Handling Uncertainty in Clustering Art-exhibition Visiting Styles

Francesco Gullo[1], Giovanni Ponti[2], Andrea Tagarelli[3], Salvatore Cuomo[4],
Pasquale De Michele[4], and Francesco Piccialli[5]

[1] UniCredit, R&D Department, Via Molfetta 101, I-00171 Rome
gullof@acm.org
[2] DTE-ICT-HPC, ENEA Portici, P.le E. Fermi 1, I-80055 Portici (NA)
giovanni.ponti@enea.it
[3] DIMES, University of Calabria, I-87036 Rende (CS)
andrea.tagarelli@unical.it
[4] DMA, University of Naples "Federico II", Via Cupa Cintia 21, I-80126, Naples
{salvatore.cuomo,pasquale.demichele}@unina.it
[5] DIETI, University of Naples "Federico II", Via Mezzocannone 8, I-80100, Naples
francesco.piccialli@unina.it

Abstract. Uncertainty is one of the most critical aspects that affect the quality of Big Data management and mining methods. Clustering uncertain data has traditionally focused on data coming from location-based services, sensor networks, or error-prone laboratory experiments. In this work we study for the first time the impact of clustering uncertain data on a novel context consisting in visiting styles in an art exhibition. We consider a dataset derived from the interaction of visitors of a museum with a complex Internet of Things (IoT) framework. We model this data as a set of uncertain objects, and cluster them by employing the well-established UK-medoids algorithm. Results show that clustering accuracy is positively impacted when data uncertainty is taken into account.

Key words: uncertain objects, clustering, data mining, cultural heritage data

1 Introduction

In the last decade, "Veracity" has been named the fourth "V" referred to the Big Data paradigm in addition to Volume, Velocity and Variety. This attribute emphasizes the importance addressing managing of uncertainty inherent within several types of data. It refers to the level of reliability associated with certain types of data. In this scenario, handling uncertainty in data management requires more and more importance if we consider the wide range of Big Data applications. Some data can be considered inherently uncertain, for example: sentiment in humans; GPS sensors bouncing among the skyscrapers of New York; weather conditions; and clearly the future. 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 [22]. 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.

Clustering of uncertain objects has traditionally been employed to categorize data coming from location-based services, sensor networks, or error-prone laboratory experiments. In this work we focus for the first time on studying how handling data uncertainty impacts the performance of clustering methods in a novel context of visiting styles in art exhibition. We consider a dataset derived from the analysis of how visitors of a museum interact with mobile devices such as smartphones or tablets. We model this data as a set of uncertain objects, and apply the UK-medoids algorithm [19] to obtain clusters of similar visiting styles. We compare such a visiting-style grouping with a ground truth obtained by a well-established classification methodology, which classifies visiting styles into four categories (ant, butterfly, fish, grasshopper) based on the values of some exemplar parameters, such as the percentage of viewed artworks or the average time spent in interacting with artworks [5, 8, 9, 10]. F-measure results confirm the claim that clustering accuracy increases when data uncertainty is taken into account in the process.

The rest of the paper is organized as follows: Section 2 describes some preliminaries on clustering techniques of uncertain data and Section 3 presents the case study. Moreover in Section 4 we report some experiments on accuracy and efficiency of K-medoids algorithm applied to our case study. Finally conclusions close the paper.

2 Preliminaries on clustering of uncertain data

Data clustering is a central problem in pattern recognition, knowledge discovery, and data management disciplines. Given a set of objects represented in a multidimensional space, the objective is to infer an organization for these objects into groups, also called *clusters*, according to some notion of affinity or proximity among the objects. Two general desiderata for any clustering algorithm is that each of the discovered clusters should be cohesive (i.e., comprised of objects that are very similar to each other) and that the clusters are well-separated from each other. A major family of clustering algorithms is referred to as partitional clustering [2, 7, 17], whose general scheme is to produce a partitioning of the input set of objects by iteratively refining the assignment of objects to clusters based on the optimization of some criterion function. This approach can be computationally efficient when a proper notion of cluster prototype is defined and used to drive the assignment of objects to clusters. Typically, a cluster prototype is defined as the mean object in the cluster (centroid), or an object that is closest to each of the other objects in the cluster (medoid). K-means [20] and K-medoids [19] are two classic algorithms that exploit the notions of centroid and medoid, respectively.

In this paper we exploit a clustering approach originally designed in the research of uncertain data mining. To this purpose, we can refer to a relatively large corpus of studies developed in the last decade [3, 13, 14, 15, 16, 18, 25]. In this work we focus on the uncertain counterpart of K-medoids, named *UK-medoids*, which was proposed in [13]. This algorithm overcomes two main issues of the uncertain K-means (UK-means) [3]: (i) the centroids are regarded as deterministic objects obtained by averaging the expected values of the pdfs of the uncertain objects assigned to a cluster, which may result in loss of information; (ii) the adopted Expected Distance between centroids and uncertain objects requires numerical integral estimations, which are computationally inefficient.

Given a dataset D of uncertain objects and a number k of desired output clusters, the UK-medoids algorithm starts by selecting a set of k initial medoids (uniformly at random or, alternatively, by any ad-hoc strategy for obtaining well-separated medoids). Then, it iterates through two main steps. In step 1, every object is assigned to the cluster corresponding to the medoid closest to the object. In step 2, all cluster medoids are updated to reflect the object assignments of each cluster. The algorithm terminates when cluster stability is reached (i.e., no relocation of objects has occurred with respect to the previous iteration).

One of the strength point of UK-medoids is that it employs a particularly accurate distance function designed for uncertain objects, which hence overcomes the limitation in accuracy due to a comparison of the expected values of the object pdfs. Also, the uncertain distance for every pair of objects are computed once in the initial stage of the algorithm, and subsequently used at each iteration. The combination of the above two aspects has shown that UK-medoids outperforms UK-means in terms of both effectiveness and efficiency.

3 A case study: styles of visit in an art show

As case study has been considered the art show "*The Beauty and the Truth*"[1]. Here, Neapolitan works of art dating from the late XIX and early XX centuries have been shown. The sculptures have been exposed in the monumental complex of *San Domenico Maggiore*, located in the historical centre of Naples. During the event, we have collected log files related to 253 visitors thanks to an Internet of Things deployed framework [4, 6]. The analysis of their behaviours within the cultural space has enabled us to define a classification of the visiting styles. In the literature, there exist several research papers that focus on this objective.

As a starting point for our classification we have considered the work in [23], where authors have proposed a classification method based on a comparison between behaviours of museum visitors and four "typical" animals (i.e., ant, fish, butterfly and grasshopper). Moreover, we have resorted to the work presented in [24], where, recalling the above mentioned approach, authors have introduced a methodology based on two unsupervised learning approaches for validating empirically their model of visiting styles. Finally, in [5, 8, 9, 10], we have proposed a classification technique able to discover how visitors interact with a complex Internet of Thinghs (IoT) framework, redefining the visiting styles' definition. We have considered the behaviours of spectators in connection with the use of the available supporting technology, i.e., smart-phones, tablets and other devices. For completeness, we report a brief description below.

A visitor is considered:

- an *ant* (**A**), if it tends to follow a specific path in the exhibit and intensively enjoys the furnished technology;
- a *butterfly* (**B**), if it does not follow a specific path but rather is guided by the physical orientation of the exhibits and stops frequently to look for more media contents;
- a *fish* (**F**), if it moves around in the center of the room and usually avoids looking at media content details;
- a *grasshopper* (**G**), if it seems to have a specific preference for some preselected artworks and spends a lot of time observing the related media contents.

The four visiting styles are characterized by three different parameters, assuming values in $[0, 1]$: a_i, τ_i and v_i. More in detail, for the i-th visitor, we denote by:

- a_i, the percentage of viewed artworks;
- τ_i, the average time spent by interacting with the viewed artworks;
- v_i, that measures the quality of the visit, in terms of the sequence of crossed sections (i.e., path).

The classification of the visiting styles is obtained following the scheme summarized in Table 1.

[1] http://www.ilbellooilvero.it

Table 1. Characterization of the visiting styles.

Visiting Style	a_i	τ_i	v_i
A	≥ 0.1	negligible	≥ 0.58
B	≥ 0.1	negligible	< 0.58
F	< 0.1	< 0.5	negligible
G	< 0.1	≥ 0.5	negligible

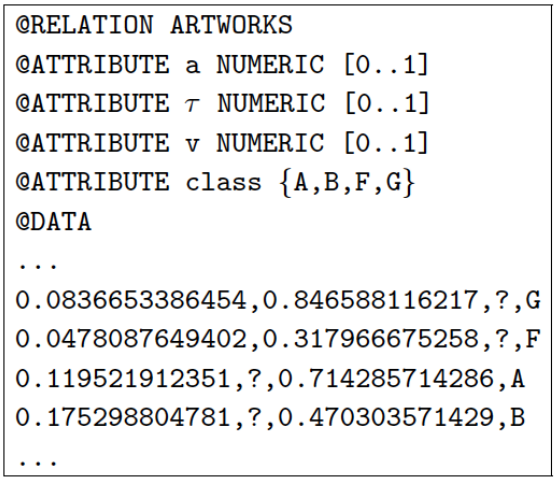
As we can observe, values $a_i \geq 0.1$ characterize both **A**s and **B**s, while values $a_i < 0.1$ are related to **F**s and **G**s. Moreover, the parameter τ_i does not influence the classification of **A**s and **B**s, while values $\tau_i < 0.5$ are typical for **F**s and values $\tau_i \geq 0.5$ are inherent in **G**s. Finally, the parameter v does not influence the classification of **F**s and **G**s, whereas values $v \geq 0.58$ are related to **A**s and values $v < 0.58$ characterize **B**s. We recall that, each parameter is associated with a numerical value normalized between 0 and 1. The thresholds values $\bar{a} = 0.1$, $\bar{\tau} = 0.5$ and $\bar{v} = 0.58$ have been obtained after a tuning step, in which we have resorted to the K-means clustering algorithm to discover data groups reflecting visitors' behaviours in all the sections of the exhibit. More details, about how these values have been set, are reported in [11].

4 Experimental evaluation

We devised an experimental evaluation aimed to assess the ability in clustering uncertain objects of the algorithm proposed in [13] and discussed in Section 2 We consider the dataset derived from the analysis of how visitors of a museum interact with an IoT framework, according to the methodology described in Section 3. We model this data as a set of uncertain objects, and apply the UK-medoids algorithm [19] to obtain clusters of similar visiting styles. The ultimate goal of our evaluation is to compare such a visiting-style grouping with a ground truth obtained by a well-established classification methodology defined [5, 8, 9, 10] (described in Section 3), and show that our method outperforms a baseline clustering method that does not take uncertainty into consideration.

4.1 Evaluation methodology

Dataset. Experiments were executed by exploiting the dataset populated with data coming from the above mentioned log files. In the following, we report a description of the dataset resorting, for simplicity of representation, to the ARFF Weka format (see Figure 1 for more details). Notice that, the dataset is characterized by: (i) 253 objects (i.e., the visitors); (ii) 3 attributes (i.e., a, τ, and v), which reflect, for each visitor, the parameters a_i, τ_i and v_i, described in Section 3; (iii) 4 classes (i.e, `A`, `B`, `F` and `G`), corresponding to the already cited typical animals. Moreover, observe that tuples contain the symbol "`?`" for some attribute values that are not significant for the classification. In other words, accordingly with the classification rules summarized in Table 1, for **A**s and **B**s we neglect attribute `tau` and for **F**s and **G**s we neglect attribute `v`.



```
@RELATION ARTWORKS
@ATTRIBUTE a NUMERIC [0..1]
@ATTRIBUTE τ NUMERIC [0..1]
@ATTRIBUTE v NUMERIC [0..1]
@ATTRIBUTE class {A,B,F,G}
@DATA
...
0.0836653386454,0.846588116217,?,G
0.0478087649402,0.317966675258,?,F
0.119521912351,?,0.714285714286,A
0.175298804781,?,0.470303571429,B
...
```

Fig. 1. The dataset in the ARFF Weka format

The selected dataset is originally composed by deterministic values. For this reason, we needed to synthetically generate the uncertainty. Notice that, in order to adapt the dataset to the algorithm in [13], the neglected values have been substituted with the numerical approximation 0.0. In substance, this can be assimilated to a first kind of perturbation.

For the univariate case, we needed to define the region for the interval of uncertainty $I^{(h)}$ and the related pdf $f^{(h)}$ for the region $I^{(h)}$, for all the $a^{(h)}, h \in [1..m]$ attributes of the o object. We randomly chose the interval region $I^{(h)}$ as in the subinterval $[min_{o_h}, max_{o_h}]$, and these two boundaries are the minimum (i.e., min_{o_h}) and the maximum (i.e., max_{o_h})) deterministic values for h (i.e., the attribute) taking into account the objects that are part of the same ideal classification for o. Regarding $f^{(h)}$, a continuous formulation of the density function has been taken into account, that is *Uniform*, together with a discrete mass function, that is *Binomial*. We properly set the parameters for the Binomial distribution in order to have the mode in correspondence of the original deterministic value of the attribute h-th of the o object.

Clustering validity criteria. In order to evaluate quality the clustering in output, we resorted to the availability of the classification originally provided in the dataset. Indeed, following the natural cluster paradigm, the higher the clustering solution is similar to the reference classification, the higher is the quality achieved. *F-measure* [21] is a well known external criterion used to evaluate clustering solution, which exploits *Recall* and *Precision* notions from the Information Retrieval field.

Overall Recall (R) and Precision (P) can be computed by means of a macro-averaging strategy performed on local values as:

$$R = \frac{1}{H}\sum_{i=1}^{H} \max_{j\in[1..K]} R_{ij}, \qquad P = \frac{1}{H}\sum_{i=1}^{H} \max_{j\in[1..K]} P_{ij},$$

Overall F-measure is defined as the harmonic mean of P and R as:

$$F = \frac{2PR}{P+R}$$

Settings. The calculation of the distances involves integral computation, and we do it by exploiting the sample list coming from the pdfs. We resorted to a sampling method based on the classical *Monte Carlo*.[2] A tuning phase has been preliminary done in order to set in the proper way the sample number S; the strategy was based on a choice of S producing an accuracy level that another $S' > S$ was not able to improve significantly.

4.2 Results

Accuracy. Accuracy tests have the objective evaluate the impact of dealing with uncertainty in a clustering-based analysis. For this reason, we are interesting in comparing clustering results achieved by our UK-medoids algorithm on the dataset with uncertainty w.r.t. the ones achieved by K-medoids algorithm on the dataset with deterministic values.

In Table 2 we report only results on the univariate model (multivariate model carried out similar results). More in detail, here we highlight the differences, in terms of F-measure percentage gains, between UK-medoids (both binomial and uniform) and deterministic K-medoids. It can be observed that UK-medoids achieves higher accuracy results w.r.t. K-medoids, that are slight for binomial distribution (0.043%), but relevant for uniform one (6.227%). In general, we can notice that introducing uncertainty in the dataset and handling it in the clustering task with our proper UK-medoids algorithm leads to improve the effectiveness of the results.

Efficiency. To evaluate the efficiency of UK-medoids, we measured time performances in clustering uncertain objects.[3] Figure 2 shows the total execution times (in milliseconds) obtained by UK-medoids on our dataset. Notice that, we calculated the sum of the times obtained for the pre-computing phase (i.e., uncertain distances computation), together with the algorithm runtimes. Here,

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

[3] Experiments were conducted on an ENEA server of CRESCO4 HPC cluster hosted in Portici [12] – http://www.cresco.enea.it/

Table 2. UK-medoids' performance results compared with deterministic K-medoids in terms of F-measure percentage.

pdf	*UK-medoids gain*
Binomial	0.04298805%
Uniform	6.22712192%

it can be noted that by using a uniform pdf we obtain execution times about 2 times faster than those achieved with a binomial pdf. This is due to the fact that a binomial pdf requires to process a higher number of samples w.r.t. a uniform pdf.

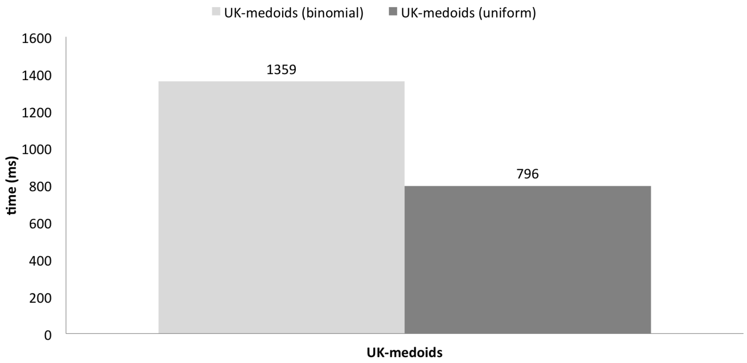


Fig. 2. Clustering time performances

5 Conclusion

In this paper we addressed the topic of how data collected by an IoT system through mobile devices in a cultural environment could be opportunely exploited and analysed. The main goal is to infer useful knowledge about visitors. Real data are generally affected of a large degree of uncertainness and to deal with this drawback, here we propose a clustering approach based on K-medoids algorithms. Nevertheless the limitation of a not very large dataset, first results encourage us to deeply investigate this approach, in order to better analyse data collected from a real cultural heritage scenario. Moreover, with the aim to improve the performance of the proposed method, in future works we will intend to better adapt the uncertain interval and the pdf, defined on this set, to our problem.

References

1. Aggarwal, C.C.: Managing and Mining Uncertain Data, Advances in Database Systems, vol. 35. Kluwer (2009), http://dx.doi.org/10.1007/978-0-387-09690-2
2. Bello-Orgaz, G., Jung, J., Camacho, D.: Social big data: Recent achievements and new challenges. Information Fusion 28, 45–59 (2016)
3. Chau, M., Cheng, R., Kao, B., Ng, J.: Uncertain Data Mining: An Example in Clustering Location Data. In: Proc. PAKDD Conf. pp. 199–204 (2006)
4. Chianese, A., Marulli, F., Piccialli, F., Benedusi, P., Jung, J.: An associative engines based approach supporting collaborative analytics in the internet of cultural things. Future Generation Computer Systems (2016)
5. Chianese, A., Piccialli, F.: Improving user experience of cultural environment through iot: The beauty or the truth case study. Smart Innovation, Systems and Technologies 40, 11–20 (2015)
6. Chianese, A., Piccialli, F., Riccio, G.: Designing a smart multisensor framework based on beaglebone black board. Lecture Notes in Electrical Engineering 330, 391–397 (2015)
7. Cuomo, S., De Michele, P., Galletti, A., Piccialli, F.: A cultural heritage case study of visitor experiences shared on a social network. pp. 539–544 (2015)
8. Cuomo, S., De Michele, P., Galletti, A., Pane, F., Ponti, G.: Visitor dynamics in a cultural heritage scenario. In: DATA 2015 - Proceedings of 4th International Conference on Data Management Technologies and Applications, Colmar, Alsace, France, 20-22 July, 2015. pp. 337–343 (2015), http://dx.doi.org/10.5220/0005579603370343
9. Cuomo, S., De Michele, P., Galletti, A., Ponti, G.: Visiting styles in an art exhibition supported by a digital fruition system. In: 11th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2015, Bangkok, Thailand, November 23-27, 2015. pp. 775–781 (2015), http://dx.doi.org/10.1109/SITIS.2015.87
10. Cuomo, S., De Michele, P., Galletti, A., Ponti, G.: Data Management Technologies and Applications: 4th International Conference, DATA 2015, Colmar, France, July 20-22, 2015, Revised Selected Papers, Communications in Computer and Information Science, vol. 584, chap. Classify Visitor Behaviours in a Cultural Heritage Exhibition, pp. 17–28. Springer International Publishing (2016), http://link.springer.com/chapter/10.1007/978-3-319-30162-4_2
11. Cuomo, S., De Michele, P., Galletti, A., Ponti, G.: Intelligent Interactive Multimedia Systems and Services 2016, Smart Innovation, Systems and Technologies, vol. 55, chap. Influence of Some Parameters on Visiting Style Classification in a Cultural Heritage Case Study, pp. 567–576. Springer International Publishing (June 2016), http://dx.doi.org/10.1007/978-3-319-39345-2_50, iIMSS - IS07: Internet of Things: Architecture, Technologies and Applications Invited Session of KES 2016
12. G. Ponti et al.: The role of medium size facilities in the HPC ecosystem: the case of the new CRESCO4 cluster integrated in the ENEAGRID infrastructure. In: International Conference on High Performance Computing & Simulation, HPCS 2014, Bologna, Italy, 21-25 July, 2014. pp. 1030–1033 (2014)
13. Gullo, F., Ponti, G., Tagarelli, A.: Clustering Uncertain Data Via K-Medoids, pp. 229–242. Springer Berlin Heidelberg, Berlin, Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-87993-0_19

14. Gullo, F., Ponti, G., Tagarelli, A.: Minimizing the variance of cluster mixture models for clustering uncertain objects. In: Proc. IEEE ICDM Conf. pp. 839–844 (2010)
15. 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)
16. Gullo, F., Tagarelli, A.: Uncertain centroid based partitional clustering of uncertain data. PVLDB 5(7), 610–621 (2012)
17. Jain, A., Dubes, R.: Algorithms for Clustering Data. Prentice-Hall (1988)
18. Jiang, B., Pei, J., Tao, Y., Lin, X.: Clustering uncertain data based on probability distribution similarity. IEEE Trans. Knowl. Data Eng. 25(4), 751–763 (2013)
19. L. Kaufman and P. J. Rousseeuw: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley (1990)
20. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proc. Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297 (1967)
21. van Rijsbergen, C.J.: Information Retrieval. Butterworths (1979)
22. Sarma, A.D., Benjelloun, O., Halevy, A.Y., Nabar, S.U., Widom, J.: Representing uncertain data: models, properties, and algorithms. VLDB J. 18(5), 989–1019 (2009), `http://dx.doi.org/10.1007/s00778-009-0147-0`
23. Veron, E., Levasseur, M., Barbier-Bouvet, J.: Ethnographie de l'exposition. Paris, Bibliothèque Publique d'Information, Centre Georges Pompidou (1983)
24. Zancanaro, M., Kuflik, T., Boger, Z., Goren-Bar, D., Goldwasser, D.: Analyzing museum visitors' behavior patterns. In: User Modeling 2007, 11th International Conference, UM 2007, Corfu, Greece, June 25-29, 2007, Proceedings. pp. 238–246 (2007)
25. Züfle, A., Emrich, T., Schmid, K.A., Mamoulis, N., Zimek, A., Renz, M.: Representative clustering of uncertain data. In: Proc. KDD Conf. pp. 243–252 (2014)