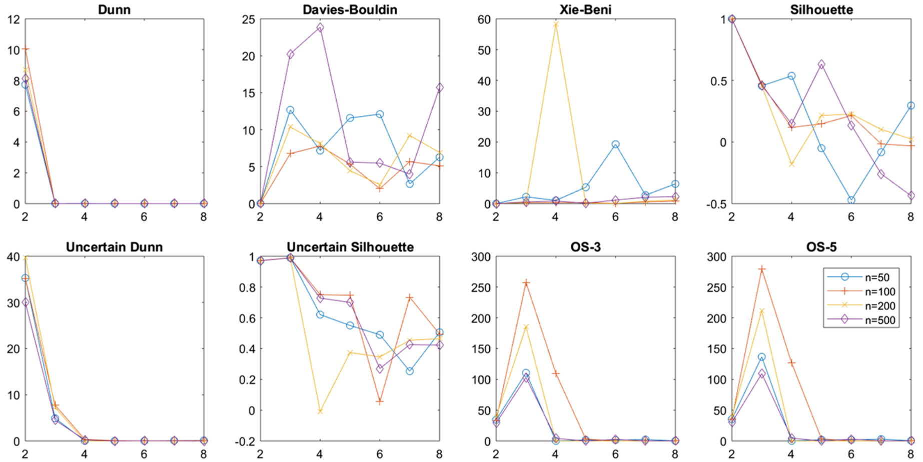


Fig. 10 Values of the studied indices with respect to *k, k*=2, 3,…, 8, for the SD3 data set with different number of objects

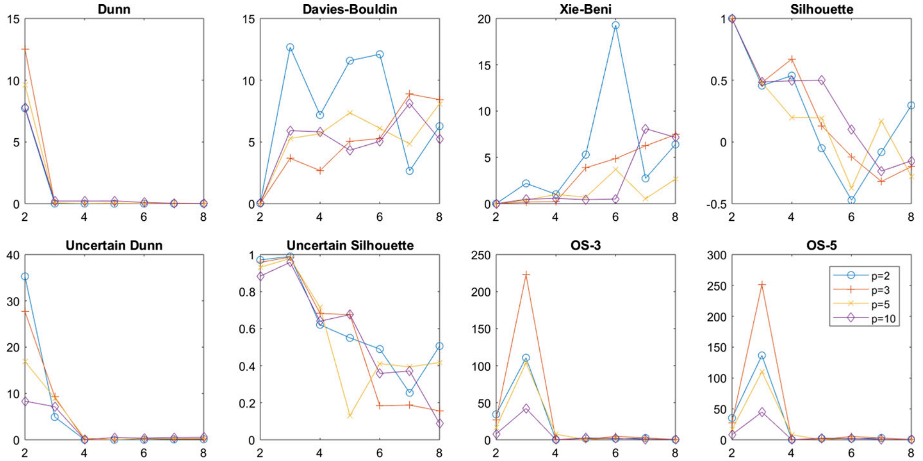


Fig. 11 Values of the studied indices with respect to *k, k*=2, 3,…, 8, for the SD3 data set with different dimensions

indicate different clusters of objects. As it can be also verified from the figures, the optimal number of clusters should be respectively 3, 2, 3 for SD4–SD6.

Table 4 contains the results for SD4 and has one more index compared to Tables 1, 2 and 3. That index is OS with $s=t=t=1$. The results demonstrate that uncertain Silhouette and OS with $s=t=5$ work well in the case of existing a major outlier. As it can be seen from the table, in addition to the validity indices for certain data objects and uncertain Dunn that fail in detecting the correct number of clusters, if OS with $s=t=1$ or $s=t=3$ are used, the correct number of clusters which is 3, is not detected and 2 clusters are detected as the correct number of clusters instead.

Table 5 results for SD5 also demonstrate that uncertain Silhouette and OS with $s=t=5$ work well in the case of existing a major outlier again. As it can be seen from the table, in addition to the validity indices for certain data objects in detecting the correct number of

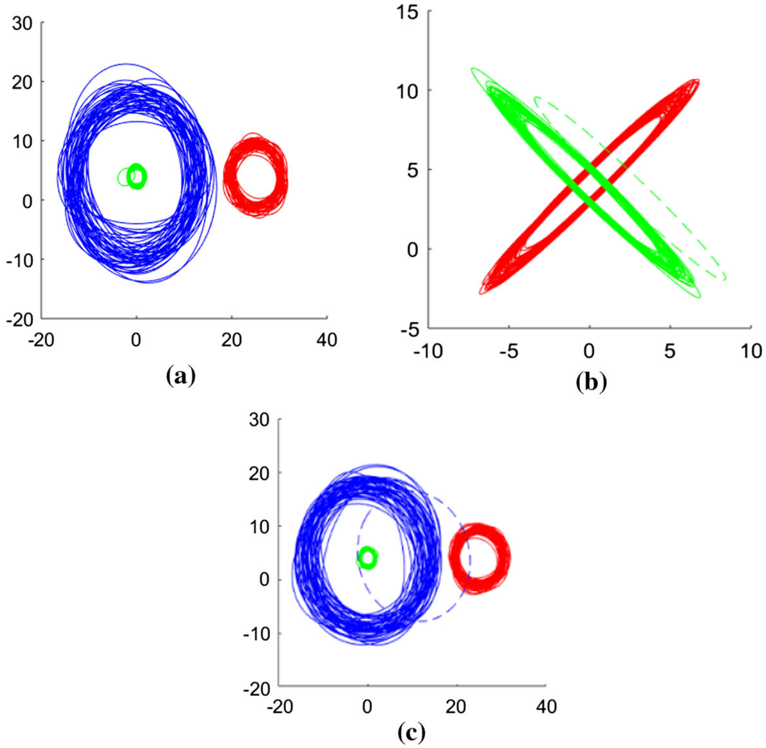


Fig. 12 three two-dimensional synthetic data sets of uncertain data objects with major outliers, **a** SD4, **b** SD5, and **c** SD6. The correct number of clusters for (**a**) to (**c**) are respectively 3, 2, and 3

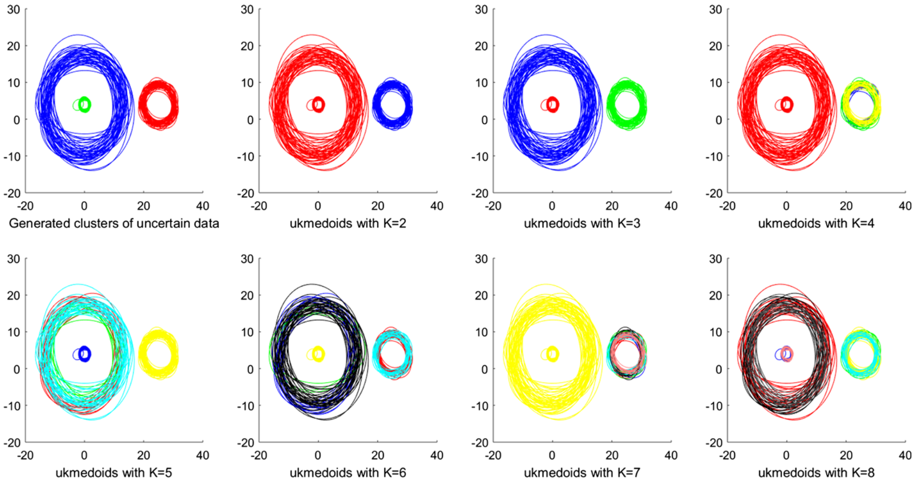


Fig. 13 The optimal formed clusters for *k, k*=2, 3,…, 8, after applying the uncertain K-medoids algorithm on the two-dimensional data set SD4

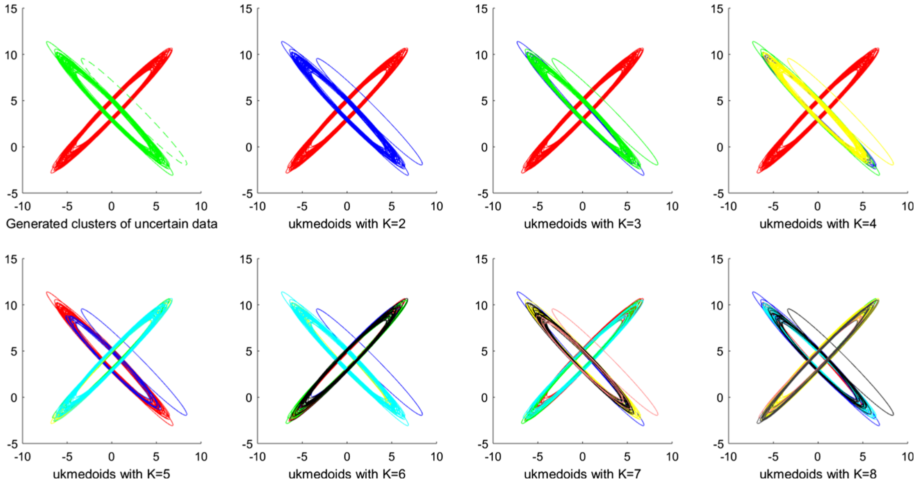


Fig. 14 The optimal formed clusters for *k*, *k* = 2, 3,…, 8, after applying the uncertain K-medoids algorithm on the two-dimensional data set SD5

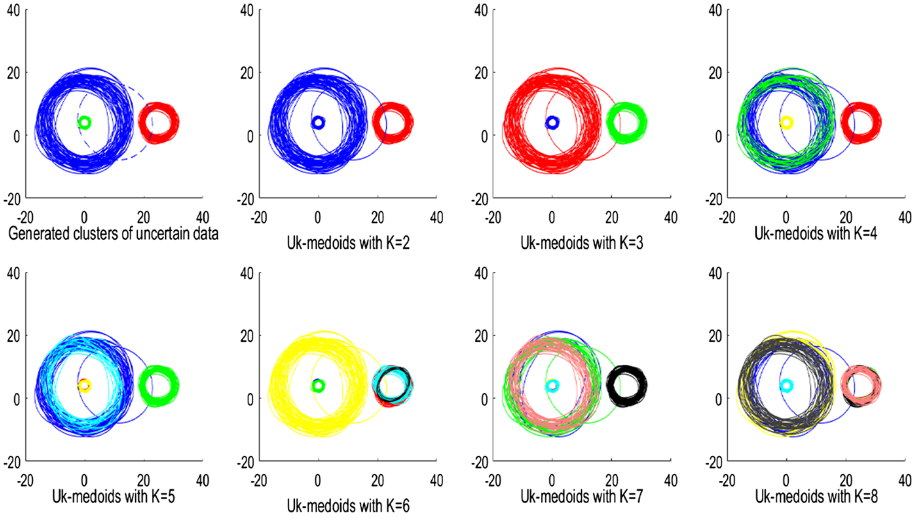


Fig. 15 The optimal formed clusters for *k*, *k* = 2, 3,…, 8, after applying the uncertain K-medoids algorithm on the two-dimensional data set SD4

clusters, if OS with $s = t = 1$ or $s = t = 3$ are used, the correct number of clusters which is 2 is not detected.

Finally, Table 6 results for SD6 demonstrate that uncertain Silhouette, OS with $s = t = 3$, and OS with $s = t = 5$ all work well in the case of an existing outlier. As it can be seen from the table, in addition to the validity indices for certain data objects and uncertain Dunn that fail in detecting the correct number of clusters, if OS with $s = t = 1$ is used, the correct number of clusters which is 3, is also not detected and 2 clusters are detected as the correct number of clusters instead.

Table 4 Applying certain and uncertain clustering validity indices on the two-dimensional data set SD4

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=1$	OS $s=t=3$	OS $s=t=5$
2	**6.4799**	**0.0518**	**0.0000**	**0.9983**	**9.4989**	0.8986	**8.3253**	**8.8975**	9.4989
3	0.0222	4.7583	0.2537	0.4544	1.1551	**0.9809**	2.9053	7.9952	**13.0676**
4	0.0029	2.0038	0.1160	0.7032	0.0182	0.6391	0.0499	0.0628	0.0752
5	0.0097	5.2450	1.0220	0.4789	0.0257	0.7056	0.2192	0.5833	0.9150
6	0.0031	6.7019	2.9211	0.1934	0.0224	0.4263	0.2565	0.6582	1.0274
7	0.0029	2.8873	0.2703	0.4968	0.0150	0.6641	0.0358	0.0664	0.0984
8	0.0029	6.3078	2.3841	−0.3768	0.0229	0.0977	0.0268	0.0934	0.1632

In addition to all the certain validity indices which fail in detecting the correct number of clusters, OS with $s=t=1$ and also OS with $s=t=3$ fail due to the existing outlier. Uncertain clustering validity indices uncertain Silhouette, OS with $s=t=5$ are both successful in detecting the correct number of clusters

The returned value by each index for the number of clusters that it detects as correct is shown in bold

Table 5 Applying certain and uncertain clustering validity indices on the two-dimensional data set SD5

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=1$	OS $s=t=3$	OS $s=t=5$
2	0.0026	28.6897	4.1184	**– 0.0023**	**0.5740**	**0.9674**	0.1123	0.3451	**0.5740**
3	0.0027	10.4394	**1.2829**	– 0.1046	0.0267	0.3973	**0.1138**	**0.3466**	0.5706
4	0.0014	16.2415	3.7877	– 0.4553	0.0061	0.3410	0.0046	0.0225	0.0475
5	**0.0030**	8.2306	2.8111	– 0.1111	0.0132	0.4454	0.0087	0.0351	0.0720
6	0.0020	20.0309	60.3606	– 0.1477	0.0220	0.2665	0.0427	0.1716	0.1983
7	0.0016	**6.2229**	3.1294	– 0.5326	0.0069	0.2518	0.0111	0.0381	0.0806
8	0.0017	7.0794	5.5323	– 0.6347	0.0144	0.1506	0.0247	0.0846	0.1730

In addition to most of the certain validity indices which fail in detecting the correct number of clusters, Silhouette, OS with $s=t=1$, and also OS with $s=t=3$ fail due to the existing outlier. Uncertain clustering validity indices uncertain Silhouette, OS with $s=t=5$ are both successful in detecting the correct number of clusters

The returned value by each index for the number of clusters that it detects as correct is shown in bold

Table 6 Applying certain and uncertain clustering validity indices on the two-dimensional data set SD6

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=1$	OS $s=t=3$	OS $s=t=5$
2	**1.1663**	**0.0721**	**0.0000**	**0.9939**	**8.3014**	0.8868	**7.9472**	8.1401	8.3014
3	0.0085	5.5387	0.4185	0.4156	3.2962	**0.9850**	4.5835	**16.3612**	**25.5424**
4	0.0080	3.0342	0.1601	0.4479	0.0177	0.6729	1.5203	2.9987	4.0152
5	0.0010	3.2160	1.7455	0.1968	0	0.5173	0.5910	1.2890	1.8138
6	0.0010	5.3456	− 0.3045	0.0625	0.4267	0.4267	0.5919	1.2668	1.7419
7	0.0010	3.8207	20.9575	0.0819	0	0.4341	0.0613	0.1793	0.2975
8	0.0016	4.3963	3.4262	0.1286	0.0524	0.4476	0.0587	0.1811	0.2991

In addition to all the certain validity indices which fail in detecting the correct number of clusters, OS with $s=t=1$ also fails due to the existing outlier. Uncertain clustering validity indices uncertain Silhouette, OS with $s=t=3$, and OS with $s=t=5$ are all successful in detecting the correct number of clusters

The returned value by each index for the number of clusters that it detects as correct is shown in bold

In conclusion, the results of the experiments on SD4, SD5, and SD6 show that uncertain Silhouette and the OS index with larger values of s and t can be more reasonable for dealing with outliers.

5.2 Experiments with real data sets

In this section, we provide experiments on two sets of real data. The first set of experiments are on a weather data set and the second set is on two sets of images.

5.2.1 The weather data set

The weather data set in this paper is a data set that was collected from the National Center for Atmospheric Research data archive (https://rda.ucar.edu/datasets/ds512.0/). The collected data set contains the daily weather information (average temperature and precipitation level) of 1522 weather stations around the world for the year 2011. Each station in this data set, can be considered as an uncertain object with 365 two-dimensional points. Based on Kӧppen–Geiger climate classification (Peel et al. 2007), these stations are of five climate types: polar, cold, temperate, tropical, and dry. Figure 16 demonstrates examples of stations from the five climate types.

We performed the uncertain K-medoids algorithm with Bhattacharyya pdm on the weather data set with $k=2, 3,\ldots,8$. For each k, we ran the algorithm 10 times and compared the performance of nine indices: Dunn, Davies–Bouldin, Xie–Beni, Silhouette, Uncertain Dunn, Uncertain Silhouette, OS with $s=t=3$, OS with $s=t=5$, and OS with $s=t=10$. The numbers for each particular k in Table 7, demonstrate the best results out of the 10 runs for each index.

As it can be seen from the table, our developed uncertain clustering validity indices, uncertain Silhouette and OS perform very well in detecting the correct number of clusters (five). The four clustering validity indices for certain data, i.e. Dunn, Davies–Bouldin, Xie–Benie, and Silhouette fail in detecting the correct number of clusters. Also, we can see that Uncertain Dunn, which is a simple case of the OS algorithm, fails, possibly because of its sensitivity to outlier values or either dominant separation values, or compactness values. The values of the nine indices with respect to the number of clusters k, $k=2, 3,\ldots,8$, are plotted in Fig. 17a. From the figure, it can be seen that only the developed clustering validity indices for uncertain data uncertain Silhouette, OS with $s=t=3$, OS with $s=t=5$, and OS with $s=t=10$ produce sharp peaks for the correct number of clusters which is five.

For the weather data set, we also considered modeling uncertain objects with bivariate normal pdfs. For each weather station we fitted a bivariate normal pdf to the 365 two-dimensional points. Figure 18 demonstrates examples of stations from the five climate types, modeled with pdfs. Similar to the experiment on the weather data set considered with points, we performed the uncertain K-medoids algorithm with Bhattacharyya pdm with $k=2, 3,\ldots,8$. For each k, we ran the algorithm 10 times and compared the performance of the nine indices. Again, the numbers for each particular k in Table 8, demonstrate the best results out of the 10 runs for each index.

As it can be seen from the table, again the developed uncertain clustering validity indices perform very well in detecting the correct number of clusters. The four clustering validity indices for certain data fail in detecting the correct number of clusters. The values of the nine indices with respect to the number of clusters k, $k=2, 3,\ldots,8$, are plotted in Fig. 17b. Again, from the figure, it can be seen that only the developed clustering validity indices for

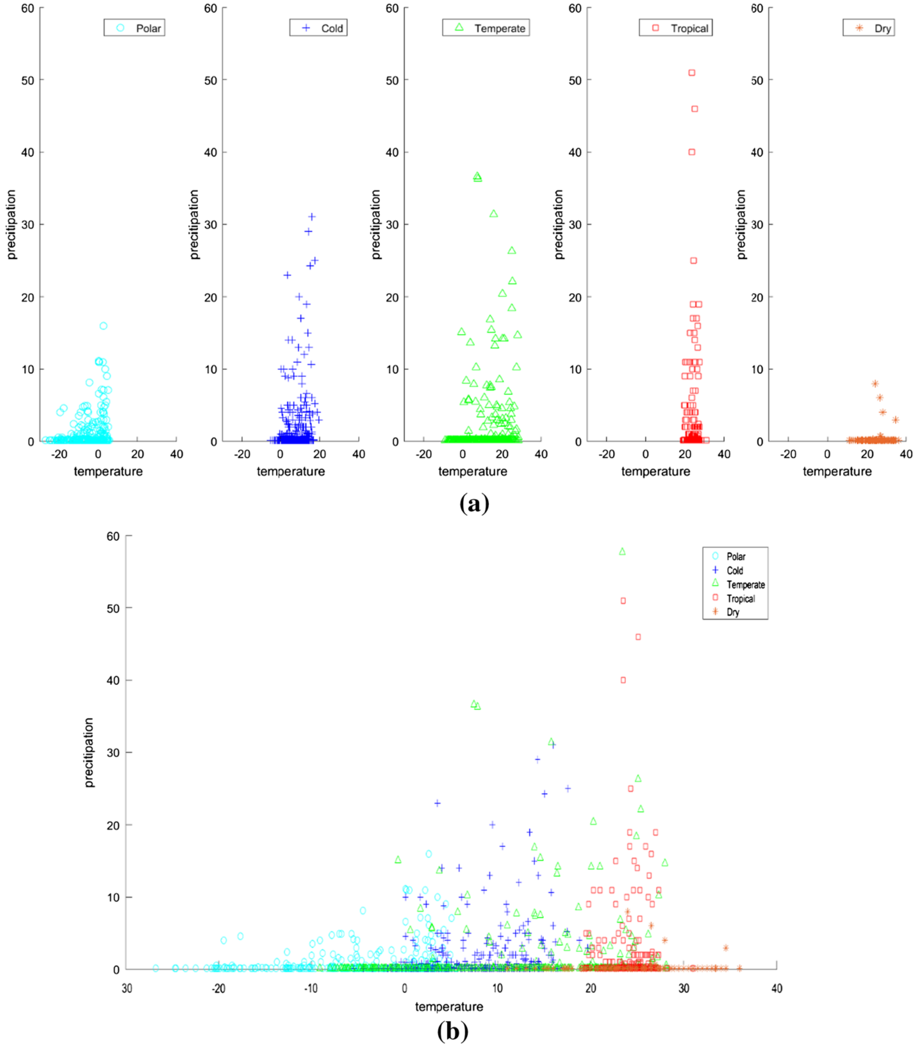


Fig. 16 Examples of stations from the five climate types: polar, cold, temperate, tropical, dry, **a** plotted separately, **b** plotted together

uncertain data uncertain Silhouette, OS with $s=t=3$, OS with $s=t=5$, and OS with $s=t=10$ produce sharp peaks for the correct number of clusters which is five.

Comparing the results of the two cases of modeling the weather data set with points and pdfs shows that the performance of the indices does not change substantially. The certain validity indices perform almost the same as they only use the same single statistic i.e. the mean point of each object. The developed validity indices for uncertain objects are more different in the two cases.

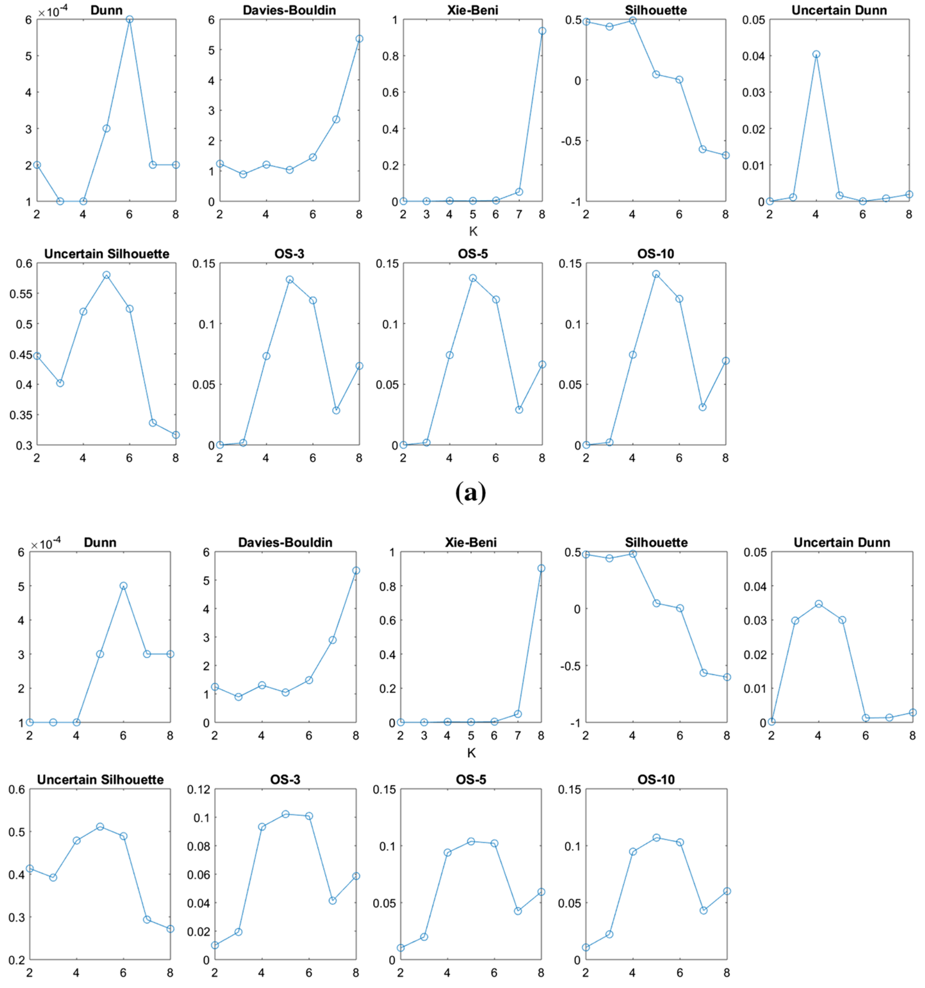
For the case with data considered with points, the magnitude of the peaks for indices values are larger compared to the magnitude of the peaks in the data set modeled with pdfs. Also, for the case modeled with points, the peaks are sharper which is more desirable. These can be seen from Fig. 19.

Table 7 Applying certain and uncertain clustering validity indices on the weather data set

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $r=K-1$ $s=t=3$	OS $r=K-1$ $s=t=5$	OS $r=K\ -1$ $s=t=10$
2	0.0002	1.2389	0.0005	**0.4795**	0	0.4466	0	0	0
3	0.0001	**0.8873**	**0.0004**	0.4387	0.0011	0.4017	0.0017	0.0019	0.0021
4	0.0001	1.2076	0.0029	0.4899	**0.0404**	0.5197	0.0732	0.0740	0.0743
5	0.0003	1.0342	0.0025	0.0452	0.0016	**0.5802**	**0.1363**	**0.1375**	**0.1408**
6	**0.0006**	1.4505	0.0041	0.0025	0	0.5244	0.1191	0.1198	0.1204
7	0.0002	2.6987	0.0518	− 0.5712	0.0008	0.3363	0.0284	0.0290	0.0311
8	0.0002	5.3586	0.9361	− 0.6211	0.0019	0.3165	0.0651	0.0663	0.0693

Uncertain clustering validity indices uncertain Silhouette, OS with $s=t=3$, and OS with $s=t=5$, and OS with $s=t=10$ successfully detect the correct number of clusters which is five while others fail

The returned value by each index for the number of clusters that it detects as correct is shown in bold



(a)

(b)

Fig. 17 Values of the studied indices with respect to *k*, *k* = 2, 3,…, 8, for the weather data set **a** considered with points, **b** modeled with pdfs. The developed clustering validity indices for uncertain data uncertain Silhouette, OS with $s = t = 3$, OS with $s = t = 5$, and OS with $s = t = 10$ produce sharp peaks for the correct number of clusters

It is also notable from Tables 6 and 7 that in both cases, for the OS index, as the values of s and t increase, the magnitude of the values along with the sharpness of the peaks increase and the correct number of clusters can be detected more precisely. Overall, on this data set, considering the data with points is more accurate than modeling them with bivariate normal pdfs.

5.2.2 The image data sets

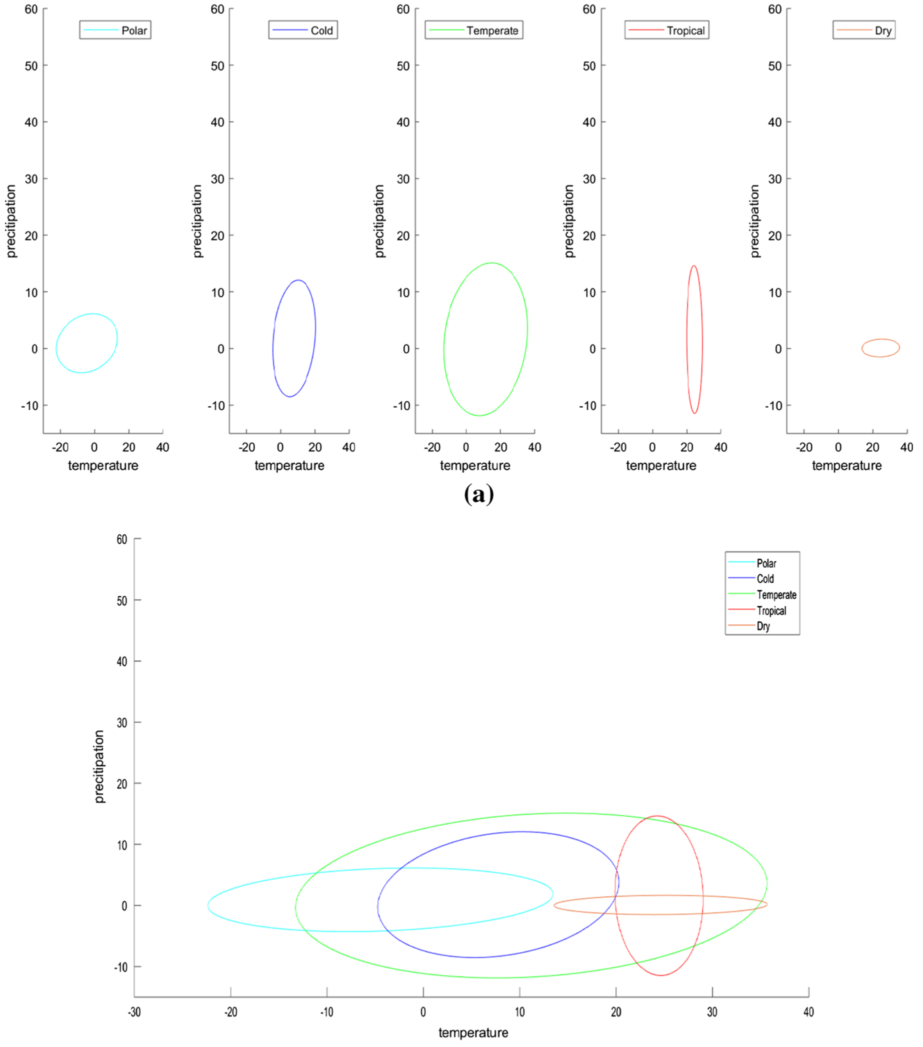
Images are the type of objects that can be considered as uncertain data objects as well. Given an image with at least one object, each image object can be converted to a group of two-dimensional points. Then the whole image can be considered as the combination of all the points from all the objects in the image.

Table 8 Applying certain and uncertain clustering validity indices on the weather data set modeled with pdfs

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=3$	OS $s=t=5$	OS $s=t=10$
2	0.0001	1.2491	0.0006	**0.4751**	0.0002	0.4133	0.0101	0.0103	0.0107
3	0.0001	**0.8943**	**0.0003**	0.4411	0.0298	0.3921	0.0194	0.0199	0.0222
4	0.0001	1.3067	0.0033	0.4809	**0.0347**	0.4788	0.0933	0.0940	0.0948
5	0.0003	1.0524	0.0024	0.0462	0.0300	**0.5114**	**0.1022**	**0.1038**	**0.1071**
6	**0.0005**	1.4832	0.0040	0.0031	0.0013	0.4891	0.1009	0.1021	0.1030
7	0.0003	2.8944	0.0491	−0.5647	0.0014	0.2936	0.0414	0.0426	0.0431
8	0.0003	5.3370	0.9026	−0.6026	0.0029	0.2718	0.0587	0.0595	0.0601

Again, uncertain clustering validity indices uncertain Silhouette, OS with $s=t=3$, and OS with $s=t=5$, and OS with $s=t=10$ successfully detect the correct number of clusters which is five while others fail

The returned value by each index for the number of clusters that it detects as correct is shown in bold



(a)

(b)

Fig. 18 Examples of stations modeled with pdf from the five climate types: polar, cold, temperate, tropical, dry, **a** plotted separately, **b** plotted together

We conducted experiments by considering two sets of images. The first set named Crosswords includes three images. The second set named Cards includes four images. Figures 20 and 21 show the images in the first and second sets respectively. In our experiments, first we scaled the resolution of all the images to $100 * 100$ pixels. Then we considered each image as an uncertain object by converting it to a sample of 100 two-dimensional points normalized to be between 0 and 1. Fifty replicates of each image (or uncertain object) were created by adding random numbers in $[-0.05, 0.05]$ to each dimension of the original images. Essentially, 150 image replicates or uncertain objects were created for the Crosswords set and 200 were created for the Cards set. Next, we performed clustering with different values of k, $k=2, 3, \ldots, 8$ by using the uncertain K-medoids algorithm. As it can be noted, three should

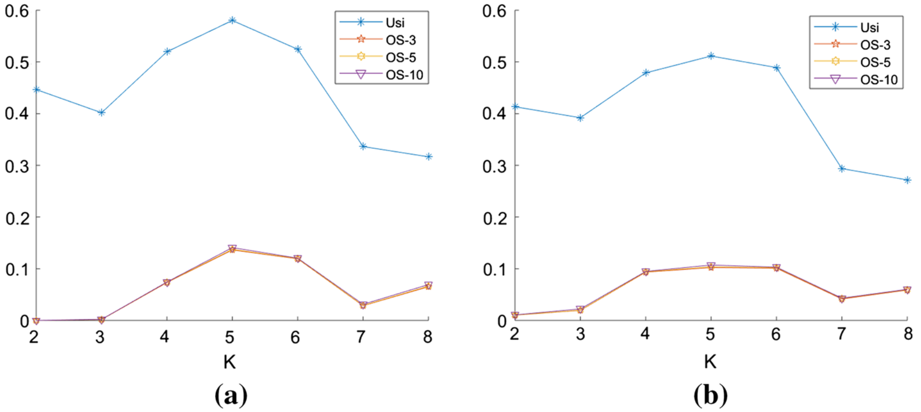


Fig. 19 Values of the proposed indices with respect to *k*, *k* = 2, 3,…, 8, for the weather data set **a** considered with points **b** with pdfs

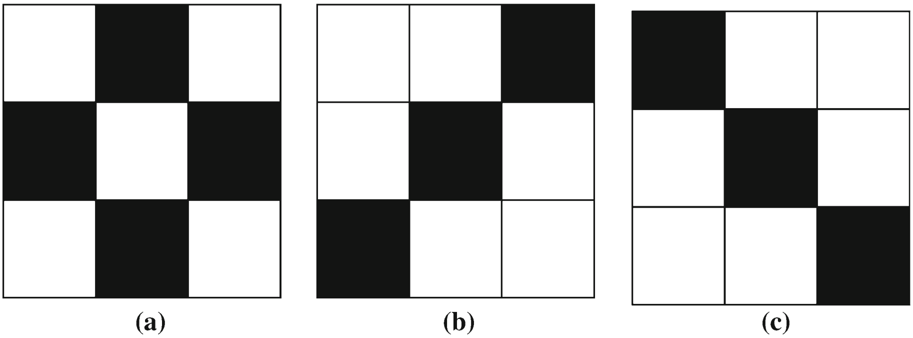


Fig. 20 The three crossword images

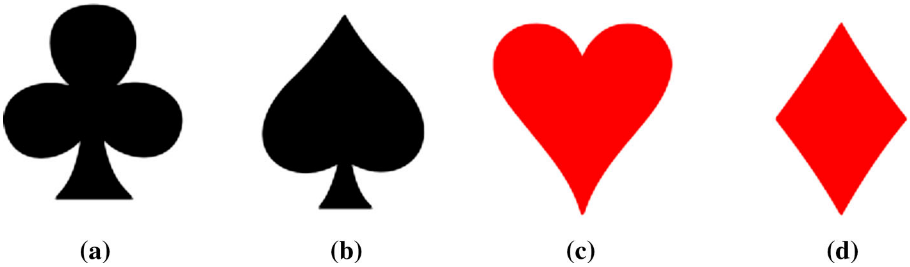


Fig. 21 The three card images

be identified as the correct number of clusters for the Crosswords set and four should be identified as the correct number of clusters for the Cards set. The results of applying the validity indices are reported in Tables 9 and 10. As it can be seen once again, for both the Crosswords and Cards sets, the existing clustering validity indices for certain data and the uncertain Dunn index fail in detecting the correct numbers of clusters which are three and four respectively. However, as it can be seen, all the proposed indices successfully detect the correct number of clusters in both cases.

Table 9 Applying certain and uncertain clustering validity indices on the Crosswords set

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=3$	OS $s=t=5$	OS $s=t=10$
2	**2.7433**	**0.1940**	**0.0023**	**0.9801**	**8.1013**	0.7461	1.4466	1.4466	1.4467
3	0.4337	0.2840	0.0042	0.9144	1.0064	**0.9923**	**79.7271**	**98.6653**	**103.0141**
4	0.1641	0.5967	0.2652	0.7861	0.9185	0.9507	71.6272	87.1223	96.2333
5	1.3096	1.0059	0.1442	0.6758	0.9023	0.7408	58.4279	60.9654	63.9215
6	0.9138	1.2377	0.2118	0.6923	0.8817	0.7029	55.2414	57.9566	60.0447
7	1.4981	0.2554	0.0028	0.7891	0.7014	0.6615	51.1777	54.8188	57.5005
8	2.5926	0.2716	0.0025	0.9024	0.5039	0.6424	50.0201	52.6039	55.4862

Uncertain clustering validity indices uncertain Silhouette, OS with $s=t=3$, OS with $s=t=5$, and OS with $s=t=10$ successfully detect the correct number of clusters which is five while others fail

The returned value by each index for the number of clusters that it detects as correct is shown in bold

Table 10 Applying certain and uncertain clustering validity indices on the Cards set

K	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=3$	OS $s=t=5$	OS $s=t=10$
2	**5.9967**	**0.0444**	**0.0005**	**0.9984**	**9.9708**	0.5513	0.9344	0.9369	0.9392
3	0.1046	0.9478	0.0729	0.3559	4.3133	0.7576	1.8991	1.9115	1.9323
4	1.0631	0.3149	0.0048	0.8447	1.2565	**0.9914**	**79.0935**	**119.4419**	**145.6607**
5	0.3536	0.5827	0.1523	0.7222	0.9124	0.8071	41.6637	63.7608	81.0969
6	0.2073	0.8099	0.1760	0.6816	0.8994	0.7606	76.1157	90.4836	104.8198
7	0.1723	0.8628	0.1939	0.6561	0.7822	0.7337	39.2816	57.0948	74.4646
8	0.0919	0.9710	0.2651	0.6217	0.5361	0.6741	30.7455	47.7212	69.9669

Uncertain clustering validity indices uncertain Silhouette, OS with $s=t=3$, OS with $s=t=5$, and OS with $s=t=10$ successfully detect the correct number of clusters which is five while others fail

The returned value by each index for the number of clusters that it detects as correct is shown in bold

Table 11 Summary of the performance of the studied clustering validity indices on all the data sets

Data set	True clusters	Dunn	Davies–Bouldin	Xie–Beni	Silhouette	Uncertain Dunn	Uncertain Silhouette	OS $s=t=3$	OS $s=t=5$
SD1	3	1	7	6	4	**3**	**3**	**3**	**3**
SD2	5	3	3	3	3	3	**5**	**5**	**5**
SD3	3	2	2	2	2	2	**3**	**3**	**3**
Weather	5	6	3	3	2	4	**5**	**5**	**5**
Crosswords	2	2	2	2	2	2	**3**	**3**	**3**
Cards	2	2	2	2	2	2	**4**	**4**	**4**

The developed uncertain clustering validity indices successfully detect the correct number of clusters for all the studied data sets while uncertain Dunn is only successful for one data set and the certain clustering validity indices fail for all the data sets

For each data set, the correct number of clusters detected by each index is shown in bold if it matches with the true number of clusters

Table 11 shows a summary of the performance of the studied clustering validity indices on all the data sets. As it can be seen from the table, our proposed clustering validity indices for uncertain data objects, i.e. uncertain Silhouette and OS are both successful in detecting the correct number of clusters for all the data sets. Uncertain Dunn which is a reduced and simplified form of the OS is only successful for one data set (SD1), while all the clustering validity indices for certain data objects, i.e. Dunn, Davies–Bouldin, Xie–Beni, and Silhouette, fail to detect the correct number of clusters for all data sets.

6 Conclusion

In this paper, we proposed two clustering validity indices, named uncertain Silhouette and Order Statistics index (OS), for validation of clusters of uncertain data objects. To our best knowledge, prior to this work, there was not any clustering validity indices in the literature, designed to handle uncertain objects given in forms of multiple points or probability density functions. Both proposed indices outperform existing certain clustering validity indices in validating clusters of uncertain data objects.

The advantage of the uncertain Silhouette index over the OS index is that it does not depend on any parameters such as t and s that the OS index uses. Also, the uncertain Silhouette index considers the contribution of every single object to compactness and separation of clusters and since it uses scaled values for every object's contribution (a value between -1 and 1), it is very robust to outliers.

The advantage of the OS index over the uncertain Silhouette index is that for the inter-cluster and intra-cluster distances it only uses the averages of the s smallest and t largest distances respectively rather than using all the objects. Sometimes, the key characteristics of clusters are determined by only a few objects and considering other unimportant objects might fade away the contribution of the key objects and weaken the performance of the index.

The advantage of the OS index over uncertain Dunn is that it is the general case of uncertain Dunn and is capable of correctly detecting the correct number of clusters in cases where there is either a very large dominant compactness value (a very spread cluster), or there is a small dominant separation value (two very close clusters).

The effectiveness of our developed indices was evaluated through several experiments on synthetic and real data sets.

Besides the two developed clustering validity indices in this paper, more indices for uncertain data objects can be significant in conducting a comprehensive validation of clusters of uncertain data objects. In the future, we will work on developing more uncertain data clustering validity indices.

Acknowledgements The authors would like to thank the editor and anonymous reviewers for their valuable comments and suggestions which helped to improve the quality of this paper.

References

Aggarwal, C. C., & Philip, S. Y. (2009). A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering, 21*(5), 609–623.

Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal Processing, 18*(4), 349–369.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics, 7*(4), 401–406.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods, 3*(1), 1–27.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City, 1*(2), 1.

Chau, M., Cheng, R., Kao, B., & Ng, J. (2006). Uncertain data mining: An example in clustering location data. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 199–204). Springer.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics, 23*(4), 493–507.

Chiang, M.-C., Tsai, C.-W., & Yang, C.-S. (2011). A time-efficient pattern reduction algorithm for k-means clustering. *Information Sciences, 181*(4), 716–731.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. Hoboken: Wiley.

Csiszar, I., & Körner, J. (2011). *Information theory: Coding theorems for discrete memoryless systems*. Cambridge: Cambridge University Press.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2,* 224–227.

Devijver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach*. Upper Saddle River: Prentice Hall.

Duan, L., Xu, L., Liu, Y., & Lee, J. (2009). Cluster-based outlier detection. *Annals of Operations Research, 168*(1), 151–168.

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

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal, 41*(8), 578–588.

Gullo, F., Ponti, G., & Tagarelli, A. (2008a). Clustering uncertain data via k-medoids. In *Proceedings of the 2nd international conference on scalable uncertainty management, ser. SUM'08* (pp. 229–242). Berlin: Springer.

Gullo, F., Ponti, G., Tagarelli, A., & Greco, S. (2008b). A hierarchical algorithm for clustering uncertain data via an information-theoretic approach. In *Data mining, 2008. ICDM'08. Eighth IEEE international conference on* (pp. 821–826). IEEE.

Gullo, F., Ponti, G., & Tagarelli, A. (2010). Minimizing the variance of cluster mixture models for clustering uncertain objects. In *Data mining (ICDM), 2010 IEEE 10th international conference on* (pp. 839–844). IEEE.

Gullo, F., Ponti, G., & Tagarelli, A. (2013). Minimizing the variance of cluster mixture models for clustering uncertain objects. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 6*(2), 116–135.

Gullo, F., Ponti, G., Tagarelli, A., & Greco, S. (2017). An information-theoretic approach to hierarchical clustering of uncertain data. *Information Sciences, 402,* 199–215.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*(2), 107–145.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28*(1), 100–108.

Jiang, B., Pei, J., Tao, Y., & Lin, X. (2013). Clustering uncertain data based on probability distribution similarity. *IEEE Transactions on Knowledge and Data Engineering, 25*(4), 751–763.

Kao, B., Lee, S. D., Lee, F. K., Cheung, D. W., & Ho, W.-S. (2010). Clustering uncertain data using voronoi diagrams and r-tree index. *IEEE Transactions on Knowledge and Data Engineering, 22*(9), 1219–1233.

Kriegel, H.-P., & Pfeifle, M. (2005). Density-based clustering of uncertain data. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 672–677). ACM.

Lee, S. D., Kao, B., & Cheng, R. (2007). Reducing UK-means to K-means. In *Data mining workshops, 2007. ICDM workshops 2007. Seventh IEEE international conference on* (pp. 483–488). IEEE.

Marinakis, Y., Marinaki, M., Doumpos, M., Matsatsinis, N., & Zopounidis, C. (2011). A hybrid ACO-GRASP algorithm for clustering analysis. *Annals of Operations Research, 188*(1), 343–358.

Nydick, S. (2012). *The wishart and inverse wishart distributions*. http://www.tc.umn.edu/~nydic001/docs/unpubs/WishartDistribution.pdf. Accessed 21 Mar 2017.

Pakhira, M. K., Bandyopadhyay, S., & Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition, 37*(3), 487–501.

Pakhira, M. K., Bandyopadhyay, S., & Maulik, U. (2005). A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy Sets and Systems, 155*(2), 191–214.

Pal, N. R., & Biswas, J. (1997). Cluster validation using graph theoretic concepts. *Pattern Recognition, 30*(6), 847–857.

Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen–Geiger climate classification. *Hydrology and Earth System Sciences Discussions, 4*(2), 439–473.

Qin, B., Xia, Y., & Li, F. (2009). DTU: A decision tree for uncertain data. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 4–15). Berlin: Sringer.

Qin, Z., Wan, T., & Zhao, H. (2017). Hybrid clustering of data and vague concepts based on labels semantics. *Annals of Operations Research, 256*(2), 393–416.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20,* 53–65.

Tavakkol, B. (2018). Data Mining methodologies with uncertain data (Doctoral dissertation, Rutgers University-School of Graduate Studies-New Brunswick).

Tavakkol, B., Jeong, M. K., & Albin, S. L. (2017). Object-to-group probabilistic distance measure for uncertain data classification. *Neurocomputing, 230,* 143–151.

Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(8), 841–847.

Yang, B., & Zhang, Y. (2010). Kernel based K-medoids for clustering data with uncertainty. In *International conference on advanced data mining and applications* (pp. 246–253). Berlin: Springer.

Zhou, S., & Chellappa, R. (2004). *Probabilistic distance measures in reproducing kernel Hilbert space*. SCR Technical Report, University of Maryland.