International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:12, 2013

Representing Data without Lost Compression Properties in Time Series: A Review

Nabilah Filzah Mohd Radzuan, Zalinda Othman, Azuraliza Abu Bakar, Abdul Razak Hamdan

***Abstract*—**Uncertain data is believed to be an important issue in building up a prediction model. The main objective in the time series uncertainty analysis is to formulate uncertain data in order to gain knowledge and fit low dimensional model prior to a prediction task. This paper discusses the performance of a number of techniques in dealing with uncertain data specifically those which solve uncertain data condition by minimizing the loss of compression properties.

***Keywords*—**Compression properties, uncertainty, uncertain time series, mining technique, weather prediction.

I. Introduction

TIME series is known as a stretch of values on a similar scale, indexed by a time that occurs naturally in many application domains. There are two categories of time series dataset, i.e. certain time series dataset and uncertain time series dataset. A certain time series dataset is defined as a sequence of data points measured typically at successive points in time spaced at uniform time intervals. While, uncertain time series dataset is defined as a non-negative and precisely different ways in a number of fields [1], [2].

In reality, time series also deals with uncertainty. Particularly, uncertain dataset refers to data in which the ambiguity on whether it really takes place or not is present, or data for that the attribute values are not ascertained with 100 percent probability [3]. The combination of uncertainties are significant [1], [2], [4] and it brings important knowledge for the attached domain. Uncertain time series data also refers to continuous data [5], where collected data often inaccurate and are based on incomplete or inaccurate information.

Here, the uncertain data is considered as data for which we may not be sure about the observation whether it really took place or not, or data for which the attribute values are not ascertained with 100 percent probability [3]. The combination of uncertainties is significant and brings important knowledge for the area involved. According to [1], [7]–[10], frequent pattern mining algorithm is implemented in order to mine the uncertain data.

The data properties in dataset act as a mirror of information for each area involved. The aim of time series analysis is to explore time series data in order to gain knowledge, fit low dimensional models, and make predictions. Besides, the paper is also focusing on lost compression properties during data handling. Lost compression properties in data bring a big impact on prediction task.

This paper is more towards a domain of weather time series data and represents a brief overview of uncertain time series and methods being used especially in weather prediction. Besides, the discussion moves further on the performance of a number of techniques in dealing with uncertain time series data specifically those which solves uncertain data condition by minimizing the lost of compression properties

This paper is organized as follows. First, the related research is reviewed in Section II. Then, some analysis is compared in Sec. III. In Section IV, there is a discussion of techniques and benefits. Finally, the conclusion is drawn in Section V.

II. Related Research

Time series data is a class of temporal data objects that can be obtained from various fields such as environmental, financial, economics and medicine [6]. The data is continuous and large in data size, and collected at regular time intervals. One of the focus areas in time series is uncertain time series especially in weather. It is important in helping to build up the prediction and influences changeable climate that provides more useful information. Important knowledge can be tackled from this changeable gap that exists in uncertain time series data.

A. Uncertain Time Series in Weather

Weather area shows that the main problem of indexing uncertain time series are the low query selectivity due to the curve of positional, only correlated uncertain time series can be treated as positional uncertain vectors and none of the approaches such as statistical modeling and Gaussian; is able to support dynamic time warping [7]. Traditional distance measures such as Euclidean distance or dynamic time warping are not always effective for analyzing uncertain time series data cause of limitation in applicability [8]. The change of precipitation from climate are often complex, uncertain, and changeable with the dynamic characteristic and rendered as a complex nonlinear dynamic system [9].

As focus is on weather area, the artificial intelligence methods for weather prediction currently include model output statistics, fuzzy logic, expert systems, genetic algorithm, and particle swarm optimization [10]. Particle swarm optimization

Nabilah FilzahMohd Radzuan is with Data Mining and Optimization, Center for Artificial Intelligence Technology, Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia, 46300 Bangi, Selangor, Malaysia (phone: 603-8921-6182; fax: 603-8921-6184; e-mail: nabilah.filzah@ gmail.com).

Zalinda Othman, Azuraliza Abu Bakar and Abdul RazakHamdan are with Data Mining and Optimization, Center for Artificial Intelligence Technology, Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia, 46300 Bangi, Selangor, Malaysia (e-mail: {zalinda, aab, arh}@ftsm.ukm.my).

(PSO) considered as good method when played in certain time series data based on the yield of prediction.

Uncertainty analysis is one of the practices that have been used to validate the precipitation data apart from systematic bias, model output evaluation with different precipitation inputs and inter-comparisons of magnitude and spatial pattern. It is also used for analysis of spatial similarities between Next Generation Radar (NEXRAD) and North American Land Data Assimilation System (NLDAS) [11]. Table I shows the comparative study on this issue.

TABLE I
COMPARISON OF PREVIOUS METHOD FOR UNCERTAIN TIME SERIES IN WEATHER DOMAIN

Method / Technique	Advantage	Disadvantage
Euclidean distance (Dynamic Time Warping)	The techniques proposed for modelling and processing uncertain time series	-Not always effective; -Limitation in applying
PSO	-Applied for analysis and prediction data; -Provide experimental results; -Work in an unsupervised manner; -Discover relations hidden inside the dataset; -Decrease mean error	Implemented in certain (normal) time series data type
Regression model	-Forecast process; -Capture the main modes of spatial variability of isotopic composition of precipitation; -Improve the forecasting accuracy	-Need framework to assess prediction; -Difficult to predict climate because of the dynamic characteristics of the sample set; -Usually implement in certain time series
Monte Carlo Simulation	-Repetitive random sampling to subtract results; -Applied in prediction as alternative methods for human perception	Bias with the limited range
Monte Carlo approach	The treatment leads to nonlinear terms	The Monte Carlo approach is less appealing

B. The Performance of Frequent Pattern Mining Technique

According to [1], [7]–[10], frequent pattern mining algorithm has been implemented in order to mine the uncertain data. The approaches can be seen in Table II. Even though, there are problems that matter to uncertain data approaches, but these approaches supposedly can be implemented in uncertain time series data.

TABLE II
THE UNCERTAIN DATA APPROACHES

FP-growth (FP-tree)	H-mine (UH-mine)	Apriori (UApriori)
Efficient & scalable especially for dense dataset	Efficient & scalable specially for uncertain dataset	
- Lost of compression properties - The largest number of false positive is generated - The elimination of such candidates further affects the efficiency	- Can avoid generating a large number of candidate itemsets - Reduce memory requirements - Best trade-offs in terms of running time and memory usage	- (UCP-Apriori&UApriori) extended from Apriori Algorithm - Efficient by employing pruning method

C. Compressed and Loss Data Properties

Compression is a well-established research field. Frequently, data compression is more focused on static data or on certain data type [12]. The compression brings benefit by influencing the reliability and replication of the data [12]. However, the compression impacts query performance when lost its properties. There are some techniques that allow algorithms, involved during compression process [13] and helps in increasing efficiency of the yield. Unfortunately, it still occurs the loss of its properties [14], [15]. The compression presently has been implemented on certain data types and none involved in uncertain time series data.

A compressed property of the data is important in order to simplify the dataset for next data handling. Compression of data can cause loss of data properties during the process. The coding methods include arithmetic coding, adaptive transmission of the model, Markov modeling, Huffman coding and Lempel-Ziv codes [13]. It acts as a powerful method of data compression since linear treatment is inappropriate. The compression technique is also performed through signal decomposition, thresholding of wavelet transform coefficients, and signal reconstruction [16]. The compression causes significant reduction, and related to the cost of storing and transmitting of the data. Table III shows the comparison of previous compression method for certain data types.

TABLE III
COMPARISON OF PREVIOUS COMPRESSION METHOD FOR CERTAIN DATA TYPE

Method / Technique	Advantage	Disadvantage
Three-scale wavelet compression technique [16].	The coefficients are kept with temporal information through wavelet transform methods.	-Discarded result when it's not associated with disturbing events, -Lost information during decomposition, -The choice of threshold is important when involve in noise in the wavelet domain.
Refine Model Accuracy (RMA) technique [17], [18]	-Reduces storage requirement, -Does not require the discretization hierarchy, -Treats each dimension as a potential target measure -Support multiple aggregation functions without additional storage costs, -Touch a clustering method of handling data, -Much information about all parameters can be formed if the noise in the dataset is independent.	-Makes no assumption regarding the type of attributes or the density distribution, -Sacrifice in accuracy.

According to [17] past studies, the methods of data analysis and statistics are not losable but the increase in error bars where it shows to be minimal and superior with respect to an alternative compression system, Principal Component Analysis (PCA). However, this method cannot run with the existence of error in terms of lost data properties during compression. This situation is shown in Fig. 1, the effect of compression which is implemented on galaxy spectra dataset. Compression can enable a data warehouse to store several times more raw data without increasing the total disk storage or impacting query performance when involved in term of data platform [19]. However, in the data itself, compression involves the minimizing of data properties. Lossless compression cause the yield of poor reliability [12], [20].

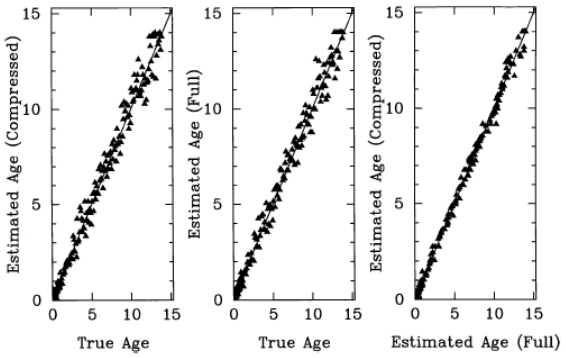

Fig. 1 The effect of the linear data compression algorithm. Here, the left-hand panel shows the recovered age from the two number y1 and y2 with age and normalization weightings, plotted against the true model age. The middle panel shows how well the full dataset can recover the parameters. The right-hand panel shows the estimated age from the y1 and y2 plotted against the age recovered from the full dataset with error and difficult to date accurately [17]

The compression algorithm brings impact on the overall design and test flow [13], [21] included compression codes, LFSR reseeding and Pseudo-random BIST techniques, which cause loss of information during the compression process. The differential compression method provides a computationally efficient compression technique and try hard at increasing the efficiency of prediction [14], [15], [22].

III. DISCUSSION

Knowledge from uncertain time series data brings important meaning for future prediction especially in weather. In a real situation, unpredictable events happen without being noticed by humans about it. In this paper, the methods have been reviewed that previously brought benefits to weather area through predicting the uncertain time series data. The analytical and experimental comparisons of techniques performed revealed that the experiment should be extended on this method in order to get the accuracy of predicting uncertain time series and improve the quality of yield in the area being focused. The methods and techniques involved in uncertain time series data are significant.

The previous experiment on compression of data properties has shown the yield of final outcome, and effect on a possibility of loss information and poor reliability. The compressed property of the data is important in order to simplify the dataset for next data handling. The frequent pattern mine algorithm has been implemented in order to mine the uncertain data type. The same technique can be implemented in uncertain time series data after the process of determining the type of dataset.

The compression techniques include three-scale wavelet, RMA and PCA that enable to compress the dataset but still bring disadvantages in term of loss of properties of data. Therefore, the combination in the implementation of algorithms and some techniques from previous research on compression data can help to reduce the loss of properties of the dataset in the uncertain time series type of dataset.

IV. CONCLUSION

This review paper produced an evaluation of methods in uncertain time series towards attaching domain. There are methods that have been used in the uncertain time series dataset but still brought limitation in predicting processes. Uncertain time series dataset is important in helping to build up the prediction. The comparison of analysis being done in order to determine the yield of predicting uncertain time series through the PSO algorithm, Monte Carlo simulation, and regression model in data mining method. The methods reviewed brought benefits to weather area through predicting. However, still future action must be taken in obtaining the accuracy of predicting uncertain time series and improve the

quality of yield in the focused area. On the other hands, frequent pattern mining algorithm has been implemented in order to mine the uncertain data even though it involves the loss properties during the compression process. The compression techniques included three-scale wavelet, RMA and PCA that enable to compress the dataset but still bring disadvantages in term of loss of properties of data. The combined study on some methods and techniques can be implemented and will help in order to fix this problem.

REFERENCES

[1] S. P. 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.

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

[3] M. Hooshsadat and O. R. Za, "An Associative Classifier For Uncertain Datasets," in in *Advances in Knowledge Discovery and Data Mining*, 2012, p. pp 342–353.

[4] 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.

[5] D. J. Gagne, A. McGovern, and M. Xue, "Machine learning enhancement of storm scale ensemble precipitation forecasts," in *Proceedings of the 2011 workshop on Knowledge discovery, modeling and simulation - KDMS '11*, 2011, p. 45.

[6] T. Fu, "A review on time series data mining," *Eng. Appl. Artif. Intell.*, vol. 24, no. 1, pp. 164–181, Feb. 2011.

[7] J. Aßfalg, H. Kriegel, P. Kr, and M. Renz, "Probabilistic Similarity Search for Uncertain Time Series," in *SSDBM 2009 Proceedings of the 21st International Conference on Scientific and Statistical Database Management*, 2009, pp. 435 – 443.

[8] S. R. Sarangi and K. Murthy, "DUST: A Generalized Notion of Similarity between Uncertain Time Series Similarity of Uncertain Time Series," *IBM India Res. Lab*, vol. 1, 2010.

[9] C. Qing, Z. Xiaoli, and Z. Kun, "Research on Precipitation Prediction Based on Time Series Model," in *2012 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring*, 2012, pp. 568–571.

[10] S. Esfandeh and M. Sedighizadeh, "Meteorological Data Study and Forecasting Using Particle Swarm Optimization Algorithm," *World Acad. Sci. Eng. Technol.*, vol. 59, pp. 2117–2119, 2011.

[11] Z. Nan, S. Wang, X. Liang, T. E. Adams, W. Teng, Y. Liang, and S. Member, "Analysis of Spatial Similarities Between NEXRAD and NLDAS Precipitation Data Products," *IEEE J. Sel. Top. Appl. EARTH Obs. Remote Sens.*, vol. 3, no. 3, pp. 371–385, 2010.

[12] C. M. Sadler and M. Martonosi, "Data compression algorithms for energy-constrained devices in delay tolerant networks," in *Proceedings of the 4th international conference on Embedded networked sensor systems - SenSys '06*, 2006, p. 265.

[13] K. C. Barr and K. Asanović, "Energy-aware lossless data compression," *ACM Trans. Comput. Syst.*, vol. 24, no. 3, pp. 250–291, Aug. 2006.

[14] A. et al., "System and Method for Differential Compression of Data from A Plurality of Binary Sources," *U.S.Patent*, vol. 2, no. 12, pp. 1–35, 2002.

[15] Levine, "Lossless Data Compression with Low Complexity," *United States Pat.*, pp. 1–25, 2000.

[16] S. Member and S. Member, "Power Quality Disturbance Data Compression using Wavelet Transform Methods," *IEEE Trans. Power Deliv.*, vol. 12, no. 3, pp. 1250–1257, 1997.

[17] A. F. Heavens, R. Jimenez, and O. Lahav, "Massive lossless data compression and multiple parameter estimation from galaxy spectra," *Mon. Not. R. Astron. Soc.*, vol. 317, no. 4, pp. 965–972, Oct. 2000.

[18] J. Shanmugasundaram, U. Fayyad, and P. S. Bradley, "Compressed data cubes for OLAP aggregate query approximation on continuous dimensions," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*, 1999, pp. 223–232.

[19] M. Poess, "Data Compression in Oracle," in *Proceedings of the 29th VLDB Conference, Berlin, Germany*, 2003.

[20] T. A. Welch, "A Technique for Hogh-Performance Data Compression," *IEEE Comput.*, pp. 8–20, 1984.

[21] E. H. Volkerink, A. Khoche, and S. Mitra, "Packet-based input test data compression techniques," in *Proceedings. International Test Conference*, 2002, pp. 154–163.

[22] B. Ryabko, J. Astola, and M. Malyutov, *Compression-Based Methods of Prediction and Statistical Analysis of Time Series: Theory and Applications*. 2010, pp. 1–109.