A Hierarchical Algorithm for Clustering Uncertain Data via an Information-Theoretic Approach

Francesco Gullo Giovanni Ponti Andrea Tagarelli Sergio Greco

DEIS - University of Calabria
Via P. Bucci 41c, Arcavacata di Rende (CS) I87036, Italy
{fgullo,gponti,tagarelli,greco}@deis.unical.it

Abstract

In recent years there has been a growing interest in clustering uncertain data*. In contrast to traditional, "sharp" data representation models, uncertain data objects can be represented in terms of an uncertainty region over which a probability density function (pdf) is defined. In this context, the focus has been mainly on partitional and density-based approaches, whereas hierarchical clustering schemes have drawn less attention.*

We propose a centroid-linkage-based agglomerative hierarchical algorithm for clustering uncertain objects, named U-AHC*. The cluster merging criterion is based on an information-theoretic measure to compute the distance between cluster prototypes. These prototypes are represented as mixture densities that summarize the pdfs of all the uncertain objects in the clusters. Experiments have shown that our method outperforms state-of-the-art clustering algorithms from an accuracy viewpoint while achieving reasonably good efficiency.*

1. Introduction

Handling uncertainty in data management has been requiring more and more importance in a wide range of application contexts. Indeed, data uncertainty naturally arises from, e.g., implicit randomness in a process of data generation/acquisition, imprecision in physical measurements, and data staling. In general, uncertainty can be considered at table, tuple or attribute level, and is usually specified by fuzzy models, evidence-oriented models, or probabilistic models [15].

In this paper, we focus on data containing attribute-level uncertainty, which can be recognized in several application domains such as, e.g., biomedical measurement, financial and market data analysis, sensor networking, motion tracking, meteorological forecasting. We hereinafter refer to attribute-level uncertain data as *uncertain objects*. An uncertain object is usually represented by means of *probability density functions* (pdfs), which describe the probability that the object appears at any position in a multidimensional space [8, 4], rather than by a traditional vectorial form of deterministic values.

Dealing with uncertain objects has raised several issues in data management and knowledge discovery. In particular, organizing uncertain objects is challenging due to the intrinsic difficulty underlying the various notions of uncertainty. As a consequence to this challenge, *clustering uncertain objects* has been attracting increasing interest in recent years (e.g., [9, 8, 4, 13, 14]). While most existing algorithms for clustering uncertain data differ on the clustering strategy and the cluster model, the adopted notions of distance between uncertain objects come into two main approaches: computing the distance between aggregated values (e.g., expected values) extracted from the pdfs of the uncertain objects, or directly comparing the whole pdfs. However, both approaches have some drawbacks in their own: the first approach, as stated in, e.g., [7], has an accuracy issue, whereas the second one suffers from slow integration estimations and/or operations quadratic w.r.t. the size of the sample lists commonly used to approximate the pdfs of the uncertain objects. Moreover, traditional measures for comparing pdfs, such as the Ali-Silvey class of information-theoretic distance measures [1], cannot be used to directly define distances for uncertain objects. Indeed, these information-theoretic measures require that the pdfs come from random variables defined over a common event space, i.e., common domain region; unfortunately, the domain regions of the pdfs associated to the uncertain objects usually do not have wide intersections.

We propose a centroid-linkage-based agglomerative hierarchical algorithm for clustering uncertain objects, named *U-AHC*. To the best of our knowledge, the proposed algorithm represents the first agglomerative hierarchical approach to the problem of clustering uncertain objects. In

U-AHC, the cluster merging step is accomplished by a centroid-linkage criterion [12] which has the following main features: *i)* the cluster prototypes (i.e., cluster centroids) are computed as mixture densities that summarize the pdfs of all the objects in the clusters, and *ii)* the pair of closest clusters is chosen according to an information-theoretic measure that computes the distance between the cluster prototypes.

The centroid-linkage-based criterion does not require a notion of distance between the objects to be clustered, unlike other traditional linkage criteria in agglomerative hierarchical clustering. This allows us to avoid defining a notion of distance between uncertain objects, which is crucial in uncertainty similarity detection; instead, the adoption of cluster prototypes as mixture densities enables our algorithm to be equipped with a notion of information-theoretic distance measure that exploits an advantageous characteristic of the cluster prototypes: the overlaps between the cluster prototypes' domain regions are generally larger than the overlaps between the individual objects' regions.

We have conducted experiments in order to assess accuracy and efficiency of our algorithm, and to compare it to state-of-the-art methods for clustering uncertain data. Experimental results have shown that our U-AHC outperforms existing algorithms from an accuracy viewpoint while achieving efficiency comparable to density-based algorithms.

2. Related work

One of the earliest attempts to solve the problem of clustering uncertain data is the partitional algorithm UK-means [4], which is an adaptation of the popular K-means designed for handling uncertain objects. UK-means suffers from an expensive computation of the expected distances (EDs) between uncertain objects and cluster centroids, which is repeated at each iteration of the algorithm. In order to improve the UK-means efficiency, pruning techniques have been developed to avoid the computation of redundant EDs [13]. Such techniques make use of lower- and upper-bounds ad-hoc defined for each ED to be calculated. In [14], the CK-means is proposed as a variant of UK-means that exploits the moment of inertia of rigid bodies in order to reduce the execution time needed for computing EDs.

Density-based approaches have been also proposed for clustering uncertain objects. In [8], the fuzzy version of the popular DBSCAN, $\mathcal{F}$DBSCAN, uses fuzzy distance functions to compute the core object and reachability probabilities, which are at the basis of the density-based clustering strategy of the algorithm. A similar approach is presented in [9], which describes the $\mathcal{F}$OPTICS algorithm. Like the well-known hierarchical density-based clustering algorithm OPTICS, $\mathcal{F}$OPTICS produces an augmented ordering of the objects based on the notion of fuzzy object reachability-distance; this ordering can be eventually used to derive a cluster hierarchy.

In contrast to the majority of algorithms for clustering uncertain objects which are based on partitional or density-based schemes, it should be noted that there is relatively poor research on hierarchical clustering of uncertain data. Moreover, our U-AHC produces a cluster hierarchy, unlike $\mathcal{F}$OPTICS which outputs a reachability plot. Another important remark is that U-AHC does not require any input parameter, such as, e.g., a threshold for the neighbor distance or the minimum number of points in the object neighborhoods.

3. Clustering uncertain objects

Representing uncertain objects is traditionally accomplished by using two types of models, namely *multivariate uncertainty* and *univariate uncertainty* models.

Definition 1 (multivariate uncertain object) *A* multivariate uncertain object *o is a pair* (R, f)*, where* $R = [l_1, u_1] \times \cdots \times [l_m, u_m]$ *is the* m*-dimensional region in which o is defined and* $f : \Re^m \rightarrow \Re_0^+$ *is the probability density function of o at each point* $\vec{x} \in R$*, such that:*

$$\int_{\vec{x} \in R} f(\vec{x}) \mathrm{d}\vec{x} = 1 \qquad \textit{and} \qquad \int_{\vec{x} \in \Re^m \setminus R} f(\vec{x}) \mathrm{d}\vec{x} = 0$$

Definition 2 (univariate uncertain object) *A* univariate uncertain object *o is a tuple* $(a^{(1)}, \ldots, a^{(m)})$*. Each attribute* $a^{(h)}$ *is a pair* $(I^{(h)}, f^{(h)})$*, for each* $h \in [1..m]$*, where* $I^{(h)} = [l^{(h)}, u^{(h)}]$ *is the interval of definition of* $a^{(h)}$*, and* $f^{(h)} : \Re \rightarrow \Re_0^+$ *is the probability density function that assigns a probability value to each* $x \in I^{(h)}$*, such that:*

$$\int_{x \in I^{(h)}} f^{(h)}(x) \mathrm{d}x = 1 \qquad \textit{and} \qquad \int_{x \in \Re \setminus I^{(h)}} f^{(h)}(x) \mathrm{d}x = 0$$

3.1. Uncertain prototype

We introduce the notion of *uncertain prototype* as a new uncertain object computed from a set of uncertain objects, which summarizes the features of all the objects in the set. Basically, an uncertain prototype is represented by mixture densities from the pdfs associated to each object in the set to be summarized.

Definition 3 (multivariate uncertain prototype) *Let* $\mathcal{C} = \{o_1, ..., o_n\}$ *be a set of multivariate uncertain objects, where* $o_i = (R_i, f_i)$*,* $R_i = [l_{i_1}, u_{i_1}] \times \ldots \times [l_{i_m}, u_{i_m}]$*, for each* $i \in [1..n]$*. The* multivariate uncertain prototype *of* $\mathcal{C}$ *is a multivariate uncertain object* $\mathcal{P}_\mathcal{C} = (R_\mathcal{C}, f_\mathcal{C})$*, where*

$$R_\mathcal{C} = \Big[\min_{i \in [1..n]} l_{i_1}, \max_{i \in [1..n]} u_{i_1} \Big] \times \cdots \times \Big[\min_{i \in [1..n]} l_{i_m}, \max_{i \in [1..n]} u_{i_m} \Big],$$

$$f_{\mathcal{C}}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\vec{x})$$

Definition 4 (univariate uncertain prototype) *Let* $\mathcal{C} = \{o_1, ..., o_n\}$ *be a set of univariate uncertain objects, where* $o_i = ((I_i^{(1)}, f_i^{(1)}), \ldots, (I_i^{(m)}, f_i^{(m)}))$, $I_i^{(h)} = [l_i^{(h)}, u_i^{(h)}]$, *for each* $h \in [1..m]$, $i \in [1..n]$. *The* univariate uncertain prototype *of* $\mathcal{C}$ *is a univariate uncertain object* $\mathcal{P}_{\mathcal{C}} = ((I_{\mathcal{C}}^{(1)}, f_{\mathcal{C}}^{(1)}), \ldots, (I_{\mathcal{C}}^{(m)}, f_{\mathcal{C}}^{(m)}))$ *such that, for each* $h \in [1..m]$*:*

$$I_{\mathcal{C}}^{(h)} = \Big[\min_{i\in[1..n]} l_i^{(h)}, \max_{i\in[1..n]} u_i^{(h)}\Big], \quad f_{\mathcal{C}}^{(h)}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i^{(h)}(x)$$

3.2. Distance between uncertain prototypes

To define a distance measure between uncertain prototypes, we employ a function that exploits the full information stored in the pdfs. Two of the most frequently used distance measures between probability densities are the Kullback-Leibler divergence [11, 10] and the Chernoff distance [5], which fall into the Ali-Silvey class of *information-theoretic* distance measures [1]. However, such distances may be disadvantageous in our setting for a number of reasons—for instance, the Kullback-Leibler divergence is not symmetric, the Chernoff distance is typically hard to compute, and both measures do not satisfy the triangle inequality.

Our definition of distance between prototypes exploits a measure based on the *Bhattacharyya coefficient* [3, 6], which is defined as follows:

$$\rho(p(\vec{x}), q(\vec{x})) = \int_{\vec{x}\in\Re^m} \sqrt{p(\vec{x})\, q(\vec{x})}\, \mathrm{d}\vec{x} \qquad (1)$$

The original definition of ρ in [3] has an important geometric interpretation: it can be seen as the cosine between the two vectors for p and q, whose components are the square root of the probabilities of the k classes that compose p and q. This interpretation also holds in an extension of the definition in Equation (1) to use the Bhattacharyya coefficient for continuous pdfs.

Among the various distance measures that can be defined based on the Bhattacharyya coefficient [6], in this work we use the following

$$\mathrm{B}(p(\vec{x}), q(\vec{x})) = \sqrt{1 - \rho(p(\vec{x}), q(\vec{x}))} \qquad (2)$$

which has a number of advantages w.r.t. other Bhattacharyya distances, such as the commonly used $-\log\rho$ definition. In particular, the Bhattacharyya distance in Equation (2) obeys the triangle inequality, ranges within the interval $[0, 1]$ and, unlike the Chernoff distance (which is a more general case), is easier to compute and satisfies the additive property even if the random variables are not identically distributed.

In the following we give our definitions of distance between uncertain prototypes both for the multivariate and univariate case.

Definition 5 (multivariate uncertain prototype distance) *Given a set* $\mathcal{D}$ *of multivariate uncertain objects, let* $\mathcal{P}_{\mathcal{C}_i} = (R_{\mathcal{C}_i}, f_{\mathcal{C}_i})$ *and* $\mathcal{P}_{\mathcal{C}_j} = (R_{\mathcal{C}_j}, f_{\mathcal{C}_j})$ *be the multivariate uncertain prototypes of the sets* $\mathcal{C}_i, \mathcal{C}_j \subseteq \mathcal{D}$*, respectively. The* multivariate uncertain prototype distance *between* $\mathcal{P}_{\mathcal{C}_i}$ *and* $\mathcal{P}_{\mathcal{C}_j}$ *is defined as*

$$\Delta(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = \gamma\, \Delta'(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) + (1-\gamma)\, \Delta''(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) \quad (3)$$

where

$$\Delta'(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = \mathrm{B}(f_{\mathcal{C}_i}, f_{\mathcal{C}_j}),$$

$$\Delta''(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = \frac{1}{E_{max}(\mathcal{D})} d(E[f_{\mathcal{C}_i}], E[f_{\mathcal{C}_j}]),$$

$$\gamma = \frac{\mathcal{V}(R_{\mathcal{C}_i} \cap R_{\mathcal{C}_j})}{\min\{\mathcal{V}(R_{\mathcal{C}_i}), \mathcal{V}(R_{\mathcal{C}_j})\}}$$

In Definition 5, d is a function that measures the distance between m-dimensional points (e.g., Euclidean norm), $E[f]$ denotes the expected value of the pdf f, $\mathcal{V}(R)$ is the hypervolume of the m-dimensional region R, and E_{max} is a normalization term, which is defined as:

$$E_{max}(\mathcal{D}) = \max_{o_u, o_v \in \mathcal{D}} d(E[f_u], E[f_v])$$

It should be noted that Δ ranges within $[0, 1]$, since Δ' and Δ'' range within $[0, 1]$ in turn.

Let us now explain the reasons for introducing Δ' and Δ'' in Equation (3). The Bhattacharyya distance (Equation (2)) compares two pdfs by considering their portions defined over a common event space (i.e., common domain region). Thus, if the event spaces of the two pdfs do not have any intersection, the Bhattacharyya distance does not work, i.e., it is always equal to one. Although these cases are quite infrequent because of the way uncertain prototypes are defined, we introduce the term Δ'' in Equation (3) to discriminate among those cases by considering the distance between the expected values of the prototype pdfs. We weight the terms Δ' and Δ'' by involving the coefficient $\gamma \in [0, 1]$, which quantifies the importance of Δ' and Δ'' in the definition of Δ. In particular, γ is proportional to the width of the domain region shared between the prototypes to be compared. This definition of γ represents a reasonable choice, since the larger the portion of the pdfs involved into the Bhattacharyya distance calculation, the smaller the need for comparing the pdfs by also considering the corresponding expected values, and vice versa.

Definition 6 (univariate uncertain prototype distance) *Given a set* $\mathcal{D}$ *of univariate uncertain objects, let* $\mathcal{P}_{\mathcal{C}_i} = ((I_{\mathcal{C}_i}^{(1)}, f_{\mathcal{C}_i}^{(1)}), \ldots, (I_{\mathcal{C}_i}^{(m)}, f_{\mathcal{C}_i}^{(m)}))$ *and*

$\mathcal{P}_{\mathcal{C}_j} = ((I^{(1)}_{\mathcal{C}_j}, f^{(1)}_{\mathcal{C}_j}), \ldots, (I^{(m)}_{\mathcal{C}_j}, f^{(m)}_{\mathcal{C}_j}))$ *be the univariate uncertain prototypes of the sets* $\mathcal{C}_i, \mathcal{C}_j \subseteq \mathcal{D}$*, respectively. The* univariate uncertain prototype distance *between* $\mathcal{P}_{\mathcal{C}_i}$ *and* $\mathcal{P}_{\mathcal{C}_j}$ *is defined as*

$$\Delta(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = f_{dist}(\delta^{(1)}, \ldots, \delta^{(m)}) \quad (4)$$

where

$$\delta^{(h)} = \gamma^{(h)}\ \mathrm{B}(f^{(h)}_{\mathcal{C}_i}, f^{(h)}_{\mathcal{C}_j}) + \\ +(1-\gamma^{(h)})\left(\frac{1}{E^{(h)}_{max}(\mathcal{D})}\left|E\left[f^{(h)}_{\mathcal{C}_i}\right] - E\left[f^{(h)}_{\mathcal{C}_j}\right]\right|\right)$$

and

$$\gamma^{(h)} = \frac{\mathcal{V}(I^{(h)}_{\mathcal{C}_i} \cap I^{(h)}_{\mathcal{C}_j})}{\min\{\mathcal{V}(I^{(h)}_{\mathcal{C}_i}), \mathcal{V}(I^{(h)}_{\mathcal{C}_j})\}},$$

$$E^{(h)}_{max}(\mathcal{D}) = \max_{o_u, o_v \in \mathcal{D}} |E[f^{(h)}_u] - E[f^{(h)}_v]|$$

for each $h \in [1..m]$.

In Equation (4), $f_{dist} : \Re^m \rightarrow \Re$ is a function that computes a scalar value from the components of an m-dimensional vector. In this work, we define $f_{dist}(\delta^{(1)}, \ldots, \delta^{(m)}) = \sqrt{(1/m)\sum_{h=1}^{m}(\delta^{(h)})^2}$.

3.3. The U-AHC algorithm

Algorithm 1 U-AHC

Input: a set of uncertain objects $\mathcal{D} = \{o_1, \ldots, o_n\}$
Output: a set of partitions $\mathbf{D}$
1: $\mathbf{C} \leftarrow \{\{o_1\}, \ldots, \{o_n\}\}$
2: $\mathbf{D} \leftarrow \{\mathbf{C}\}$
3: **repeat**
4: let $\mathcal{C}_i, \mathcal{C}_j$ be the pair of clusters in $\mathbf{C}$ such that $\frac{1}{2}(\Delta(\mathcal{P}_{\mathcal{C}_i \cup \mathcal{C}_j}, \mathcal{P}_{\mathcal{C}_i}) + \Delta(\mathcal{P}_{\mathcal{C}_i \cup \mathcal{C}_j}, \mathcal{P}_{\mathcal{C}_j}))$ is minimum
5: $\mathbf{C} \leftarrow \{\mathcal{C} \in \mathbf{C} : \mathcal{C} \neq \mathcal{C}_i, \mathcal{C} \neq \mathcal{C}_j\} \cup \{\mathcal{C}_i \cup \mathcal{C}_j\}$
6: $\mathbf{D} \leftarrow \mathbf{D} \cup \{\mathbf{C}\}$
7: **until** $|\mathbf{C}| > 1$
8: **return** $\mathbf{D}$

Algorithm 1 outlines our AHC-based algorithm for clustering uncertain objects, named *U-AHC*. Given a dataset $\mathcal{D}$ of n uncertain objects, the algorithm follows the classic AHC scheme to produce a hierarchy of clusters $\mathbf{D}$. The merge score used to decide for the pair of clusters to be merged at each step of the U-AHC algorithm (Line 4) employs the notions of distance between uncertain prototypes (Definitions 5-6). In particular, for any pair of clusters $\mathcal{C}_i, \mathcal{C}_j$ belonging to the current clustering $\mathbf{C}$, we compute the prototype of the cluster given by the union of the objects in $\mathcal{C}_i$ and $\mathcal{C}_j$, and evaluate the uncertain distances between this prototype and the prototypes of $\mathcal{C}_i$ and $\mathcal{C}_j$. We use the mean of these distances as a merge score, since intuitively the smaller these distances, the smaller the error of merging $\mathcal{C}_i$ and $\mathcal{C}_j$ to form a new cluster.

Table 1. Datasets used in the experiments

dataset	*objects*	*attributes*	*classes*
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6
Ecoli	327	7	5

4. Experimental evaluation

We evaluated effectiveness and efficiency of the U-AHC algorithm in clustering uncertain data. The experimental evaluation was also conducted to give a comparison of U-AHC with existing K-means based algorithms (i.e., UK-means and CK-means) and density-based algorithms (i.e., $\mathcal{F}$DBSCAN and $\mathcal{F}$OPTICS).

4.1. Evaluation methodology

Datasets. We used benchmark datasets available from the UCI Machine Learning Repository.[1] For the main experiments, we selected four datasets with real-value attributes, namely Iris, Wine, Glass, and Ecoli.

Table 1 shows the main characteristics of the datasets. Iris contains measurements on different iris plants. Wine describes results of a chemical analysis of Italian wines derived from three different cultivars. In Glass, each glass instance is described by the values of its chemical components. Ecoli contains data on the Escherichia Coli bacterium, which are identified with values coming from different analysis techniques.

All the selected datasets were originally created to contain deterministic values. We synthetically generated uncertainty in the data, obtaining both univariate and multivariate uncertain objects as follows.

For each univariate object o, we produced the uncertain interval $I^{(h)}$ and the pdf $f^{(h)}$ defined over $I^{(h)}$, for each attribute $a^{(h)}$, with $h \in [1..m]$. The interval $I^{(h)}$ was randomly chosen as a subinterval within $[min_{o_h}, max_{o_h}]$, where min_{o_h} (resp. max_{o_h}) is the minimum (resp. maximum) deterministic value of the h-th attribute, over all the objects belonging to the same ideal class of o. As concerns $f^{(h)}$, we considered *Uniform*, *Normal* and *Gamma* pdfs. We set the parameters of Normal and Gamma pdfs in such a way that their mode corresponded to the deterministic value of the h-th attribute of object o.

For multivariate objects o, the uncertainty region R was defined as the product of the intervals randomly generated for each attribute of o. For the sake of brevity, we shall present results obtained using univariate models, nevertheless the relative performances of the algorithms were confirmed in the multivariate settings.

Clustering validity criteria. To assess the quality of clustering solutions we exploited the availability of reference

[1] http://archive.ics.uci.edu/ml/

Table 2. Accuracy results (F-measure) for univariate models

dataset	*pdf*	UK-means	CK-means	$\mathcal{F}$DBSCAN	$\mathcal{F}$OPTICS	**U-AHC**
Iris	Uniform	*0.93*	0.92	0.92	0.92	**0.93**
	Normal	0.84	0.85	*0.90*	*0.90*	**0.92**
	Gamma	0.60	0.50	*0.79*	0.77	**0.87**
Wine	Uniform	0.75	*0.76*	0.65	0.68	**1**
	Normal	0.70	0.71	*0.77*	0.76	**0.89**
	Gamma	*0.67*	0.58	0.64	0.64	**0.73**
Glass	Uniform	0.55	*0.69*	0.43	0.47	**0.81**
	Normal	0.58	0.55	0.60	*0.61*	**0.83**
	Gamma	0.46	0.51	0.62	*0.64*	**0.92**
Ecoli	Uniform	0.39	0.40	0.48	*0.51*	**0.79**
	Normal	0.73	*0.74*	0.68	0.68	**0.83**
	Gamma	*0.48*	0.41	0.47	0.47	**0.83**
	avg. score	0.64	0.635	0.663	0.67	0.863
	avg. gain	22.25%	22.75%	20%	19.17%	–

classifications for the datasets. The objective was to evaluate how well a clustering fits a predefined scheme of known classes (natural clusters). To this purpose, we resorted to the well-known *F-measure* [16] (ranging within $[0, 1]$), which is defined as the harmonic mean of the total precision and recall values, which in turn are computed by averaging over the classes the values of precision and recall obtained for each pair cluster-class.

Settings of the clustering methods. In UK-means and CK-means, the resulting accuracy and efficiency measurements were averaged over 100 different runs, in order to avoid that clustering results were biased by random chance due to the initial centroid selection.

We performed a tuning phase for the parameters ϵ (i.e., the threshold for the distance of the neighbors of an object) and μ (i.e., the minimum number of points within the neighborhood of an object) required by $\mathcal{F}$DBSCAN and $\mathcal{F}$OPTICS. We set these parameters to the values that allowed each method to achieve the best accuracy results.

For our U-AHC and $\mathcal{F}$OPTICS, we also needed for selecting a partition of appropriate size from the clustering solution obtained by each of these methods. In case of U-AHC, which directly produces a cluster hierarchy, we cut the solution at the level corresponding to the number of output clusters equal to the number of ideal classes for each dataset. For $\mathcal{F}$OPTICS, we further derived a cluster hierarchy by exploiting the method defined in [2].

We computed the integrals involved into the distance calculation by taking into account lists of samples derived from the pdfs. To this purpose, we employed the classic *Monte Carlo* sampling method.[2] We also performed a preliminary tuning phase to properly set the number of samples s for accuracy evaluation; in particular, for each method and dataset, we chose s in such a way that there was no significant improvement in accuracy for any $s' > s$.

[2]We used the SSJ library at http://www.iro.umontreal.ca/∼simardr/ssj/

4.2. Results

Accuracy. Table 2 summarizes the F-measure results obtained by U-AHC and the other methods on the various datasets, for the univariate models. In the table, the last two rows contain, respectively, the average F-measure score obtained by each method and the average percentage gain (in terms of quality) of U-AHC in relation to each method.

Overall, U-AHC outperformed the other methods with average gains from 19% ($\mathcal{F}$OPTICS) to about 23% (CK-means). Looking at the performances on each dataset, U-AHC achieved the following maximum quality improvements (i.e., gains w.r.t. the relative worst methods): from 1% (Iris) to 40% (Ecoli), for Binomial pdfs; from 8% (Iris) to 28% (Glass), for Normal pdfs; from 15% (Wine) to 46% (Glass), for Gamma pdfs. Also, U-AHC achieved the following minimum quality improvements (i.e., gains w.r.t. the relative best methods): from 0% (Iris) to 28% (Ecoli), for Binomial pdfs; from 2% (Iris) to 22% (Glass), for Normal pdfs; from 6% (Wine) to 35% (Ecoli), for Gamma pdfs.

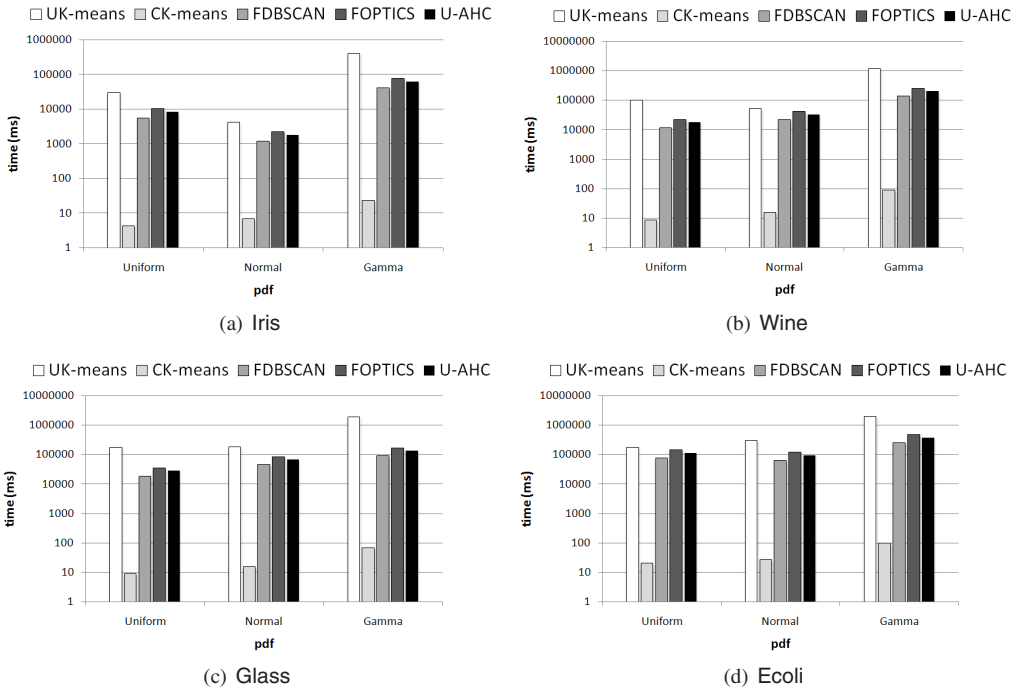
As far as the competing methods, we observed the better performance of the two density-based algorithms w.r.t. the K-means based algorithms, which is about 3%. It should be also noted that $\mathcal{F}$OPTICS and $\mathcal{F}$DBSCAN behaved closely each other, as well as UK-means and CK-means; actually, this was not surprising since the two couples of algorithms relatively employ similar clustering schemes.

Efficiency. We measured the runtime behaviors of U-AHC and the competing methods.[3] Figure 1 shows the total execution times (in milliseconds) obtained by the various methods on all the datasets. In the figure, we can see that our U-AHC performed one order of magnitude faster than UK-means on average. UK-means was the slowest method on all the datasets, which can be explained by the fact that each iteration of UK-means requires the EDs computation for all the objects in the dataset.

In general, the performances of CK-means, $\mathcal{F}$DBSCAN, $\mathcal{F}$OPTICS and our U-AHC followed the corresponding computational complexities, i.e., $\mathcal{O}(s\ n)$ for CK-means, $\mathcal{O}(n^2)$ for $\mathcal{F}$DBSCAN, and $\mathcal{O}(s\ n^2)$ for $\mathcal{F}$OPTICS and our U-AHC. Indeed, CK-means outperformed all the other methods on all the datasets, with a significant gain w.r.t. UK-means mainly due to the optimization employed for the EDs calculation. However, the CK-means algorithm is less general than the other methods, as it works only if the mean squared error for the definition of the EDs is used and the distance function is based on the Euclidean norm.

The performances of $\mathcal{F}$DBSCAN, $\mathcal{F}$OPTICS and our U-AHC showed to be comparable, with $\mathcal{F}$DBSCAN slightly better than U-AHC, and U-AHC better than $\mathcal{F}$OPTICS in

[3]Experiments were conducted on a platform Intel Pentium IV 3GHz with 2GB memory and running Microsoft WinXP Pro



(a) Iris (b) Wine

(c) Glass (d) Ecoli

Figure 1. Clustering time performances

turn. We also observed that the relative performance between U-AHC and $\mathcal{F}$OPTICS, and between $\mathcal{F}$DBSCAN and U-AHC did not substantially vary with the dataset.

5. Conclusion

We have studied the problem of clustering uncertain data and proposed *U-AHC*, a centroid-linkage-based agglomerative hierarchical algorithm. According to univariate and multivariate uncertainty models, we have introduced a notion of uncertain (cluster) prototype which is based on mixture densities from the pdfs associated to the objects belonging to a cluster. The cluster merging criterion in U-AHC exploits a new information-theoretic-based distance between uncertain prototypes. Our U-AHC has shown to outperform other existing methods in terms of accuracy, regardless of the choice of uncertainty density function. Also, from an efficiency viewpoint, U-AHC performs comparably to density-based clustering algorithms.

References

[1] S. M. Ali and S. D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *J. Roy. Stat. Soc.*, 28(1):131–142, 1966.

[2] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander. OPTICS: Ordering Points To Identify the Clustering Structure. In *Proc. ACM SIGMOD Conf.*, pages 49–60, 1999.

[3] A. Bhattacharyya. On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. *Bull. Calcutta Math. Soc.*, 35:99–110, 1943.

[4] M. Chau, R. Cheng, B. Kao, and J. Ng. Uncertain Data Mining: An Example in Clustering Location Data. In *Proc. PAKDD Conf.*, pages 199–204, 2006.

[5] H. Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *Ann. Math. Stat.*, 23(4):493–507, 1952.

[6] T. Kailath. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Trans. on Comm. Tech.*, 15(1):52–60, 1967.

[7] H. P. Kriegel, P. Kunath, M. Pfeifle, and M. Renz. Approximated Clustering of Distributed High-Dimensional Data. In *Proc. PAKDD Conf.*, pages 432–441, 2005.

[8] H. P. Kriegel and M. Pfeifle. Density-Based Clustering of Uncertain Data. In *Proc. ACM SIGKDD Conf.*, pages 672–677, 2005.

[9] H. P. Kriegel and M. Pfeifle. Hierarchical Density-Based Clustering of Uncertain Data. In *Proc. IEEE ICDM Conf.*, pages 689–692, 2005.

[10] S. Kullback. *Information theory and statistics*. Wiley, 1959.

[11] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Ann. Math. Stat.*, 22(1):79–86, 1951.

[12] F. Murtagh. A survey of recent advances in hierarchical clustering algorithm. *The Computer Journal*, 26(4):354–359, 1983.

[13] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip. Efficient Clustering of Uncertain Data. In *Proc. IEEE ICDM Conf.*, pages 436–445, 2006.

[14] S. D. Lee and B. Kao and R. Cheng. Reducing UK-means to K-means. In *Proc. IEEE ICDM Workshops*, pages 483–488, 2007.

[15] Y. Tao, X. Xiao, and R. Cheng. Range Search on Multidimensional Uncertain Data. *ACM TODS*, 32(3):15–62, 2007.

[16] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.