Hierarchical Clustering of Uncertain Data

Francesco Gullo and Giovanni Ponti

Dept. of Electronics, Computer and Systems Sciences, University of Calabria, Italy,
email: {fgullo, gponti}@deis.unical.it

Abstract. *Uncertain data* objects are usually represented in terms of an uncertainty region over which a probability distribution is defined. Dealing with such data has raised several issues in data management and knowledge discovery, mainly due to the intrinsic difficulty underlying the various notions of uncertainty. Recently, there has been a growing interest in clustering uncertain data, with a special emphasis on partitional and density-based approaches, whereas there has been a poor research on hierarchical methods.
The goal of this work is to address the problem of hierarchical clustering of uncertain objects. We developed a centroid-linkage-based agglomerative hierarchical method for clustering uncertain objects, named *U-AHC*, [1] whose major novelty lies in a well-founded information-theoretic approach to the computation of distance between uncertain objects. Experiments conducted on various datasets have shown that our method outperforms state-of-the-art methods for clustering uncertain data from an accuracy viewpoint.

Key words: data mining, clustering, uncertain data, information theory, probability density function

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 [2], and is usually specified by fuzzy models [3], evidence-oriented models [4], or probabilistic models [5].

We focus on data containing attribute-level uncertainty that is modeled according to a probabilistic model. We hereinafter refer to this data as *uncertain objects*. An uncertain object is usually represented by means of *probability distributions*, which describe the likelihood that the object appears at each position in a multidimensional space [6], rather than by a traditional vectorial form of deterministic values. Attribute-level uncertainty expressed by means of probabilistic models is present in several application domains, such as sensor data [7], moving objects [8], or gene expression data [9].

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 [10, 11, 12, 13, 14].

While the proposed clustering algorithms mainly differ on the clustering strategy and the cluster model, the adopted notions of distance between uncertain objects come into two main approaches. The first approach consists in computing the distance between

[1] An early version of the U-AHC algorithm was originally presented in [1].

aggregated values extracted from the probability distributions of the uncertain objects (e.g., expected values); the second approach instead requires to somehow compare the whole distributions. However, both these approaches have some drawbacks in their own: as stated in, e.g., [15], the first approach has an accuracy issue, whereas the second one may suffer from expensive operations for approximating the probability distributions of the uncertain objects.

Information-theory represents a fruitful research area to devise distance measures for comparing probability distributions. However, most of the existing information-theoretic measures, such as the popular ones falling into the Ali-Silvey class of distance measures [16], cannot be used to directly define distances for uncertain objects. Indeed, such measures commonly require that the probability distributions being compared hold for random variables defined over a common event space (i.e., common domain region); unfortunately, the domain regions of the probability distributions associated to uncertain objects may not have wide intersections.

Within this view, we address the problem of clustering uncertain objects by proposing a different approach to the computation of distance between probability distributions for uncertain data, which is employed in a hierarchical clustering framework. Our main contributions can be summarized as follows:

1. We propose a compound distance measure which is established based on two different ways of comparing probability distributions that represent uncertain data objects: *(i)* measuring the distance by involving the whole probability distributions, and *(ii)* computing the difference between the expected values of the distributions. The intuition behind this definition of distance lies in the fact that comparing two distributions by an information-theoretic distance is powerful but, in principle, not always applicable to uncertain objects; on the contrary, expected value of distributions is always computable but represents a concise information that is not able to capture the real proximity (which also depends on the shapes) between probability distributions.
   We introduce a notion of *adequacy* of computing the distance between any two probability distributions (of uncertain objects) by means of a given information-theoretic distance. Intuitively, this notion expresses to what degree an information-theoretic distance measure is worth comparing two uncertain objects by involving only their pdfs. We also prove that the adequacy of comparing any two probability distributions provides an upper-bound for the computation of the information-theoretic measure adopted in our framework.
2. We present a *centroid-linkage-based agglomerative hierarchical* algorithm for clustering uncertain objects, named *U-AHC*. The prototype (centroid) of any given cluster is computed as a mixture model that summarizes the probability distributions of all the objects within that cluster. At each iteration of the algorithm, the pair of closest clusters is chosen according to an information-theoretic measure that computes the distance between the cluster prototypes. The whole information represented in the probability distributions is exploited for comparing uncertain objects, while computing certain information-theoretic measures is a reasonably efficient operation.

2 State of the art

One of the earliest attempts to solve the problem of clustering uncertain data is the partitional algorithm *UK-means* [12], which is an adaptation of the popular K-means [17] designed for handling uncertain objects. UK-means suffers from a major weakness, that is the expensive computation of the *expected distance* (ED) between uncertain objects and cluster centroids, at each iteration of the algorithm. In order to improve the UK-means efficiency, [13, 18] proposes some pruning techniques to avoid the computation of redundant EDs, whereas in [14], the *CK-means* algorithm 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.

Besides the efficiency issue due to EDs calculation, UK-means suffers also from an accuracy issue. Indeed, cluster centroids are computed as deterministic objects using the expected values of the pdfs of the clustered objects. In [19], the *UK-medoids* algorithm is proposed to overcome the above issue by employing distance functions properly defined for uncertain objects and exploiting a K-medoids scheme.

Devising a fuzzy distance function is a key aspect in density-based approaches to clustering uncertain objects [11, 10]. In [11], the *$\mathcal{F}$DBSCAN* is proposed as a fuzzy version of the popular DBSCAN, which uses fuzzy distance functions to compute the core object and reachability probabilities. A similar approach is presented in *$\mathcal{F}$OPTICS* [10]. Like the well-known density-based clustering algorithm OPTICS, $\mathcal{F}$OPTICS produces an augmented ordering of the objects based on the notion of fuzzy object reachability-distance, which can eventually be used to derive a cluster hierarchy.

3 Uncertain objects

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 $f : \Re^m \to \Re_0^+$ is the probability density function of o at each point $\mathbf{x} \in \Re^m$, such that $\int_{\mathbf{x} \in \Re^m \setminus R} f(\mathbf{x}) \mathrm{d}\mathbf{x} = 0$ and $f(\mathbf{x}) > 0, \ \forall \mathbf{x} \in R$.*

3.1 Distance measures for pdfs

Probability density functions are usually compared by using *information-theoretic* (IT) measures, such as those falling into the Ali-Silvey class of distance functions [16] (e.g., the *Kullback-Leibler* (KL) divergence [20] and the *Chernoff* distance [21]).

Using an IT proximity measure represents a natural solution for devising a notion of distance between uncertain objects; in particular, this choice is essential to establish a function that is able to compare two pdfs by exploiting the whole information stored in the pdfs. However, this holds provided that the comparison makes sense: indeed, it should be taken into account that IT measures work out for pdfs that share a common event space (domain region). By contrast, if the two pdfs do not have any intersection in their event spaces (i.e., there is no region in which both pdfs are greater than zero), the distance according to any IT measure is always equal to the maximum value possible.

We introduce a notion, called *IT-adequacy*, which quantifies to what degree an information-theoretic distance measure is worth comparing two uncertain objects by involving only their pdfs.

Definition 2 (IT-adequacy). *Let g_1 and g_2 be two m-dimensional pdfs ($m \geq 1$), and $R_1 \subseteq \Re^m$, $R_2 \subseteq \Re^m$ be two m-dimensional regions such that (for $i \in \{1, 2\}$): $\int_{\mathbf{x} \in \Re^m \setminus R_i} g_i(\mathbf{x}) \mathrm{d}\mathbf{x} = 0 \quad and \quad g_i(\mathbf{x}) > 0 \ \ , \forall \mathbf{x} \in R_i$ The* IT-adequacy *between g_1 and g_2 with respect to R_1 and R_2 is defined as:*

$$\Upsilon_{R_1, R_2}(g_1, g_2) = \frac{1}{2} \Bigg(\int\limits_{\mathbf{x} \in R_1 \cap R_2} g_1(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int\limits_{\mathbf{x} \in R_1 \cap R_2} g_2(\mathbf{x}) \, \mathrm{d}\mathbf{x} \Bigg) \tag{1}$$

$\Upsilon_{R_1,R_2}(g_1, g_2)$ (which ranges within $[0, 1]$) expresses the adequacy of computing the distance between g_1 and g_2 by using a certain IT measure. In particular, the higher $\Upsilon_{R_1,R_2}(g_1, g_2)$, the more the information coming from g_1 and g_2 that is exploited in the comparison. For the sake of simplicity of notation, we will use the symbols Υ to denote the IT-adequacy relative to the comparison of any two multivariate uncertain objects. Formally, $\Upsilon(o_i, o_j) = \Upsilon_{R_i,R_j}(f_i, f_j)$ for any two multivariate uncertain objects $o_i = (R_i, f_i)$, $o_j = (R_j, f_j)$.

3.2 Distance measure for uncertain objects

According to Definition 1, it can be straightforwardly noted that a poor IT-adequacy may be computed when the pdfs of uncertain objects being compared have small (or empty) overlapping areas. To address such cases, it may be advisable to express the proximity between pdfs by resorting to the difference of their expected values. Within this view, the main intuition underlying our notion of distance measure between uncertain objects is to suitably combine an IT measure (which in principle is not always applicable) with a concise (but always available) information given by the expected values. Formally, we propose a distance measure Δ for uncertain objects o_i and o_j which is expressed as a function of two different terms:

$$\Delta(o_i, o_j) = \mathrm{f}(\Delta_{IT}(o_i, o_j), \Delta_{ED}(o_i, o_j)) \tag{2}$$

where Δ_{IT} involves a comparison by means of a certain IT measure, and Δ_{ED} measures the distance proportionally to the difference of the expected values.

In Equation 2, the IT measure we chose for computing Δ_{IT} is the Bhattacharyya distance [22]:

$$\mathrm{B}(g_1, g_2) = \sqrt{1 - \rho(g_1, g_2)} \tag{3}$$

where

$$\rho(g_1, g_2) = \int_{\mathbf{x} \in \Re^m} \sqrt{g_1(\mathbf{x})\, g_2(\mathbf{x})}\, \mathrm{d}\mathbf{x} \tag{4}$$

represents the *Bhattacharyya coefficient* [23], which compares any two continuous pdfs g_1, g_2 from a similarity viewpoint. The basic motivations for which B has been preferred to other IT measures (such as, e.g., $-\log \rho$, Kullback-Leibler or Chernoff) are the following. B ranges within $[0, 1]$, which makes this measure easily comparable and combinable with other distance functions, which represents a major focus on this work. Furthermore, B can be proved to be strictly related to Υ (Definition 2); indeed, it is based on ρ (the Bhattacharyya coefficient), for which the following nice property holds.

Proposition 1. *Let g_1 and g_2 be two m-dimensional pdfs ($m \geq 1$), and $R_1 \subseteq \Re^m$, $R_2 \subseteq \Re^m$ be two m-dimensional regions such that (for $i \in \{1, 2\}$) $\int_{\mathbf{x} \in \Re^m \setminus R_i} g_i(\mathbf{x}) \mathrm{d}\mathbf{x} = 0$, and $g_i(\mathbf{x}) > 0$, $\forall \mathbf{x} \in R_i$. It holds that $\rho(g_1, g_2) \leq \Upsilon_{R_1,R_2}(g_1, g_2)$.*

Proposition 1 shows that the upper bound of the computation of ρ for any two given pdfs is equal to the IT-adequacy between those pdfs. This represents a key aspect in our framework since it supports the theoretical validity for combining Δ_{IT} and Δ_{ED}. Indeed, in order to define a way to properly combine Δ_{IT} and Δ_{ED}, we resort to a factor that is proportional to the degree of overlap (i.e., the IT-adequacy) of the pdfs of the objects to be compared.

Definition 3 (multivariate uncertain distance). *The* multivariate uncertain distance *between two multivariate uncertain objects $o_i = (R_i, f_i)$ and $o_j = (R_j, f_j)$ is defined as*

$$\Delta(o_i, o_j) = \mathrm{B}(f_i, f_j) - \sqrt{1 - \Upsilon(o_i, o_j)}\; e^{-dist(E[f_i], E[f_j])} \tag{5}$$

In Equation (5), $dist : \Re^m \rightarrow \Re_0^+$ is a function that measures the distance between m-dimensional points (e.g., Euclidean norm), and $E[f]$ denotes the expected value of the pdf f. Note that the exponential function is used to make the distance between expected values ranging within $[0, 1]$.

Proposition 2. *Given any two uncertain objects o_i and o_j, $\Delta(o_i, o_j) \in [0, 1]$.*

Definition 4 (multivariate uncertain prototype). *Let $\mathcal{C} = \{o_1, \ldots, o_n\}$ 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 pair $\mathcal{P}_\mathcal{C} = (R_\mathcal{C}, f_\mathcal{C})$, where*

$$R_\mathcal{C} = \left[\min_{i\in[1..n]} l_i^{(1)}, \max_{i\in[1..n]} u_i^{(1)}\right] \times \cdots \times \left[\min_{i\in[1..n]} l_i^{(m)}, \max_{i\in[1..n]} u_i^{(m)}\right], \; f_\mathcal{C}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$$

Proposition 3. *Let $\mathcal{C} = \{o_1, \ldots, o_n\}$ be a set of uncertain objects. The uncertain prototype $\mathcal{P}_\mathcal{C}$ is an uncertain object according to Definition 1.*

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 $\Delta(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j})$ is minimum
5:     $\mathbf{C} \leftarrow \mathbf{C} \setminus \{\mathcal{C}_i, \mathcal{C}_j\} \cup \{\mathcal{C}_i \cup \mathcal{C}_j\}$
6:     $\mathbf{D} \leftarrow \mathbf{D} \cup \{\mathbf{C}\}$
7: until $|\mathbf{C}| = 1$
```

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 proposed notion of distance between uncertain objects (Definition 3).

4 Experimental evaluation

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

Datasets. Table 1 shows the main characteristics of the datasets used in the experiments. We selected eight benchmark datasets with real-value attributes available from the UCI Machine Learning Repository, [2] namely Iris, Wine, Glass, Ecoli, Yeast, Segmentation, Abalone, and Letter.

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

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
Yeast	1,484	8	10
Segm.	2,310	19	7
Abalone	4,124	7	17
Letter	7,648	16	10

Table 2. Accuracy results (F-measure)

dataset	*pdf*	UK-means	CK-means	UK-medoids	$\mathcal{F}$DBSCAN	$\mathcal{F}$OPTICS	**U-AHC**
Iris	U	0.948	*0.962*	0.907	0.929	0.907	**1**
	N	0.859	0.897	0.888	*0.929*	0.907	0.962
Wine	U	0.735	0.747	0.761	*0.767*	0.713	**0.826**
	N	0.707	0.705	*0.749*	0.691	0.713	0.795
Glass	U	0.677	*0.703*	0.653	0.575	0.636	0.779
	N	0.540	0.551	0.579	*0.868*	0.828	**0.891**
Ecoli	U	0.787	*0.790*	0.728	0.443	0.477	0.743
	N	*0.745*	0.740	0.560	0.416	0.477	**0.795**
Yeast	U	0.533	0.538	*0.622*	0.599	0.528	**0.684**
	N	0.455	*0.457*	0.318	0.374	0.420	0.486
Segm.	U	0.780	*0.801*	0.765	0.482	0.419	**0.837**
	N	0.628	0.637	*0.649*	0.415	0.419	0.684
Abalone	U	0.288	0.290	*0.531*	0.499	0.439	0.492
	N	0.215	0.217	0.288	0.497	*0.558*	**0.572**
Letter	U	0.637	0.636	*0.763*	0.320	0.318	**0.798**
	N	0.442	0.435	*0.595*	0.353	0.318	0.613

All the selected datasets were originally created to contain deterministic values. We synthetically generated uncertainty in the data as follows. For each object o, we produced an uncertain interval $I^{(h)}$ for each attribute $a^{(h)}$, $h \in [1..m]$. 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. The uncertainty region R was finally defined as the product of the intervals randomly generated for each attribute of o. As concerns f, we considered *Uniform* and *Normal* pdfs. We set parameters of Normal pdfs so that their mode corresponded to the deterministic value of object o.

Clustering validity criteria. To assess the quality of clustering solutions we exploited the availability of reference 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* [24] (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.

Results. Table 2 summarizes the F-measure results obtained by U-AHC and the other methods on the various datasets. On average, U-AHC outperformed the other methods on all the datasets, with average gains that ranged from 10% (vs. UK-medoids) to 18% (vs. $\mathcal{F}$OPTICS). Among the competing methods, UK-medoids behaved better than the other ones—6 out of 16 times—obtaining an average quality gain up to 8%. Also, the partitional algorithms performed slightly better than the density-based algorithms in the univariate case.

5 Conclusion

We addressed the problem of clustering uncertain data by proposing *U-AHC*, a centroid-linkage-based agglomerative hierarchical algorithm. We 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 been experimentally shown to outperform major competing methods in terms of accuracy on all the datasets used in the evaluation.

References

1. Gullo, F., Ponti, G., Tagarelli, A., Greco, S.: A Hierarchical Algorithm for Clustering Uncertain Data via an Information-Theoretic Approach. In: Proc. ICDM Conf. (2008) 821–826
2. Tao, Y., Xiao, X., Cheng, R.: Range Search on Multidimensional Uncertain Data. TODS **32**(3) (2007) 15–62
3. Galindo, J., Urrutia, A., Piattini, M.: Fuzzy Databases: Modeling, Design, and Implementation. Idea Group Publishing (2006)
4. Lee, S.K.: An Extended Relational Database Model for Uncertain and Imprecise Information. In: Proc. VLDB Conf. (1992) 211–220
5. Sarma, A.D., Benjelloun, O., Halevy, A., Widom, J.: Working Models for Uncertain Data. In: Proc. ICDE Conf. (2006) 7–18
6. Cheng, R., Kalashnikov, D.V., Prabhakar, S.: Evaluating probabilistic queries over imprecise data. In: Proc. SIGMOD Conf. (2003) 551–562
7. Cantoni, V., Lombardi, L., Lombardi, P.: Challenges for Data Mining in Distributed Sensor Networks. In: Proc. ICPR Conf. (2006) 1000–1007
8. Li, Y., Han, J., Yang, J.: Clustering Moving Objects. In: Proc. KDD Conf. (2004) 617–622
9. Hein, A.M.K., Richardson, S., Causton, H.C., Ambler, G.K., Green, P.J.: BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data. Biostatistics **6** (2005) 349–373
10. Kriegel, H.P., Pfeifle, M.: Hierarchical Density-Based Clustering of Uncertain Data. In: Proc. ICDM Conf. (2005) 689–692
11. Kriegel, H.P., Pfeifle, M.: Density-Based Clustering of Uncertain Data. In: Proc. KDD Conf. (2005) 672–677
12. Chau, M., Cheng, R., Kao, B., Ng, J.: Uncertain Data Mining: An Example in Clustering Location Data. In: Proc. PAKDD Conf. (2006) 199–204
13. Ngai, W.K., Kao, B., Chui, C.K., Cheng, R., Chau, M., Yip, K.Y.: Efficient Clustering of Uncertain Data. In: Proc. ICDM Conf. (2006) 436–445
14. S. D. Lee and B. Kao and R. Cheng: Reducing UK-means to K-means. In: Proc. IEEE ICDM Workshops. (2007) 483–488
15. Kriegel, H.P., Kunath, P., Pfeifle, M., Renz, M.: Approximated Clustering of Distributed High-Dimensional Data. In: Proc. PAKDD Conf. (2005) 432–441
16. Ali, S.M., Silvey, S.D.: A General Class of Coefficients of Divergence of One Distribution from Another. J. Roy. Stat. Soc. **28**(1) (1966) 131–142
17. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proc. Berkeley Symp. on Mathematical, Statistics and Probability. (1967) 281–297
18. Kao, B., Lee, S.D., Cheung, D.W., Ho, W.S., Chan, K.F.: Clustering Uncertain Data using Voronoi Diagrams. In: Proc. ICDM Conf. (2008) 333–342
19. Gullo, F., Ponti, G., Tagarelli, A.: Clustering Uncertain Data via K-medoids. In: Proc. SUM Conf. (2008) 229–242
20. Kullback, S., Leibler, R.A.: On Information and Sufficiency. Ann. Math. Stat. **22**(1) (1951) 79–86
21. Chernoff, H.: A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. Ann. Math. Stat. **23**(4) (1952) 493–507
22. Kailath, T.: The Divergence and Bhattacharyya Distance Measures in Signal Selection. IEEE Trans. on Comm. Tech. **15**(1) (1967) 52–60
23. Bhattacharyya, A.: On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. Bull. Calcutta Math. Soc. **35** (1943) 99–110
24. van Rijsbergen, C.J.: Information Retrieval. Butterworths (1979)