DOI 10.29042/2018-3957-3960

Helix Vol. 8(5): 3957- 3960

A Review on Time Series Dimensionality Reduction

*[1]Sagar S. Badhiye, [2]Dr. P. N. Chatur
[1]Scholar, [2] Head of Department, Government College of Engineering, Amravati
*Email: sagarbadhiye@gmail.com

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract
Time series is a sequential collection of values with respect to time obtained from various applications. The time series data have basic features like huge data size, high dimensionality with characteristics like trend, cyclical, seasonal, and irregular. The cumulative use of time series data has initiated a great deal of research and development attempts in the field of data mining. Dimensionality of time series is directly proportional to the efficiency of various data mining algorithms used for time series analysis. In this paper, a widespread review on the existing time series dimensionality reduction methods is given. The chief objective of this paper is to aid interested researchers to have a general idea about the current investigation in time series dimensionality reduction methods and identify their potential research direction to advance investigation in the same. The papers also discuss about the possibilities of using automata model for time series dimensionality reduction.

Keywords: Dimensionality, Time Series, Automata, Data Mining

Introduction

Large amount of data is being produced by various organizations in the world; the inventions of social networking websites are continuously adding to the data repositories all over the world. Most of this data is time dependent i.e. the time series data and hence has some features like cyclical, seasonal, trend, irregular. Storing this huge data needs lot of memory moreover analysis of this data for extracting these features from it consumes lot of time. Thus, to reduce the time required for extracting these information the given time series has to be represented in lower dimension thus storing only necessary values that depicts some reasonable information and helps in important data mining tasks like classification, clustering, forecasting, etc. by feature extraction, matching and computation of parametric values of the time series.

Existing methods used for dimensionality reduction in time series are sampling, Piecewise Aggregate Approximation (PAA), Dynamic Time Warping (DWT), Piecewise Linear Approximation (PLA), Piecewise Trend approximation (PTA), Piecewise Cloud Approximation (PWCA), Symbolic Aggregate Approximation (SAX), [1].

Even though by applying any of the above method on time series it is possible to extract required information form time series, still there is scope of improvement in the efficiency of time series mining tasks by designing new methods for time series dimensionality reduction as no exact method is best it depends on application domain.

Overview of Basic Terminologies

Time series data mining has attracted many researchers in last few decades. The review below is necessarily brief. Time Series Mining tasks that attracted researchers to work on it are as follows:

Prediction: Time Series are generally considered to be very lengthy and smooth, i.e. the sequential values are within the predictable ranges of one another [2]. Prediction aims at modelling a variable dependency that helps to forecast next few values of the time series. Thus, Given a Time Series $T = (t_1, t_2, \dots, t_n)$, Predict Next k values of Time Series t such that $t' = (t_{n+1}, t_{n+2}, \dots, t_{n+k})$ that are expected to occur.

Classification: Assigning an unlabelled time series Q to one of two or more predefined classes [3, 4].

Clustering: Distance formula is used to measure D (Q, C) so that the time series are grouped in database [5].

Anomaly Detection: A time series T, with some model of "normal" behaviour, find all sections of T which contain anomalies [6, 7].

Summarization: A time series T is consisting of n data points such that n is an extremely large number, and then a graphical approximation of T is created which retains its essential characteristics and is visible on computer screen, executive summary, etc [8].

Indexing: If some similarity/dissimilarity measure D (Q, C) is applied to time series Q to find the most similar time series in database DB [10, 11].

The time series data used for performing any of the above tasks consists of huge amount of data, the key features when dealing such huge time-series data is its representation. Because of the large number of samples in time series it is important to formulate a lower dimensional representation that retains basic properties of the time series. This representation scheme can be used with certain algorithmic procedure to perform any of the aforesaid time series mining

tasks; it also plays a vital role in determining the efficiency of these tasks.

Any time series representation methods should have the following properties as stated:

- Substantial reduction in number of samples in the data set.
- Fundamental visual characteristics of time series data should be maintained in the representation, and should support pattern matching in time series.
- Computation cost for representation should be low.
- Efficient restoration should be possible from the reduced representation.
- Noise in time series data should be handled properly.

Study of various representation methods that are already implemented by various researchers shows different trade-offs among the aforesaid properties [11]. Classified the representations into the following three categories as non-data adaptive, data adaptive and model based.

Non-data Adaptive: The transformation parameters are same for every time series regardless of its nature.

Data Adaptive: The transformation parameters are altered subject to the available data. Approximately all non-data adaptive methods can be implemented as data adaptive.

Model Based: A model is considered to be the source of time series. The intention is to find the parameters of such a model as representation.

Related Work

Discrete Fourier Transform [12] is based on the basic principle of spectral decomposition. The extended wavelet function was used [13]. The literature consists of numerous wavelets like Haar [14], Daubechies [15], coiflets [2].

The Piecewise aggregate approximation (PAA) method [16], represents a time series by computing average of sequential fixed length segments, Adaptive Piecewise Constant Approximation (APCA) [16] allowed variable length segments represented by average value of all data points and segment length.

A novel symbolic representation technique based on Piecewise Aggregate Approximation is Symbolic Aggregate Approximation (SAX), [18] that converts the time series into an alphabet of discrete symbols and allows dimensionality reduction. An extended approach of SAX [19], where in two additional points with maximum and minimum values of data points within the equal size segments is added to the symbolic representation. iSAX includes Relative Frequency and k-Nearest Neighbor (RFknn) algorithm[20], where the number of intervals is improved by remembering more distance with the closest distance measure to remember more information as compared to SAX.

Symbolic Essential Attribute Approximation (SEAA) is used to reduce the dimensionality of multi-dimensional time series [21 is based on the concept of data series envelopes and critical attributes produced by multilayer neural network. The representation based on upper and lower envelopes permits high dimensionality reduction in time series.

A method based on PIP detection i.e. inflection points detection in time series [31] which represents the movement shape of the time series more exactly than SAX, with fixed number of deflection points, this method can perform clustering of time series based on movement shape of time series.

Piecewise Trend Approximation (PTA) method for time series representation [22], allows dimensionality reduction by retaining the main trends in time series using ratio between two consecutive points, segmentation is performed only if the two Sequential segments have diverse trend and each segment is approximated by the fraction of first and last points inside the segment.

Piecewise Cloud Approximation (PWCA) is a cloud model based technique for dimensionality reduction [23], where each cloud reflects the distribution of data points within the “frame”.

Random sketches based dimensionality reduction of large datasets is used for distance based clustering and classification of reduced dataset [24], in random sketch, the sketch is formed by multiplying the data by a vector. Sketch vector forms the outline of the entire data.

Hinging Hyper planes function is used for time series segmentation [25] after which regression is used to maintain the continuity of the time series, least square support vector machine and l1-regularization is used to detect the segmentation points due to representation of continuous PWL function by Hinging Hyper planes.

Probabilistic Finite Automata is used to characterize and predict time series as part of machine learning field [26]. The probabilistic finite automata predict new values efficiently, and it is justified by examining it on global hourly radiation data and dry bulb temperature data.

A hybrid methodology based on Random Walk (RW) and Artificial Neural Network (ANN) which intelligently combines the benefits of RW and ANN models is implemented in [27], experimental analysis on the four real world financial time series shows that hybrid model improved the overall accuracy of forecasting.

A combination of short term forecasting of Multilayer Perceptron ensembles with the long term forecasting model of non-stationary time series shows improved

results when compared to the MLP ensemble alone [28].

Probabilistic arithmetic automata (PAA) [29] is a general model to describe chains of operations whose operands depend on chance, along with two algorithms to numerically compute the distribution of the results of such probabilistic calculations.

Dual finite automata based model [30] for deep packet inspection, in which packet payloads are matched against a large set of patterns in many networking applications, the dual finite automata consists of a Linear Finite Automata (LFA) and an extended deterministic finite automaton (EDFA). The dual automaton is evaluated in real world rule sets using different synthetic payload streams, results shows that dual FA increases efficiency in terms of number of states, and memory bandwidth.

Discussions

From the overall literature survey it is observed that there is still scope of research for representation of time series with dimensionality reduction. The various representation techniques used for dimensionality reduction are not able to represent the movements of shapes in time series efficiently. It is also observed that there is scope of increasing the efficiency of the various algorithms used for data mining task such as prediction, classification, regression etc by using this representation. It is also observed that use of automata can also be considered for time series representation that could eventually lead to better result of data mining tasks.

Proposed Methodology

Range Automata Model is proposed for dimensionality reduction in time series which converts the input numerical time series into alphabetical sequences in lower dimension.

Fig 1 shows the block diagram of the proposed method.

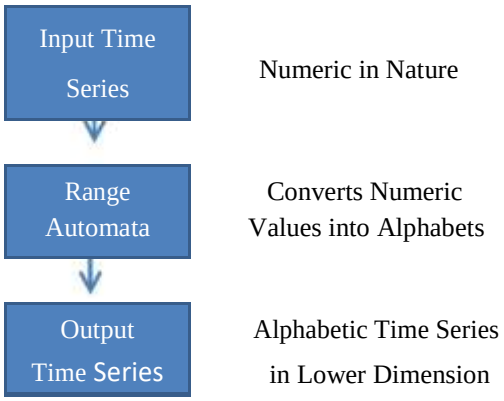

Fig. 1: Block Diagram

The output alphabetic time series is used for various time series data mining task like classification, clustering, prediction etc. The efficiency of the proposed method is determined by comparing with existing methods.

Future Work

The proposed method will be implemented and it will be tested on time series data set of temperature. Forecasting Model based on above automata model will be proposed for prediction of future values of time series.

References

1. Tak-chung – Fu. 2011. A review on time series data mining. Engineering Applications of Artificial Intelligence, Elsevier. 164-181.

2. Shasha, D., Zhu, Y. 2004. High Performance Discovery in Time Series: Techniques and Case Studies. Springer.

3. Geurts, P. 2001. Pattern Extraction for Time Series Classification. Proceedings of Principles of Data Mining and Knowledge Discovery, 5th European Conference, Freiburg, Germany. 115-127.

4. Keogh, E., Pazzani, M. 1998. An Enhanced Representation of Time Series which allows Fast and Accurate Classification, Clustering and Relevance Feedback. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, NY. 239-241.

5. Aach, J. and Church. 2001. G. Aligning gene expression time series with time warping algorithms, Bioinformatics. 495-508.

6. Keogh, E., Lonardi, S., Chiu, W. 2002. Finding Surprising Patterns in a Time Series database in Linear Time and Space. In Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada. 550-556.

7. Shashabi, C., Tian, X., Zhao, W. 2000. Time Series A Tree: A wavelet based approach to improve the efficiency of multi-level surprise and trend queries. In Proceedings of the 12th International Conference on Scientific and Statistical Database Management System, Berlin, Germany. 55-68.

8. Indyk, P., Koudas, N., Muthukrishnan, S. 2000. Identifying Representative Trends in Massive Time Series datasets using Sketches. In Proceedings of 26th International Conference on Very Large Databases, Cairo, Egypt. 363-372.

9. Chakrabarti, K., Keogh, E., Pazzani, M., Mehrotra, S. 2002. Locally Adaptive Dimensionality Reduction for Indexing large time series databases. ACM Transactions on Database Systems, Volume 27, Issue 2. 188-228.

10. Kahveci, T., Singh, A. Variable Length Queries for Time Series Data. In Proceedings of 17th

International Conference on Data Engineering, Heidelberg, Germany. 273-282.
11. Keogh, E., Lonardi, S., Ratanamahatana, C. 2004. Towards parameter free Data Mining. In Proceedings of 10th ACM International Conference on knowledge Discovery and Data Mining. 206-215
12. Agrawal, R., Faloutsos, C., Swami, A. 1993. Efficient Similarity Search in Sequence Databases. International Conference on Foundation of Data Organization (FODO)
13. Chan, K., Fu, A. W. 1999. Efficient Time Series Matching by Wavelets. Proceedings of 15th IEEE International Conference on Data Engineering, Sydney, Australia. 126-133.

14. Chan, F., Fu, A., Yu, C. 2003. Haar: Wavelets for Efficient Similarity Search of Time Series; With and Without Time Warping. IEEE Transaction on Knowledge Data Engineering, Volume 15, Issue 3. 686-705.
15. Popivanov, I., Miller, R. J. 2002. Similarity Search over Time Series Data using Wavelets. In Proceedings of the 18th International Conference on Data Engineering, San Jose, CA. 212-221.
16. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S. 2001. Dimensionality Reduction for fast similarity search in large Time Series Databases, Knowledge Information System, Volume 3, Issue 3. 263-286.
17. Keogh, E., Charkrabarti, K., Pazzani, M. 2001. Locally Adaptive Dimensionality Reduction for Indexing large Time Series Databases. In Proceedings of ACM Conference on Management of Data. 151-162.
18. Lin, J., Keogh, E., Lonardi, S., Chiv, B. 2003. A Symbolic Representation of Time Series with Implications for Streaming Algorithms. Workshop on Research issues in Data Mining and Knowledge Discovery, 8th ACM SIGMOD; San Diego, CA.
19. Battuguldur Lkhagva, Yu Suzuki and Kyoji Kawagoe. 2006. Extended SAX: Extension of Symbolic Aggregate Approximation for Financial Time Series Data Representation Data Engineering Workshop, IEEE IC.
20. Almahdi Mohammed Ahmed, Azuraliza Abu Bakar, Abdul Razak Hamdan. 2010. Improved SAX TS Data Representation based on Relative Frequency and k-Nearest Neighbor Algorithm. IEEE. 935-937.
21. Krawczak, M., and Szkatula, G. 2013. "An approach to Dimensionality reduction in TS", Information Science, Elsevier. 1-22.
22. Jingpei Dan, Weiren Shi, Fangyan Dong, and Kaoru Hirota, 2013. Piecewise Trend Approximation: A Ratio-Based Time Series Representation. Abstract and Applied Analysis, Hindawi Publishing Corporation.1-7.
23. Hailin Li, Chonghui Guo. 2011. Piecewise cloud approximation for time series mining. Knowledge-Based Systems, Elsevier. 492-500.
24. Parul Gupta, Swati Agnihotri, Suman Saha. 2013. Approximate Data Mining using Sketches for Massive Data. IC on Computational Intelligence: Modeling, Techniques and Applications, Procedia Technology, Elsevier. 781-787.
25.Xiaolin Huang, Marin Matijas and Johan A. K. Suykens. 2013. Hinging Hyperplanes for Time-Series Segmentation. IEEE Transaction on Neural Networks and Learning Systems, Volume 24 (8).
26. L. Mora-Lopez, J. Mora, R. Morales-Bueno, M. Sidrach-de-Cardono. 2005. Modeling Time Series of Climatic Parameters with Probabilistic Finite Automata. Environmental Modeling and Software, Elsevier. 753-760.
27.Ratnadip Adhikari, R. K. Agrawal. 2013. A combination of artificial neural network and random models for financial time series forcasting. Neural Computation and Applications, Springer.
28. Domingos Savio Pereira Salazar, Paulo Jorge Leitao Adeodato and Adrian Lucena Arnaud. 2014. Continuous Dynamical Combination of Short and Long-Term Forecasts for Nonstationary Time Series. IEEE Transactions on Neural Networks and Learning Systems, Volume 25 (1). 241-246.
29. Tobias Marshall, Inke Herms, Hans-Michael Kaltenbach and Sven Rahmann. 2012. Probabilistic Arithmetic Automata and Their Applications. IEEE/ACM Transaction on Computational Biology and Bioinformatics, Volume 9 (6). 1737-1750.
30. Cong Liu and Jie Wu. 2013. Fast Deep Packet Inspection with a Dual Finite Automata, IEEE Transactions on Computers, Volume 62 (2). 310-320.
31. Sang-Ho-Park, Ju Hong Lee, Se ok-Ju Chun, Jae-Won Song. Representation and Clustering of Time Series by means of Segmentation based on PIP detection IEEE Volume. 17-21.

