Trend based Approach for Time Series Representation

Sagar S. Badhiye
Department of CT, YCCE
Nagpur, India

Kalyani S. Hatwar
Department of CT, YCCE
Nagpur, India

P. N. Chatur, Ph.D
Department of CSE, GCOE
Amravati, India

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. With the help of suitable choice of representation it will address high dimensionality issues and improve the efficiency of time series data mining. Symbolic Piecewise Trend Approximation is proposed to improve efficiency of time series data mining in high dimensional large databases. SPTA represents time series in trends form and obtained its values. Sign of value indicate changing direction and magnitude indicates degree of local trend. Depending on the trend of time series, it is segmented into samples of different size which are approximated by the ratio between first and last points within the segment. Each segment then represented by alphabet. The time series is thus represented as sequence of alphabets thus reducing its dimension. Validate SPTA with naïve based classification method.

Keywords

Data mining, Time series representation, Time series, piecewise trends Approximation

1. INTRODUCTION

Time series is one of key issue in Data Mining [1]. It 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 $x(\tau_1)$, $x(\tau_2)$,…, $x(\tau_N)$.

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 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.

2. RELATED WORK

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(p1, …,pn) . 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.

3. PROPOSED METHODOLOGY

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

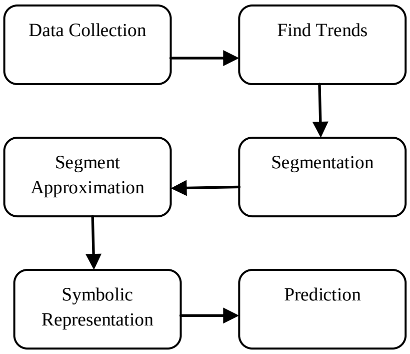


Fig 1: Phases of implementation of proposed plan

3.1 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] Knowledge of past behavior, Estimation, Study of other components.

Given a time series, X = {(x_1, t_1),…, (x_n, t_n)}, where x_i is a real numeric value and ti is the timestamp.

Stock market data is a time series data collecting information at each day. So, plot Stock market dataset in amplitude versus time axis [12].

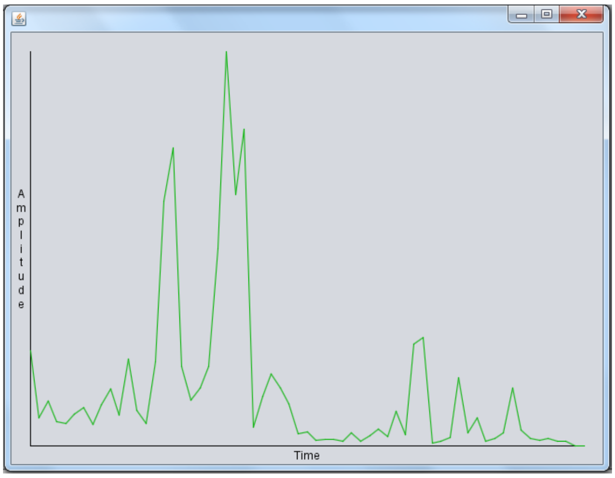


Fig 2: View Data

PTA [4] representation of X is,

$T' = \{(R_1,R_{t1})\ldots\ldots(R_m, R_{tm})\}$, $m \leq n$, $n \in N$,……(1)

Where R_{ti} is the right end point of the ith segment, Ri ($1 < i \leq m$) is the ratio between R_{i-1} and R_{ti} in the ith segment, and R_1 is the ratio between the first point ti and R_{ti} . The length of the ith segment can be calculated as $R_{ti} - R_{ti-1}$

3.1.1. Local Trend Transform

The original time series is transformed into a new series where the values of data points are ratios between any two consecutive data points in original series.

Given, Time series X = {(x_1, t_1),…, (x_n, t_n)}, $n \in N$,

New series T= {(r_1, t_2),……,(r_n, t_n)} is made from X by local trend transform, where r_i is the value of ratio between (x_{i-1}, t_{i-1}), i = 2,….,n. Ratios between each two consecutive data points in X are calculated according to the equation by justifying (1) as follows:

$$r_i = \frac{x_i - x_{i-1}}{x_{i-1}}, \quad i = 2,\ldots, n, \ldots\ldots(2)$$

T is indeed a feature space of local trends mapped from the original data space with one dimension reduced.

3.1.2. Segmentation

The transformed local trend series is divided into variable-length segments such that two conjunctive segments represent different trends, this segmentation aggregates the data points having the same changing directions.

X is segmented into $S = \{S_1, \ldots, S_m\}$ ($m \leq n$, $m \in N$), where $S_k(k = 1, \ldots, m)$

3.1.3. Segment Approximation

Each segment is represented by ratio between first and last points within segments.

3.2 Alphabetic Representation

3.2.1 Time Domain

Data compression method make data partition in the time domain. SAX implementation, time domain is divided into regular intervals and in each interval the average of amplitude value is calculated.

3.2.2 Amplitude Domain

Then partition the amplitude axis into intervals and this each interval denote with the suitable 'alphabets'. Amplitude interval may be regular or irregular, this amplitude range calculates with the help of probability of distribution of amplitude i.e. with function cumulative distribution function (CDF). CDF calculates the breakpoints for amplitude axis [9].

Each segment formed from the trends of the time series are represented with the help of alphabets.

4. OUTPUT

Proposed Symbolic Piecewise Trend Approximation methods for time series representation is more efficient and obtain best results than the exiting time series representation techniques.

In fig. 3 obtained trends from the original time series that lead to obtain new time series data with one dimension reduce.

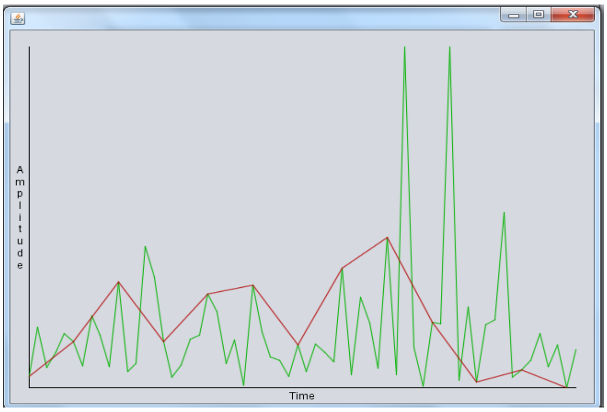
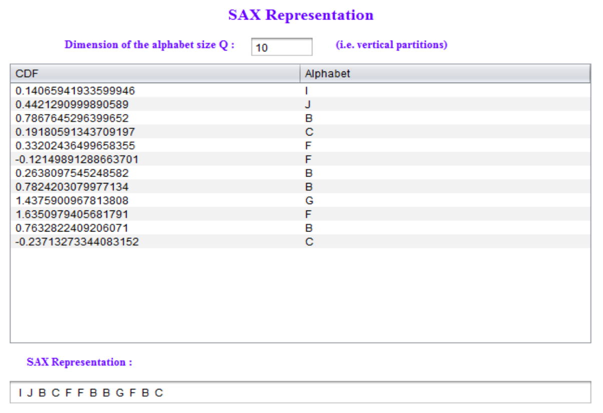


Fig 3: Trend Representation

In fig. 4 obtained trend of the time series is represented by alphabetic symbols. First divide an amplitude axis in alphabets then calculate a Cumulative distribution Function (CDF) of trends and then assign a symbol to trends.



SAX Representation

Dimension of the alphabet size Q : 10 (i.e. vertical partitions)

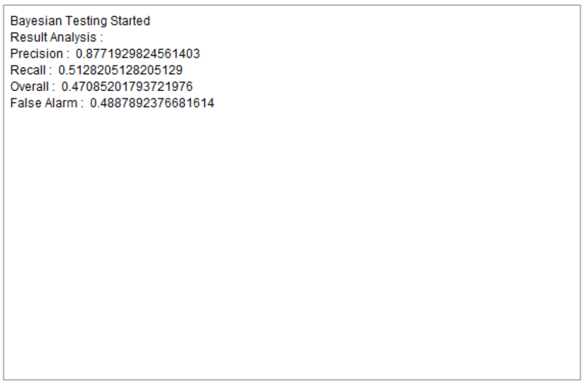
CDF	Alphabet
0.14065941933599946	I
0.4421290999890589	J
0.7867645296399652	B
0.19180591343709197	C
0.33202436499658355	F
-0.12149891288663701	F
0.2638097545248582	B
0.7824203079977134	B
1.4375900967813808	G
1.6350979405681791	F
0.7632822409206071	B
-0.23713273344083152	C

SAX Representation :

I J B C F F B B G F B C

Fig 4: Alphabetic Representation

In fig. 5 shows the analysis of the proposed work or validate the performance of the SPTA representation with the help of classification method. Naive base classification method used for the prediction of the time series analysis.



```
Bayesian Testing Started
Result Analysis :
Precision :  0.8771929824561403
Recall :  0.5128205128205129
Overall :  0.47085201793721976
False Alarm :  0.4887892376681614
```

Fig 5: Analysis

5. FUTURE SCOPE

In future, work on multidimensional parameter of time series dataset, which useful for analysis of the high dimensional data size of time series.

6. CONCLUSION

In order to improve the efficiency of time series data mining, Symbolic Piecewise Trend Approximation (SPTA) proposed to represent the time series alphabetically. There are several techniques of representation such as sampling, PPA, APCA, PLA, PTA, DWT, and SAX have some advantages and disadvantages hence there is some scope for improving the efficiency of time series representation. Trend represents every peak values of the dataset which contains the information like sign i.e. directions and magnitude i.e. local trend values of dataset. Then Segments of trend are represents as the sequence of alphabets where a single alphabet represents a single segment thus reducing dimension of the time series dataset. By applying naive based classification algorithm calculate the precision of the stock market dataset. Hence, efficiency of time series data mining is enhanced by applying SPTA representation.

7. 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] A.Notaristefano, G. Chicco, F. piglione “Data size reduction with symbolic aggregate” in IET Generation, Transmission and Distribution, July 2012.

[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.

[12] http://www.nseindia.com/live_market/dynaContent/live_watch/equities_stock_watch.htm?cat=N