A Framework for Clustering Uncertain Data

Erich Schubert, Alexander Koos, Tobias Emrich,
Andreas Züfle, Klaus Arthur Schmid, Arthur Zimek

Ludwig-Maximilians-Universität München
Oettingenstr. 67, 80538 Munich, Germany
http://www.dbs.ifi.lmu.de
{schube,koos,emrich,zuefle,schmid,zimek}@dbs.ifi.lmu.de

ABSTRACT

The challenges associated with handling uncertain data, in particular with querying and mining, are finding increasing attention in the research community. Here we focus on clustering uncertain data and describe a general framework for this purpose that also allows to visualize and understand the impact of uncertainty—using different uncertainty models—on the data mining results. Our framework constitutes release 0.7 of ELKI (`http://elki.dbs.ifi.lmu.de/`) and thus comes along with a plethora of implementations of algorithms, distance measures, indexing techniques, evaluation measures and visualization components.

1. INTRODUCTION

Given high-quality, reliable, up-to-date, exact, and sufficiently large data, clustering is often used to support advanced and educated decision making in many application domains in economics, health-care, science, and many more. Consequently, a large number of clustering algorithms has been developed to cope with different application scenarios. However, our ability to unearth valuable knowledge from large sets of data is often impaired by the quality of the data: data may be imprecise (e.g., due to measurement errors), data can be obsolete (e.g., when a dynamic database is not up-to-date), data may originate from unreliable sources (such as crowd-sourcing), the volume of the dataset may be too small to answer questions reliably [8], or it may be blurred to prevent privacy threats and to protect user anonymity [20]. Simply ignoring that data objects are imprecise, obsolete, unreliable, sparse, or cloaked, thus pretending the data were accurate, current, reliable, and sufficiently large, is a common source of false decision making. A different approach accepts these sources of error and creates models of what the true (yet admittedly unknown) data may look like. This is the notion of handling uncertain data [4]. The challenge in handling uncertain data is to obtain reliable results despite the presence of uncertainty. This challenge has received a strong research focus, by both industry and academia, in the last five years. "Veracity" has often been named as the fourth "V" of big data in addition to volume, velocity and variety. Adequate methods need to quantify the uncertainty in the data using proper models of uncertainty, and then to propagate the uncertainty through the data mining process, in order to obtain data mining results associated with significance and reliability information.

This demonstration targets the problem of how to derive a meaningful clustering from an uncertain dataset. For this purpose, we extend the ELKI framework [3] to handle uncertain data. ELKI is an open source (AGPLv3) data mining software written in Java aimed at users in research and algorithm development, with an emphasis on unsupervised methods such as cluster analysis and outlier detection. We give a short overview on our new release, ELKI 0.7, in Section 2.1. Additionally, we make the following contributions to handle uncertain data in a general way:

- ELKI 0.7 adds support for the most commonly used uncertainty models (Section 2.2). In particular, ELKI 0.7 provides an uncertain databases sampler, which derives multiple database samples from an uncertain database using the configured uncertainty model.
- The ELKI visualization tools have been extended to support (the clustering of) uncertain data. Therefore ground-truth data, observed data, as well as various sampled databases and their corresponding clusterings can be analyzed visually. This allows for getting an intuition of how uncertainty affects traditional clustering results. We describe this in more detail in Section 2.3.
- Comparison algorithms for clustering uncertain data for specific uncertainty models have been added to ELKI 0.7 (see Section 2.4). The ELKI framework can easily be extended by users to support their favorite algorithms.
- Traditional clustering algorithms as implemented in ELKI can be applied to sampled databases, and the clustering results can then be unified using the approach of Züfle et al. [22] as sketched in Section 2.5.

We outline the demonstration scenario in Section 3 and close with details on the public availability of our open source (AGPLv3) implementation in Section 4.

2. THE FRAMEWORK

This project is an extension of the ELKI framework [3] (`http://elki.dbs.ifi.lmu.de/`). Based on this framework, we aim at providing a platform to design, experiment with, and evaluate algorithms for uncertain data, as we will be instantly able to use the provided functionality.

Proceedings of the VLDB Endowment, Vol. 8, No. 12

2.1 General Functionality of ELKI

ELKI uses a modular and extensible architecture. Many algorithms in ELKI are implemented based on general distance functions and neighborhood queries, but are agnostic to the underlying data type or distance. Functionality provided by ELKI includes:[1]

- input readers for many popular file formats, such as CSV, ARFF, and the libSVM format;
- distance functions, including set-based distances, distribution-based distances, and string dissimilarities;
- clustering algorithms, including many k-means and hierarchical clustering variations, density-based algorithms such as DBSCAN and OPTICS, but also subspace clustering and correlation clustering algorithms;
- unsupervised outlier detection algorithms [2];
- data indexing methods such as R*-tree variations, M-tree variations, VA-file, and LSH that can be used to accelerate many algorithms;
- evaluation measures such as the adjusted Rand index (ARI) [17], Fowlkes-Mallows [16], BCubed [7], mutual-information-based, and entropy-based measures;
- a modular visualization architecture including scatterplots and parallel coordinates, using an SVG renderer to produce high quality vector graphics.

ELKI can be extended by implementing the appropriate interfaces. The provided UIs for ELKI will automatically detect the new implementations and allow simple configuration of experiments without the need to write further code. However, not all functionality required for analyzing uncertain data can be added using such extensions. In particular, sampling possible worlds will require an additional processing loop around the algorithms, and a second meta-clustering phase to aggregate these results. An application providing such more complex solutions is a core contribution of this demonstration and will be sketched below (Section 2.5).

2.2 Supported Uncertain Data Models

The most common discrete and continuous data models for uncertain data (cf. Figure 1) have been implemented in ELKI. Let us outline the implemented models briefly:

A pioneering uncertainty model is the *existential uncertainty model* [13, 10], where each data record is associated with a Bernoulli-event deciding whether the corresponding data object is present in the database. Our framework allows to specify an attribute column which will be used as an existential uncertainty of the respective tuple. For the case where each object is described by a discrete probability mass function (p.m.f.), the *block-independent disjoint tuples model* [14] and its common examples such as the *Uncertainty-Lineage Database model* [9] and the *X-Tuple model* or *U-Relation model* [5, 6] are implemented. Each object is represented by a finite number of alternative object-instances, each associated with the probability of being the true object, but still assuming stochastic independence between objects. In ELKI, a pmf of a data record o is represented by a list of (value, probability)-pairs, requiring that the sum of probabilities of all pairs of o does not exceed one. If the sum of probabilities is less than one, the difference to one is used as the probability that the tuple does not exist.

[1]For an extensive list of publications implemented in ELKI see: `http://elki.dbs.ifi.lmu.de/wiki/RelatedPublications`.

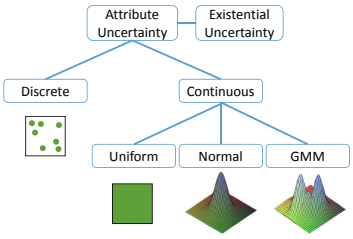


Figure 1: Uncertain Data Models

For the case of continuous probability density functions, ELKI provides classic parametric models for which the corresponding probability density function and cumulative distribution function can be specified. As standard parametric functions, ELKI offers support for uniform distributions and normal distributions. Then, for each object, the corresponding parameter values can be passed to ELKI, either by selecting attributes of a relation as parameter values, or by reading the parameter values from a file. In addition, mixture models are supported. For these models, a number of parametric probability density functions can be provided, each associated with a probability. This allows to support Gaussian mixture models as used by Böhm et al. [11].

For both cases (discrete or continuous distributions) ELKI provides data parsers and helper classes to link data records to their corresponding p.m.f. or p.d.f.

2.3 Visualization Tools

The ELKI visualization tools have been extended to support clustering of uncertain data. The corresponding view (cf. Figure 2) can switch between the following perspectives: (1) The result of clustering algorithm C on the ground-truth, if it is available, gives an intuition on how the clustering should look like without the presence of uncertainty. (2) The result of C on random samples gives insight on how the possible clusterings vary and how the uncertainty affects traditional clustering results. (3) The representative clusterings (c.f. Section 2.5)give a summarized view of the possible clusterings and allow for educated decision making. For all these perspectives we can utilize the existing visualization toolkit of ELKI including scatterplots (2-dimensional projections for each pair of attributes), 1-dimensional attribute distributions (histograms), parallel-coordinate plots, evaluation measures for data mining results, and different additional cluster model visualizations where applicable.

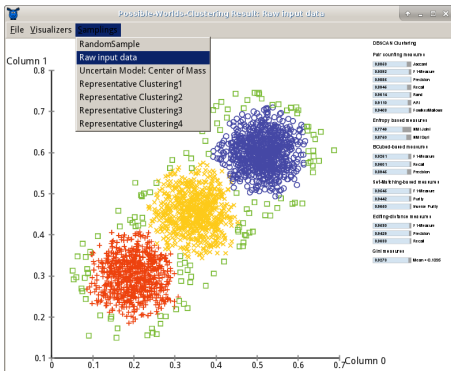


Figure 2: Views of a Dataset

2.4 Uncertain Clustering Algorithms

Based on the uncertainty models (Section 2.2), various clustering algorithms for uncertain data have been published in the past. We implemented some of the most prominent algorithms and made them available in ELKI.

A commonly used naive approach is implemented which represents each uncertain object by its expectation and uses a traditional clustering algorithm. Additionally, a fair baseline approach is to return the most probable clustering result. For this purpose, a naïve approach would be to apply clustering algorithm $\mathcal{C}$ to every possible world and return the clustering which has the highest support (the sum of probabilities of the worlds with this clustering result is the highest). However, the naïve is often not applicable since the number of possible worlds may be very large or even infinite, e.g., for continuous probability distributions. Thus we support a Monte Carlo approximation of the naïve approach, which applies $\mathcal{C}$ on a user-defined number of randomly sampled (i.i.d. according to the specified data model) databases, and returns the clustering with the highest support (most of the samples yield the clustering) in this sample.

In addition, a number of published algorithms for clustering of uncertain data have been implemented and are available in ELKI 0.7. One of the pioneering works in this field is the approach of Kriegel and Pfeifle [19]. This approach uses a discrete uncertainty model requiring each alternative of an object to have the same probability. Furthermore, the *UK-means* [12] has been implemented, which is similar to the *fuzzy c-means* algorithm [15]. This approach is applicable to arbitrary uncertainty models. Subsequent work [21] has shown how to improve the efficiency of this algorithm if uncertainty regions follow uniform distribution.

2.5 Representative Clustering

Clustering uncertain data is a daunting task, since the number of possible clusterings of a database with $|\mathcal{D}|$ objects is $O(2^{|\mathcal{D}|})$. Existing solutions to uncertain clustering either cluster aggregates (e.g., the expectation or the medoid) of the uncertain objects, or the probabilistic distance between objects [21], or the results are based on probabilistic distance thresholds [19]. These approaches have two crucial drawbacks: First they are tailored to a specific clustering algorithm, making the approach inapplicable in scenarios where the cluster algorithm does not fit the application. Second, and even more important, they have been shown to yield results that are not in accordance with possible worlds semantics.

In contrast, the approach of *representative clustering* [22] is more general and thus applicable in a wide number of scenarios, yielding outcomes that are in accordance with possible worlds semantics. The approach can be summarized in a few steps (cf. Figure 3): **First**, a set $X = \{X_1, .., X_N\}$ of possible worlds is sampled in an unbiased way from the database X, using ELKI's new uncertain database sampler. The uncertain objects in X can be represented by either of the models described in Section 2.2. **Second**, each database sample $X_i \in X$ is clustered using a traditional clustering algorithm $\mathcal{C}$, which can be chosen from the large list of clustering algorithms implemented in ELKI. Any algorithm that assigns clusters to discrete (potentially overlapping or non-exhaustive) partitions can be used, methods that perform a soft cluster assignment or yield a hierarchical result first need to be discretized into partitions. This step yields a set

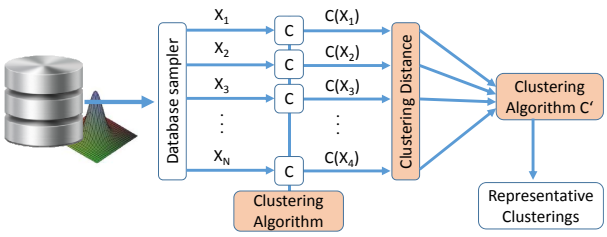


Figure 3: Workflow of Representative Clustering

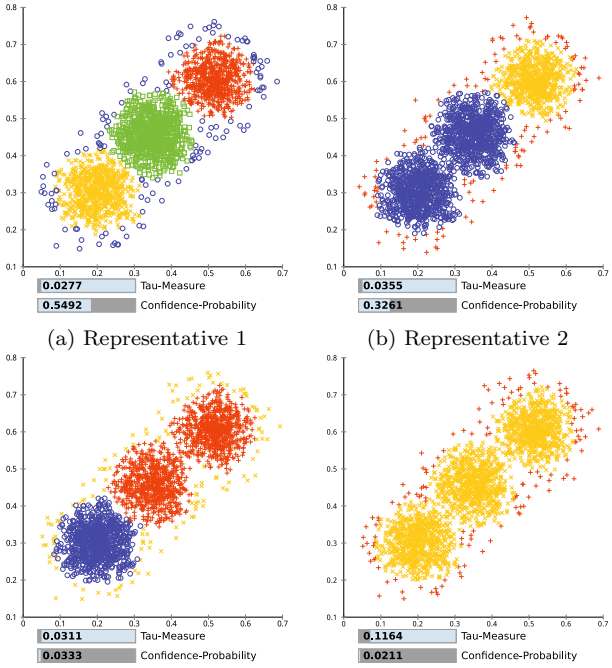
of possible clusterings $PC = \{\mathcal{C}(X_1), ..., \mathcal{C}(C_N)\}$. **Third**, using a distance measure for clusterings, *dist*, we can then apply any distance-based clustering algorithm $\mathcal{C}'$ to cluster the set of clusterings PC. For the choice of *dist*, ELKI allows to choose between pre-implemented distance measures for clusters, including Adjusted Rand Index [17], Fowlkes-Mallows index [16], BCubed [7], mutual-information-based, and entropy-based measures, and can easily be extended with custom measures (see also [1]). For the choice of $\mathcal{C}'$, ELKI allows to choose between various distance-based clustering algorithms. **Fourth**, for each of the resulting meta-clusters returned by $\mathcal{C}'$, we select a *$\tau\phi$-representative clustering* $\mathcal{C}(X_i) \in PC$, such that we can guarantee that, at a user-specified level of significance α, the probability that the (unknown) ground-truth has a distance (in terms of the chosen distance measure *dist*) to X_i of at most τ is at least ϕ [22]. **The resulting** set of $\tau\phi$-representatives can be used for educated decision making, as the returned parameters τ and ϕ allow to assess the quality of a clustering.

Since the three parts clustering algorithm $\mathcal{C}$, distance measure *dist* and distance-based clustering algorithm $\mathcal{C}'$ (shown in orange in Figure 3) in *representative clustering* are modular, ELKI is a perfect implementation environment. ELKI is not constrained to using numerical vector types, and thus clustering *results* can be treated as first-class citizens, and we can use appropriate distance functions to run clustering algorithms on the results to obtain clusters of clustering results. Additional benefits of ELKI are the large number of alternatives that are already present and its easy extensibility for inclusion of further alternatives.

3. DEMONSTRATION SCENARIO

The audience will be presented with the functionality of ELKI 0.7. Specifically, we will focus on the uncertain data mining/clustering aspect. The effect of different parameter settings will be elaborated on three interesting datasets.

To illustrate the concepts behind the demonstration we start by presenting the clustering of a small uncertain dataset. The dataset consists of uncertain objects that are arranged in three Gaussian distributions (A, B, C) next to each other. After illustrating the original clustering, that separates the objects in A, B and C (cf. Figure 2), we will sample some possible datasets and show their clustering result. It will clearly be visible that the larger the uncertainty of the objects, the more diverse the clustering results. After showing the result from the clustering of the center of mass points, we will present the *representative clustering*. The result of *representative clustering* consists of the intuitive possible results of the clustering of the uncertain data. Specifically, it will contain the clustering where A, B, and C are separated (cf. Figure 4a), A and B merge (cf. Figure 4b), B and C



(a) Representative 1 (b) Representative 2

(c) Representative 3 (d) Representative 4

Figure 4: The four representatives of the dataset

merge (cf. Figure 4c), and A, B, and C merge (cf. Figure 4d), each annotated with the similarity τ and confidence ϕ. We see that the first representative (Figure 4a) has the highest probability, whereas the latter two representatives only represent a small fraction of possible worlds. This part of the demonstration should give an intuitive introduction to the framework. Next we will show the validity of the approach on low dimensional real datasets. The variety of possible clusterings will be shown to increase. We will utilize different datasets here which will make the use of different clustering algorithms necessary. We will show the impact of centroid-based approaches (e.g., k-means) versus density-bases approaches (e.g., DBSCAN) which tend to have less variability in their clustering results.

Lastly, we will increase the complexity of the demonstration and apply the framework to multi-dimensional datasets. This demonstration will show the full potential of the ELKI framework since it is able to visualize (uncertain) clustering results even of high-dimensional spaces through 2-dimensional projections and parallel-coordinate [18] techniques.

4. AVAILABILITY

ELKI is an open source (AGPLv3) data mining software written in Java, actively and continuously developed since years. A growing community uses ELKI in related research areas such as databases and data mining as well as in other research areas where data mining methods are applied. Our visualization and evaluation tool will be available along with ELKI release 0.7 at: `http://elki.dbs.ifi.lmu.de/`

Acknowledgements: Part of this project is funded by the Deutsche Forschungsgemeinschaft (DFG) under grant number RE 266/5-1.

5. REFERENCES

[1] E. Achtert, S. Goldhofer, H.-P. Kriegel, E. Schubert, and A. Zimek. Evaluation of clusterings – metrics and visual support. In *Proc. ICDE*, pages 1285–1288, 2012.

[2] E. Achtert, H.-P. Kriegel, L. Reichert, E. Schubert, R. Wojdanowski, and A. Zimek. Visual evaluation of outlier detection models. In *Proc. DASFAA*, 2010.

[3] E. Achtert, H.-P. Kriegel, E. Schubert, and A. Zimek. Interactive data mining with 3D-Parallel-Coordinate-Trees. In *Proc. SIGMOD*, pages 1009–1012, 2013.

[4] C. C. Aggarwal. *Managing and Mining Uncertain Data*. Springer, 2010.

[5] P. Agrawal, O. Benjelloun, A.D. Sarma, C. Hayworth, S. Nabar, T. Sugihara, and J. Widom. Trio: A system for data, uncertainty, and lineage. In *VLDB*, 2006.

[6] L. Antova, T. Jansen, C. Koch, and D. Olteanu. Fast and simple relational processing of uncertain data. In *Proc. ICDE*, 2008.

[7] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL*, pages 79–85, 1998.

[8] Y. Benjamini. Simultaneous and selective inference: Current successes and future challenges. *Biometrical J.*, 52(6):708–721, 2010.

[9] O. Benjelloun, A. D. Sarma, A. Halevy, and J. Widom. ULDBs: Databases with uncertainty and lineage. In *Proc. VLDB*, pages 1249–1264, 2006.

[10] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Züfle. Probabilistic frequent itemset mining in uncertain databases. In *Proc. KDD*, 2009.

[11] C. Böhm, A. Pryakhin, and M. Schubert. The Gauss-tree: Efficient object identification of probabilistic feature vectors. In *Proc. ICDE*, page 9, 2006.

[12] M. Chau, R. Cheng, B. Kao, and J. Ng. Uncertain data mining: An example in clustering location data. In *Proc. PAKDD*, pages 199–204, 2006.

[13] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *VLDB J.*, 16:523–544, 2007.

[14] N. N. Dalvi, C. Ré, and D. Suciu. Probabilistic databases: diamonds in the dirt. *Commun. ACM*, 52(7):86–94, 2009.

[15] J. Dunn. Well separated clusters and optimal fuzzy partitions. *J. Cybernet.*, 4:95–104, 1974.

[16] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *JASA*, 78(383):553–569, 1983.

[17] L. Hubert and P. Arabie. Comparing partitions. *J. Classif.*, 2(1):193–218, 1985.

[18] A. Inselberg. *Parallel coordinates: visual multidimensional geometry and its applications*. Springer, 2009.

[19] H.-P. Kriegel and M. Pfeifle. Density-based clustering of uncertain data. In *Proc. KDD*, pages 672–677, 2005.

[20] M. F. Mokbel, C.-Y. Chow, and W. G. Aref. The new casper: A privacy-aware location-based database server. In *Proc. ICDE*, pages 1499–1500, 2007.

[21] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip. Efficient clustering of uncertain data. In *Proc. ICDM*, pages 436–445, 2006.

[22] A. Züfle, T. Emrich, K. A. Schmid, N. Mamoulis, A. Zimek, and M. Renz. Representative clustering of uncertain data. In *Proc. KDD*, pages 243–252, 2014.