Approaches of Handling Uncertain Time Series Data towards Prediction

Radzuan M. F. Nabilah, Zalinda Othman, and Bakar A. Azuraliza

***Abstract*—This paper works on clustering issues of uncertain time series data prior to prediction process. The aim of uncertainty analysis is to determine how to deal with uncertain data in order to gain knowledge, fit low dimensional model, and to predict. So as to gain a reliable prediction, uncertainty in data could not be ruled out because it may bring important knowledge. Clustering as a step before prediction process can be seen as the most popular representative of unsupervised learning, while classification together with regression are possibly the most frequently considered tasks in supervised learning. Clustering uncertain time series data posts significant challenges on both modeling similarity between uncertain objects and developing efficient computational methods. This work will benefit in many application domains.**

***Index Terms*—Clustering, prediction, time series data, uncertain time series data, uncertainty.**

I. Introduction

Clustering is among the most important problem task in data analysis, data mining and machine learning. In fact, while clustering can be seen as the most popular representative of unsupervised learning, classification together with regression is possibly the most frequently considered task in supervised learning [1]. Uncertain time series is believed to be able to avoid risks and help in making better daily decisions, can improve the quality of demand, and identify temporal patterns that emerge and persist [2]. Clustering on uncertain data, one of the essential tasks in mining uncertain data, posts significant challenges on both modeling similarity between uncertain objects and developing efficient computational methods.

The previous studies on clustering uncertain data are largely focus on various extensions of the traditional clustering algorithms designed for certain data. As an object in a certain data set is a single point, the distribution regarding the object itself is not considered in traditional clustering algorithms. Thus, the studies that extended traditional algorithms to cluster uncertain data are limited to using geometric distance based similarity measures, and cannot capture the difference between uncertain objects with different distributions [3]. Clustering uncertain data has been well recognized as an important issue [4], [5]. Generally, an uncertain data object can be represented by a probability distribution [6], [7].

The structure of this paper is organized as follows. After the introduction, the main aims of the paper and briefly related researches will be defined and described in Part II, including the researches related to uncertainty modeling of time series in clustering. Then, scability and clustering process will be explained in Part III. In Part IV, a discussion on prediction of uncertain time series data and the objectives of the investigated research are provided. The flow of approaches will be shown in Part V. The final section of the paper contains the conclusions and references.

II. Research Related

Time series is well known as a stretch of values on a similar scale, indexed by a time that occurs naturally in many application domains such as weather, manufacturing, environmental, economic, finance, and medicine. All data of time series can be processed in order to gain knowledge for future used. Time series mining is one of temporal data mining applications that can help in extracting knowledge. In time of focusing the times series data, there is existence of uncertainty in time series.

Uncertainty is a basic feature of automatic and semi-automatic data processes [8]. There are many solutions have been proposed in order to reduce uncertainty because of risk in losing information and misleading results [9]. The uncertain time series is also a non-negative and precisely different ways in some fields. Particularly, uncertain data refers to data in which the ambiguity on whether it takes place or not, the existence of the data for the particular attribute values are not ascertained with 100 percent probability [10].

Besides, uncertainty exists in a modeling process, in which it arises from the fundamental choice as seen in grid resolution and from the parameterization of processes unresolved at the grid scale [11]. Also, as example, a high uncertainty brings a big impact on prediction of regional climate change [12]. Then, in fact, a lack of model diversity can cause a limited range of projections in climate change [13]. Meanwhile, the distinct source of uncertainty in prediction includes internal variability, model uncertainty or response uncertainty, and scenario uncertainty [12], [14].

The uncertainty has been explicitly indicated as one of the future challenges in many fields [15]. The uncertainty is present in all data processes and methods. The characterizing feature of this and other early works using uncertainty theories is after probabilities have been evaluated and a threshold is used to select matching and non-matching objects [16]. Therefore, the uncertainty generated during data

Manuscript received October 1, 2016; revised December 29, 2016. This work was supported in part by the Universiti Kebangsaan Malaysia (UKM), Malaysia through grant scheme number of Project FRGS/1/2014/ICT02/UKM/02/1 for financial support.

The authors are with the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Malaysia (e-mail: nabilah.filzahr@gmail.com, zalinda@ukm.edu.my, azuraliza@ukm.edu.my).

doi: 10.18178/ijfcc.2016.5.6.477

processes is missing [17], [18]. There are relationships between uncertain and original series [19], [20]. A certain time series is extracted to represent the original uncertain time series [21]. Uncertain time series can be treated as positional uncertain vectors [22].

The combination of uncertainties is significant [23]–[25] in time series and brings important knowledge for end user. In order to gain benefits through uncertain time series data, the essential problem of uncertainty should be focused. The problem in related to time series data mining is how to manage the uncertain time series data during clustering process before next data handling. Therefore, this study aims to propose uncertainty modeling of time series during clustering.

III. Clustering

The uncertainty is not an error where error is referring to the indication of the wrongness of measurement [26]. While as the uncertainty in time series brings difference meaning as unique numbers of measured collections of data, which through technology nowadays, it is difficult to collect errors in database collections.

The accuracy of measurements affect all trade, commerce, safety, and many more especially prediction. The quality of these measurements is regulated by a variety important role included the measurement data and random variables, probability density functions, sampling distribution, estimation degrees of freedom and regression [27]. These are important to determine measurement uncertainty because i) it estimates the error associated with the measurement in numerical values, ii) it provides a level of confidence in one's measurement, iii) it is a good practice, and iv) it is required for laboratory validation process. As a theory, in order to determine uncertainty measurements, there are basic guidelines that can be followed [26]. There are some examples of typical factors that affecting the measurement included:

1) Environment (humidity, vibration, temperature)
2) Accuracy of measurement equipment
3) Stability
4) Instrument resolution
5) Instrument calibration
6) Repeatability
7) Reproducibility
8) Operator
9) Measurement setup
10) Method (procedure)
11) Software

Nowadays, handling uncertainty and modeling in an appropriate way normally comes with an increase computational complexity. Since time series datasets have a tendency to increase in size and computationally intense, resource bounded frameworks such as mining data [28] become increasingly relevant, the aspects of computational complexity. Therefore, scalability should always be kept in mind when developing methods for uncertainty handling in data analysis. Indeed, the aspect of scalability in general and its interaction with uncertainty in particular are important on ongoing research [29].

The idea of clustering is that grouping in a set of unlabeled data for each object of a given radius has to contain at least a minimum number of *n* objects [4]. Therefore throughout clustering process, in uncertainty, the functions in algorithm return values of the same type as the corresponding new module function (instead of the generally returning a value with a zero; +/- 0). A zero or non-uncertainty (certain value) is explicitly display as the integer 0 which understood the new group of data. Then abbreviations for the nominal value (*n*) are now available which means clustering process can be implemented.

Purposely, three principal categories exist in literature, namely i) partitioning clustering approaches [7], [30], ii) density-based clustering approaches [4], [5], and iii) possible world approaches [7]. The first two are along the line of the categorization of clustering methods for certain data [31], the possible world approaches are specific for uncertain data following the popular possible world semantics for uncertain data. As these approaches only examine the geometric properties of data objects and focus on instances of uncertain objects. They do not consider the similarity between uncertain objects in terms of distributions.

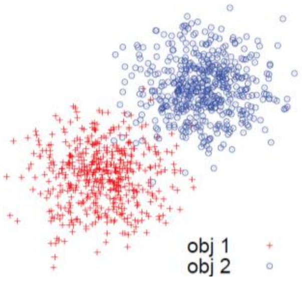

Fig. 1. The two objects have different geometric locations with probability density functions over the entire data space are different [3].

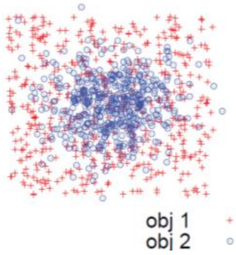

Fig. 2. The two objects are heavily overlapping with different distributions [3].

Partitioning clustering approaches extend the k-means method with the use of the expected distance to measure the similarity between two uncertain objects [7], [30]. Density-based clustering approaches extend the DBSCAN method and the OPTICS method [4], [5] in a probabilistic way [3]. The basic idea behind the algorithms does not change where objects in geometrically dense regions are grouped together as clusters and clusters are separated by sparse regions. However, the objects heavily overlap and no clear sparse regions to separate objects into clusters. Thus, the density-based approaches cannot work well.

On the other hands, a sampled possible world does not consider the distribution of a data object since a possible world only contains one instance from each object. The clustering results from different possible worlds can be drastically different. The most probable clusters calculated using possible worlds may still carry a very low probability. Therefore, the possible world approaches often cannot provide a stable and meaningful clustering result at the object level, not to mention that it is computationally infeasible due to the exponential number of possible worlds.

Uncertain objects can have any discrete or continuous distribution. The distribution differences cannot be captured by the previous methods based on geometric distances. In many case, the accurate probability distributions of uncertain objects are not known beforehand in practice. Instead, the probability distribution of an uncertain object is often derived from the observations of the corresponding random variable.

As example from [3], in Fig. 1, the two objects have different geometric locations. Their probability density functions over the entire data space are different. In Fig. 2, although the geometric locations of the two objects are heavily overlapping, they have different distributions. The difference between their distributions can also be discovered by extensions of the traditional clustering algorithms designed through captured by the extended methods.

IV. Prediction of Uncertain Time Series

Prediction of uncertain time series is important, for example in climate change problems. It influences the changeable climate that provides more useful information. Then, important knowledge can be tackled from this changeable gap that exists in uncertain time series data, in which the uncertainty can provide better results in terms of quality and efficiency [21], [32], [33]. The uncertain time series has been explored extensively in recent years.

Predicting uncertain time series appears to be a serious problem, as the existing forecast of certain time series does not purely mirror the ability of predicting future decisions, considering uncertainty in time series. The certain time series data cannot be implemented in large scale of dataset which bring error in prediction [25], [34], [35]. Uncertain time series in prediction is believed can avoid risks and help in making better daily decisions.

Hence, the determination of predicting uncertain time series should be noted as a serious action to be taken to improve the quality of many yields. Therefore, a comparison of the available methods should be carried out in determine the yield of predicting uncertain time series. Then, the limitation found from the analysis can be used as an opening of the experiment and aim at securing the limitation for enhancing the prediction outcome.

Previous studies have discovered some possible types of uncertainty in dataset. Also, clarification of uncertainty in dataset is important in identifying the type of data, so that they are not simply neglected. In normal practice, the organizers or any data collector will neglect any data that they perceive as ‘error’ without investigating uncertain data properties. Therefore, there is an initiative to work on uncertainty modeling in clustering for time series data. In conjunction to this, more work in progress study must be continued in achieve the aim of this study, which is to manage the uncertain time series data during clustering process.

V. Flow of Approaches

There are properties of uncertainty which included non-negative, loss value or null, truly different ways in a number of fields, data aggregation, privacy-preserving transforms, error-prone mining, positional uncertain vectors, exist in the modelling process where it arises from fundamental choice, and from the parameterization of processes unsolved at grid scale [2].

The data has gone through a discretization process (a process of organizing the dataset in minimizing redundancy and dependency, and makes it more informative to use). The discretization process involves scale-selective discretization (SSD) procedure as in [36]. This SSD separates small and large scales of the flow using a high-pass filter.

The flows of approaches help in order to propose uncertainty modeling of time series during clustering. The uncertain data is used to prove especially the accuracy of each prediction so that these methods can be studied for time series data. Fig. 3 shows the flow of approaches for this study starting from discretization of raw data until the prediction process.

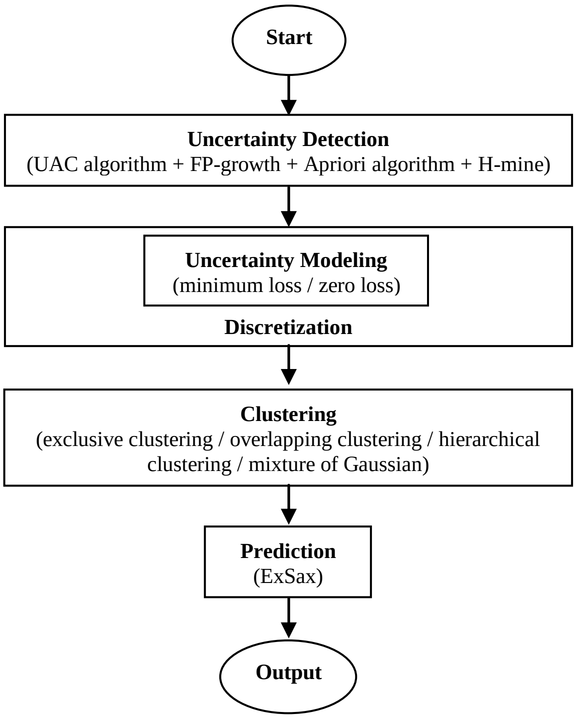
Start
Uncertainty Detection
(UAC algorithm + FP-growth + Apriori algorithm + H-mine)
Uncertainty Modeling
(minimum loss / zero loss)
Discretization
Clustering
(exclusive clustering / overlapping clustering / hierarchical clustering / mixture of Gaussian)
Prediction
(ExSax)
Output

Fig. 3. The flow of approaches.

VI. Conclusion

Time series acts as a stretch of values on a similar scale, indexed by a time that occurs naturally in many application domains. Time series mining is one of temporal data mining applications that can help in extracting knowledge. In time of focusing the times series data, the existence of uncertainty in time series could not be avoided. The essential problem in the circumstance of time series data mining is how to manage the uncertain time series data during clustering process before next data handling. Therefore, this study aims to propose uncertainty modeling of time series during clustering by taking steps throughout scability and detection of uncertainty in time series data before proceed to clustering process. Thus, the flows of approaches helps in order to propose uncertainty modeling of time series during clustering.

ACKNOWLEDGMENT

This work is acknowledged by Universiti Kebangsaan Malaysia (UKM), Malaysia through grant scheme number of Project FRGS/1/2014/ICT02/UKM/02/1 for financial support.

REFERENCES

[1] E. Hüllermeier, "Uncertainty in clustering and classification," Lecture *Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6379, 2010, pp. 16–19.

[2] N. F. M. Radzuan, Z. Othman, and A. A. Bakar, "Analysis of uncertainty in time series data: Issues and challenges," in *Proc. the Asian Conference on Technology, Information & Society 2014*, 2014, pp. 13–24.

[3] B. Jiang, J. Pei, Y. Tao, and X. Lin, "Clustering uncertain data based on probability distribution similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 751–763, Apr. 2013.

[4] H. Kriegel and M. Pfeifle, "Density-based clustering of uncertain data," in *Proc. the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD '05*, 2005, p. 672.

[5] H. Kriegel and M. Pfeifle, "Hierarchical density-based clustering of uncertain data," in *Proc. Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005, pp. 689–692.

[6] R. Cheng, D. V Kalashnikov, and S. Prabhakar, "Evaluation of probabilistic queries over imprecise data in constantly-evolving environments," *Inf. Syst.*, vol. 32, no. 1, pp. 104–130, Mar. 2007.

[7] P. B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner, "Clustering uncertain data with possible worlds," in *Proc. 2009 IEEE 25th International Conference on Data Engineering*, 2009, pp. 1625–1632.

[8] D. A. Keijzer, V. M. Keulen, and A. Dekhtyar, "Report on the first VLDB workshop on management of uncertain data (MUD)," *ACM SIGMOD Rec.*, vol. 36, no. 4, pp. 18–32, Dec. 2007.

[9] N. F. M. Radzuan, Z. Othman, and A. A. Bakar, "Uncertain time series in weather prediction," *Procedia Technol.*, vol. 11, pp. 557–564, 2013.

[10] M. Hooshsadat and O. R. Za, "An associative classifier for uncertain datasets," *Advances in Knowledge Discovery and Data Mining*, 2012, pp. 342-353.

[11] M. R. S. Angew, J. M. Murphy, D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, "Quantification of modelling uncertainties in a large ensemble of climate change simulations," *Nat. Publ. Gr.*, vol. 430, no. August 2004, pp. 768–772, 2004.

[12] E. Hawkins and R. Sutton, "The potential to narrow uncertainty in regional climate predictions," *Bull. Am. Meteorol. Soc.*, vol. 90, no. 8, pp. 1095–1107, Aug. 2009.

[13] C. Pennell and T. Reichler, "On the effective number of climate models," *J. Clim.*, vol. 24, no. 9, pp. 2358–2367, May 2011.

[14] J. C. Hargreaves, "Skill and uncertainty in climate models," *Wiley Interdiscip. Rev. Clim. Chang.*, vol. 1, no. 4, pp. 556–564, Jul. 2010.

[15] A. Halevy and J. Ordille, "Data integration : The teenage years," in *VLDB '06 Proc. the 32nd International Conference on very Large Data Bases*, 2006, vol. 12, pp. 9–16.

[16] S. Hayne and S. Ram, "Multi-user view integration system (MUVIS): An expert system for view integration," in *[1990] Proc. Sixth International Conference on Data Engineering*, 1990, pp. 402–409.

[17] D. Dey and S. Sarkar, "Generalized normal forms for probabilistic relational data," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 3, pp. 485–497, May 2002.

[18] D. Florescu, D. Koller, and A. Y. Levy, "Using probabilitistic information in data integration," in *VLDB '97 Proc. the 23rd International Conference on very Large Data Bases*, 1997, pp. 216–225.

[19] C. Russa and K. Andrews, "AC 2010-1093 : Managing a digitization project : Issues for state agency publications with folded maps," *Am. Soc. Eng. Educ.*, pp. 1–6, 2010.

[20] L. Haitao and Z. Xiaofu, "Precipitation time series predicting of the chaotic characters using support vector machines," in *Proc. 2009 International Conference on Information Management, Innovation Management and Industrial Engineering*, 2009, pp. 407–410.

[21] Y. Zuo, G. Liu, X. Yue, W. Wang, and H. Wu, "Similarity matching over uncertain time series," *2011 Seventh Int. Conf. Comput. Intell. Secur.*, pp. 1357–1361, Dec. 2011.

[22] J. Abfalg, H. Kriegel, P. Kr, and M. Renz, "Probabilistic similarity search for uncertain time series," in *SSDBM 2009 Proc. the 21st International Conference on Scientific and Statistical Database Management*, 2009, pp. 435–443.

[23] P. S. Lykoudis, A. A. Argiriou, and E. Dotsika, "Spatially interpolated time series of δ18O in eastern mediterranean precipitation," *Glob. Planet. Change*, vol. 71, no. 3–4, pp. 150–159, Apr. 2010.

[24] H. L. Cloke and F. Pappenberger, "Ensemble flood forecasting: A review," *J. Hydrol.*, vol. 375, no. 3–4, pp. 613–626, Sep. 2009.

[25] V. Jankovic, "Science migrations: Mesoscale weather prediction from Belgrade to Washington, 1970–2000," *Soc. Stud. Sci.*, vol. 34, no. 1, pp. 45–75, Feb. 2004.

[26] J. L. Bucher, *The Metrology Handbook*, 2004.

[27] S. K. Kimothi, *Uncertainty of Measurements - Physical and Chemical Metrology - Impact and Analysis - Knovel*, American Society for Quality (ASQ), 2002.

[28] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams," *ACM SIGMOD Rec.*, vol. 34, no. 2, p. 18, Jun. 2005.

[29] A. Laurent and M. Lesot, *Scalable Fuzzy Algorithms for Data Management and Analysis : Methods and Design*, 2010.

[30] C. Jin, J. X. Yu, A. Zhou, and F. Cao, "Efficient clustering of uncertain data streams," *Knowl. Inf. Syst.*, vol. 40, no. 3, pp. 509–539, Sep. 2014.

[31] J. Han, M. Kamber, and J. Pei, *Data Mining : Concepts and Techniques*, Elsevier/Morgan Kaufmann, 2012.

[32] M. Dallachiesa, B. Nushi, K. Mirylenka, and T. Palpanas, "Uncertain time-series similarity : Return to the basics," in *Proc. the VLDB Endowment*, 2012, vol. 5, no. 11, pp. 1662–1673.

[33] M. Dallachiesa, B. Nushi, T. Palpanas, and K. Mirylenka, "Similarity matching for uncertain time series," in *Proc. the 2nd ACM SIGSPATIAL International Workshop on Querying and Mining Uncertain Spatio-Temporal Data - QUeST '11*, 2011, pp. 8–15.

[34] C. Qing, Z. Xiaoli, and Z. Kun, "Research on precipitation prediction based on time series model," in *Proc. 2012 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring*, 2012, pp. 568–571.

[35] D. J. Gagne, A. McGovern, and M. Xue, "Machine learning enhancement of storm scale ensemble precipitation forecasts," in *Proc. the 2011 Workshop on Knowledge Discovery, Modeling and Simulation*, 2011, p. 45.

[36] V. Vuorinen, M. Larmi, P. Schlatter, L. Fuchs, and B. J. Boersma, "A low-dissipative, scale-selective discretization scheme for the Navier–Stokes equations," *Comput. Fluids*, vol. 70, pp. 195–205, Nov. 2012.

Nabilah Filzah Mohd Radzuan received the BS degree in e-commerce information technology from University Malaysia Sabah, Sabag in 2009, and the MS degree in information technology, from University Utara Malaysia in 2012 and now doing the PhD degree in data mining at University Kebangsaan Malaysia.

Zalinda Othman received the BS degree in quality control and instrumentation from University Science Malaysia, Penang in 1994, and the MS degree in quality engineering, from University of Newcastle upon Tyne, United Kingdom, in 1996 and the PhD degree in artificial intelligence from University Science Malaysia, Penang, in 2002. She is Deputy Director of Career Department Center (UKM-KARIER) at Universiti Kebangsaan Malaysia, where she is currently an associate professor. Her main research topics are the study of optimization and data mining.

Azuraliza Abu Bakar is a professor in data mining at University Kebangsaan Malaysia. She received her PhD degree (artificial intelligence) from University Putra Malaysia in 2002. Her research interests are in time series data mining, outbreak detection and deviation detection model employing nature inspired computing technique.