A REVIEW ON TIME SERIES REPRESENTATION

Kalyani S. Hatwar*

Prof. Sagar S. Badhiye*

Abstract-

Time series representation is one of key issues in time series data mining. Time series is simply a sequence of number collected at regular interval over a period of time and obtained from scientific and financial applications. The nature of time series data shows characteristics like large data size, high dimensional and necessity to update continuously. There are many techniques used for time series analysis and the problem of high dimension of time series data is reduced with the help of time series representation. With the help of suitable choice of representation it will address high dimensionality issues and improve the efficiency of time series data mining.

Keywords - Data mining, Time series representation, Time series data.

* Department of Computer Technology, Y.C.C.E., Nagpur, Maharashtra

I. INTRODUCTION

Time series is one of key issue in Data Mining [1]. It is an important class of data contains common kind of data which is generated sequentially in time, is simply a sequence of numbers collected at regular intervals over a period of time or Collection of observations indexed by the date of each observation. "A time series may be defined as a collection of reading belonging to different time periods of some economic or composite variables" by Ya-Lun-Chau . Mathematical representation of time series is, at some fixed interval h, at times τ_1, τ_2,…, τ_N may be denoted by $x(\tau_1)$, $x(\tau_2)$,…, $x(\tau_N)$.

Time series data is a sequence of observations collected from a process with equally spaced periods of time. Time series data [5] are like stored in financial, educational, medical and meteorological database and obtained from scientific and financial applications (e.g. electrocardiogram (ECG), weekly sales totals, and prices of mutual funds, stocks data and daily temperature). The nature [1] of time series data shows characteristics like large data size, high dimensionality and necessity to update continuously. Component of time series are i.e. the way of representation of time series are Secular trend, Seasonal variation, Cyclical variation, Irregular variation.

Data mining techniques such as clustering, classification, association rule, etc.[6] are applied on time series data to retrieve useful information and knowledge from these kinds of databases. There are various kinds of time series data related research like finding similar time series, dimensionality reduction, segmentation and subsequence searching in time series and many researchers are working on time series data analysis. Dimensionality affects destructively impacts on the result of time series data mining and a very costly querying process, so we need to overcome the problem of high dimensionality. The way of time series representation is used to reduce the dimension of the original time series data.

There are many method used to reduce high dimensionality of time series representations are sampling, Piecewise Aggregate Approximation (PAA)[2], Dynamic Time Warping (DWT)[3], Symbolic Aggregate Approximation (SAX)[4], Piecewise Linear Approximation (PLA)[11], Piecewise Trend approximation (PTA)[4], Piecewise Cloud Approximation (PWCA)[5]. Transforming the time series into any of the above representations it is possible to measure the similarity or distance between two time series in the reduced space.

Thus this many techniques used to address the issue of dimensionality reduction has been designed by various researchers and it is observed that there is still some scope[5] of increasing the efficiency of time series analysis by designing an efficient time series representation technique.

II. LITERATURE REVIEW

In this paper various dimension reduction techniques are studied, these techniques are used for the process of reducing the number of samples present in the time series. The need of addressing the problem of high dimensionality is because of its adverse effect on result of time series data mining. Query accuracy and efficiency are inversely proportional to the dimensions.

Tak-chung Fu have discussed various kinds of time series data related research. Time series is a collection of data stored in financial, educational, medical and meteorological database. Data mining techniques such as clustering, classification, association rule, etc. are applied on time series data to retrieve useful information and knowledge from it. Various traditional dimensionality reduction techniques are explained in this paper like sampling, piecewise aggregate approximation, piecewise linear approximation, symbolic aggregate approximation, discrete Fourier transform, etc. Dimension reduction can be done effectively by representing time series in various ways i.e. the number of data point of the original data is reduced [1].

The simplest method for representing time series is sampling (Astrom, 1969). In this representation method, a rate of m/n is used, where m is the length of a time series P and n is the dimension after dimensionality reduction. Drawback of this method is that distorting the shape of sampled time series, if the sampling rate is too low [1].

Another advanced method is to use the average (mean) value of each segment to represent the corresponding set of data points in time series. Piecewise aggregate approximation (PAA) in which segmented mean of starting and ending data points of each segment is consider. For example time series of n points and having p segments ($p > n$), then its PAA representation is n/p. Keogh et al.(2000a) investigate an extended version called adaptive piecewise constant approximation (APCA) in which length of each segment is not fixed but adaptive to the shape of the series. A major difference between PAA and APCA is that APCA can identify segment of variable length [10].

To reduce the dimension of time series data, another approach is to represent a time series with straight lines i.e. Piecewise Linear Approximation(PLA), in which the approximating line for the subsequence P(pi, …,pj) is simply the line connecting the data points pi and pj, i.e. the end point of consecutive segments, giving the piecewise approximation with connected lines. A piecewise linear function is a function composed of some number of linear segments defined over an equal number of intervals, generally of equal size. Advantage of this method is that reducing the dimension by preserving the salient points is a promising method. These points are called as perceptually important points (PIP). With the time series P, there are n data points: $P(p_1, \ldots, p_n)$. All the data points in P can be reordered by its importance by going through the PIP identification process. (Chung et al.(2001))[11].

Jingpei Dan, Weiren Shi, Fangyan Dong and Kaoru Hirota investigated the advanced method for representing time series i.e. Piecewise Trend Approximation (PTA). PTA represents time series in concise form while retaining main trends in original time series, thus the dimensionality of original data is reduced. PTA transforms original data space into the feature space of ratio between any two consecutive data points in original time series, where sign indicates changing direction while magnitude indicates degree of local trend. PTA algorithm consist of three main steps: local trend transform, segmentation and segment approximation [4].

DTW (Dynamic Time Warping) [3] is an effective way, used for finding similarity measure. Similarity measure is of fundamental importance for variety of time series analysis and data mining tasks. Similarly Euclidean Distance [2] is also used but it has a disadvantage that unable to handle shifting and scaling in time axis while computing similarity measure [6].

SAX is a symbolic representation of time series data. It is a first discretizing the time series into segments, then converting each segment into a symbol, i.e. an original time series of length n into a discrete symbolic string i.e. convert the result from PAA to symbol string. The whole process of SAX is completed in two phase. The first, piecewise aggregate approximation (PAA), is a process of high dimensionality reduction by considering the segmented mean for time series. In this process, time series is divided into w equal-length ‘‘frames’’, of which the length is k (k = n/w) and the values can be respectively represented by the mean of data points within the corresponding frame. The second process is symbolization. It is a transformation from mean values to a discrete symbolic string according to the equiprobably divided regions in PAA. An

important issue for symbolic representation is the evaluation of distance. SAX has two important advantages of Lower bound and dimensionality reduction [6].

III. PROPOSED PLAN

The Symbolic Piecewise Trend Approximation (SPTA) is proposed to increase the efficiency of time series representation.

A. Trend Representation

Firstly find the featured space of local trends mapped from the original data space with one dimension reduce with the help of ratios between each two consecutive data point in original time series, where sign indicates changing direction while magnitude indicates degree of local trend. After getting the trend representation of time series, discretize the obtain time series in equal samples and represent the segments in alphabets form. By trend representation will get [4]-

a. Knowledge of past behavior
b. Estimation
c. Study of other components

B. Symbolic Representation

For classification, clustering and indexing of time series, SAX is as good as or better than Fourier Transforms, wavelets, etc. Each repeated segment is represented by same alphabet symbol. Last step is to find the accuracy of the Symbolic Piecewise Trend Approximation [5].

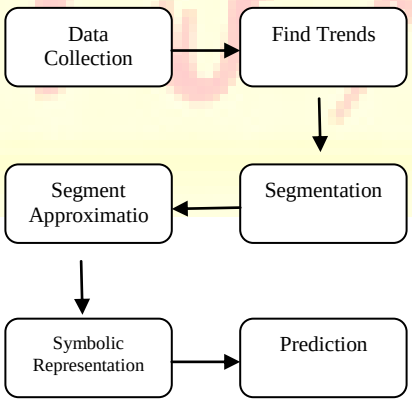

Fig. 1. Phases of implementation of proposed plan

IV. CONCLUSION

In literature survey section study of various techniques of time series representation is done. Different researches have focused on one or more problem in time series data mining, in order to improve the efficiency of time series data mining in high dimensional large size database. Several techniques of representation such as sampling, PPA, APCA, PLA, PTA, DWT, SAX have some advantages and disadvantages hence there is some scope for improving the efficiency of time series representation.

References

[1] Tak-chung Fu, "A review on time series data mining", Engneering Application of Artificial Intelligence: Elsevier Publication, Sept 2010.

[2] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," in SIGMOD Conference, 1994, pp. 419–429.

[3] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," in KDD Workshop, 1994, pp. 359–370.

[4] Jingpei Dan,1 Weiren Shi,2 Fangyan Dong,3 and Kaoru Hirota3, "Piecewise Trend Approximation: A Ratio-Based Time Series Representation", in Hindawi Publishing Corporation Abstract and Applied Analysis Volume 2013, Article ID 603629.

[5] Hailin Li, Chonghui Guo, "Piecewise cloud approximation for time series mining", Knowledge-Based Systems: Elsevier Publication, Dec 2010.

[6] Lei Sun, Yujiu Yang, Wenhuang Liu, "Trended DTW Based On Piecewise Linear Approximation for Time Series Mining", in 11th IEEE International Conference on Data Mining Workshops, 2011.

[7] Tao Sun, Hongfeng Sun, Weiheng Chen, "Dimensionality Reduction for Interval Time Series", IEEE World Congress on Information and Communication ,2012

[8] Peiman barnaghi, A. Abu Bakar, Z. Ali Othman "Enhanced symbolic Aggregate approximation method for financial time series data representation" Data Mining and Knowledge Discovery , 2007.

[9] Keogh, E., Chakrabarti, K., Mehrotra, S. Pazzani, M., "Locally adaptive dimensionality Reduction for indexing large time series databases", in Proceedings of the 2001 ACM SIGMOD International Conference on Manage- ment of Data, pp. 151–163,2001.

[10] Y. Ding, X. Yang, A. J. Kavs, and J. Li, "A novel piecewise linear segmentation for time series," in Proceedings of the 2nd International Conference on Computer and Automation Engineering, February, 2010.

[11] Chung, F. L., Fu, T.C., Luk, R., Ng, V., "Flexible time series pattern matching based on perceptually important points" , in International Joint Conference on Artificial Intelligence Workshop on Learning from Temporal and Spatial Data, pp. 1–7, 2001.