Be *Certain* of How-to before Mining *Uncertain* Data

Francesco Gullo[1], Giovanni Ponti[2], and Andrea Tagarelli[3]

[1] Yahoo Labs, Barcelona, Spain
gullo@yahoo-inc.com
[2] ENEA, Portici Research Center (NA), Italy
giovanni.ponti@enea.it
[3] DIMES, University of Calabria, Italy
tagarelli@dimes.unical.it

Abstract. The purpose of this technical note is to introduce the problems of *similarity detection* and *summarization* in uncertain data. We provide the essential arguments that make the problems relevant to the data-mining and machine-learning community, stating major issues and summarizing our contributions in the field. Further challenges and directions of research are also issued.

1 Uncertainty: What We Have to Face

The term *uncertainty* describes an ubiquitous status of the information as being produced, transmitted, and acquired in real-world data sources. Exemplary scenarios are related to the use of location-based services for tracking moving objects and sensor networks, which normally produce data whose representation (attributes) is imprecise at a certain degree. Imprecision arises from the presence of noisy factors in the device or transmission medium, but also from a high variability in the measurements (e.g., locations of a moving object) that obviously prevents an exact representation at a given time. This is the case virtually for any field in scientific computing, and consequently for a plethora of application fields, including: pattern recognition (e.g., image processing), bioinformatics (e.g., gene expression microarray), computational fluid dynamics and geophysics (e.g., weather forecasting), financial planning (e.g., stock market analysis), GIS applications to distributed network analysis [1].

For data management purposes, uncertainty has been traditionally treated at the attribute level, as this is particularly appealing for inductive learning tasks [6]. In general, attribute-level uncertainty is handled based on a probabilistic representation approach that exploits probability distributions describing the likelihood that any given data tuple appears at each position in a multidimensional domain region; the term *uncertain objects* is commonly used to refer to such data tuples described in terms of probability distributions defined over multidimensional domain regions.

Uncertainty in data representation needs to be carefully handled in order to produce meaningful knowledge patterns. Consider for instance the scenario depicted in Fig. 1—uncertain objects are represented in terms only of their domain regions for the sake of simplicity (probability distribution assumed to be uniform for all the objects). The

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 489–493, 2014.

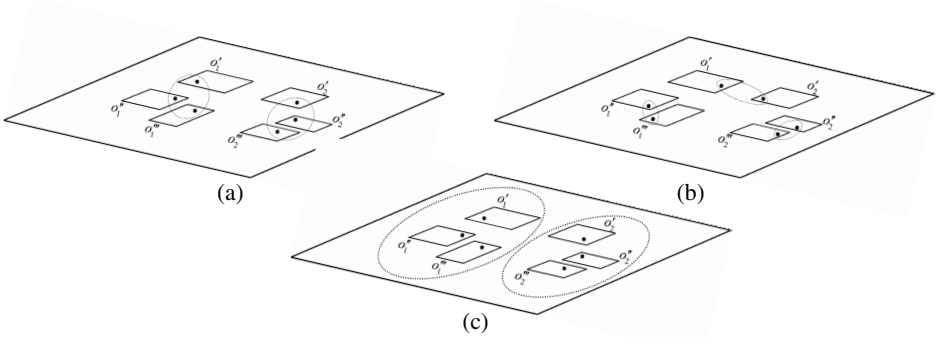


Fig. 1. Grouping uncertain data: (a) true representations of objects and their desired grouping, (b) observed representations which may lead to unexpected groupings, (c) desired grouping identified by considering the object uncertainty (domain regions).

"true" representation of each uncertain object (black circles in Fig. 1(a)) corresponds to a point within its domain region and can be in general far away from its "observed" representation (black circles in Fig. 1(b)). Thus, considering only the observed representations may lead to discover groups of similar objects (i.e., $\{o'_1, o'_2\}$, $\{o''_1, o'''_1\}$, $\{o''_2, o'''_2\}$ in Fig. 1(b)) that are substantially different from the ideal ones which would be identified by considering the true representations (i.e., $\{o'_1, o''_1, o'''_1\}$, $\{o'_2, o''_2, o'''_2\}$ in Fig. 1(a)). Instead, considering the whole domain regions (and pdfs) of the uncertain objects, may help to recognize the correct grouping (Fig. 1(c)).

The computation of proximity between uncertain objects is a fundamental primitive needed in many data-management tasks. Existing approaches fall into two main categories: (*i*) computing the distance between aggregated values extracted from the probability distributions of the uncertain objects (e.g., expected values), or (*ii*) computing the *expected distance* (ED) between distributions, which involves the whole information available from the distributions. The first approach is efficient as it has a time complexity linear in the number of statistical samples used for representing distributions, but it also has an evident accuracy issue since all the information available from the distributions is collapsed into a single numerical value; conversely, the ED-based approach is more accurate but also inefficient (it takes quadratic time). Within this view, our major contribution presented in [4] was to define a novel distance function that achieves a good tradeoff between accuracy and efficiency, by being able to capitalize on the whole information provided by the object distributions while keeping linear-time complexity.

Summarization of uncertain objects is another critical task, which is generally required in scenarios where a more compact representation is essential to analyze and/or further process a large set of (uncertain) objects that would be hard to manage otherwise. Surprisingly, a common trend in the early state-of-the-art was to employ a simple average of the expected values of the set members, which is clearly ineffective in most cases. Our contribution on this topic was to account also for the *variance* of the individual set members. In particular, we proposed a model based on a random variable derived from the realizations of the uncertain objects to be summarized [5], as well as a mixture-model-based summarization method [3,2].

Similarity detection in uncertain data is obviously central in a variety of mining tasks. Analogous consideration holds for uncertain data summarization, as it impacts

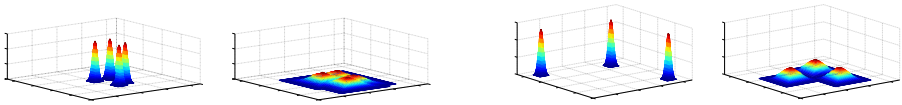
on how proximity can conveniently be computed between any uncertain object and a "prototype" object summarizing a set (e.g., cluster) of uncertain objects. Next we informally articulate the approaches we have proposed in the aforementioned contexts.

2 Uncertainty: How We Can Deal With

Similarity Detection in Uncertain Data. *Information-theory* (IT) has represented a fruitful research area to devise measures for comparing probability distributions accurately and, in most cases, in linear time with respect to the number of distribution samples. However, none of the prominent existing IT measures can be directly used to define distances for uncertain objects, mainly because of the assumption that the distributions need to share a common event space, which does not necessarily hold for distributions associated to uncertain objects. For this purpose, in [4] we developed a distance measure between uncertain objects that is able to exploit the full information stored in the object distributions, while being fast to compute. A major feature of our proposed distance is a combination of an IT measure with a measure based on aggregated information (i.e., expected value) extracted from the object distributions.

Besides representing a good tradeoff between effectiveness and efficiency, a further nice feature of the proposed distance is that the frequency of occurrence of the no-intersection event, and thus the overall accuracy of the measure, can be statistically controlled. Specifically, the width of the domain region shared between the uncertain objects to be compared represents a useful indicator of the feasibility of the distance calculation by means of the IT term only, and hence of the limited need for comparing the object distributions by also considering the expected-value-based term. This reasoning can profitably be exploited in tasks where the distances are to be computed for objects whose domain regions become larger as their processing goes on. An exemplary task where this happens is *prototype-based agglomerative hierarchical clustering*, where each cluster of uncertain objects is represented according to some notion of prototype whose domain region is ensured to increase with later steps of the clustering process.

Uncertain Data Summarization. As previously mentioned, the naïve notion of uncertain prototype as average of the expected values of the objects in a set has been widely used in the literature. Notwithstanding, it might easily result in limited accuracy, as (*i*) it has a deterministic representation, and (*ii*) it expresses only the central tendency of the objects to be summarized. This prompted us to investigate better ways of summarizing uncertain data. As a first attempt towards this matter, we proposed in [3,2] a notion of uncertain prototype as a mixture model of the set of random variables representing the uncertain objects to be summarized; that is, a notion that enables an uncertain representation while, at the same time, accounting for the variance of the individual objects rather than their central tendency only. A significant part of our work also focused on how to exploit our summarization approach in a classic data-mining task like clustering. Particularly, we demonstrated that a clustering objective criterion can be defined based on the minimization of the variance of the cluster mixture models, and that both efficiency and accuracy requirements can be satisfied. A major remark in this regard is that the proposed criterion enables the definition of fast heuristics that do not require any distance measure between uncertain objects.



(a) (b)

Fig. 2. Impact of the use of central tendency and variance of the individual objects on uncertain data summarization. (a) The two groups of objects have the same central tendency, but different variances: considering only central tendency leads to mistakenly summarize the two sets by the same prototype. (b) Considering only the variance is however not enough: even though the objects on the left have lower variance, they evidently form a less compact group than the one on the right.

Although the mixture-model-based approach has the merit of introducing a definition of uncertain prototype that is an uncertain object itself, a criterion based only on the minimization of the variance of the uncertain objects may still lead to unsatisfactory results. (Figure 2). Thus, in [5] we introduced a novel notion of uncertain prototype, named *U-centroid*: an uncertain object that is defined in terms of a random variable whose realizations correspond to all possible deterministic representations deriving from the uncertain objects to be summarized. Besides deriving the analytical expressions of domain region and probability distribution of the proposed U-centroid, a major contribution in this regard was the definition of a closed-form-computable compactness criterion that, coupled with the proposed U-centroid, naturally defines an effective yet efficient objective criterion for grouping uncertain objects.

3 Challenges and Future Directions

We provided a short review of problems related to similarity detection and summarization in uncertain data, and how we addressed them in our previous studies [3,5,2,4], which the interested reader is referred to for any technical details, including further developments we envisaged. Here we rather conclude raising a couple of concerns regarding the current trends of representing uncertainty and evaluating the induced mining results. First, we argue that a more complete treatment of uncertainty in data mining and machine learning could be obtained by integrating attribute-level with tuple-level notions of uncertainty, which have been long studied in database theory and management fields. This could imply the specification of a new yet more expressive class of models and algorithms for mining uncertain data. Second, we believe there is a strong need for the construction of benchmarks for assessing the mining results, which would avoid to bias the performance evaluation often due to the artificial, non-standardized methods for the generation of uncertainty in the selected test data. Another open problem that is worth to be addressed is the design of new assessment criteria to evaluate the many aspects inherent the quality of knowledge patterns induced from uncertain datasets.

References

1. Aggarwal, C.C.: Managing and Mining Uncertain Data. Springer (2009)
2. Gullo, F., Ponti, G., Tagarelli, A.: Minimizing the variance of cluster mixture models for clustering uncertain objects. In: IEEE ICDM, pp. 839–844 (2010)

3. Gullo, F., Ponti, G., Tagarelli, A.: Minimizing the variance of cluster mixture models for clustering uncertain objects. Statistical Analysis and Data Mining 6(2), 116–135 (2013)
4. Gullo, F., Ponti, G., Tagarelli, A., Greco, S.: A Hierarchical Algorithm for Clustering Uncertain Data via an Information-Theoretic Approach. In: IEEE ICDM, pp. 821–826 (2008)
5. Gullo, F., Tagarelli, A.: Uncertain Centroid based Partitional Clustering of Uncertain Data. PVLDB 5(7), 610–621 (2012)
6. Sarma, A.D., Benjelloun, O., Halevy, A.Y., Nabar, S.U., Widom, J.: Representing uncertain data: models, properties, and algorithms. The VLDB Journal 18(5), 989–1019 (2009)