International Journal For Technological Research In Engineering
Volume 3, Issue 7, March-2016 *ISSN (Online): 2347 - 4718*

SURVEY ON LARGE SCALE TIME SERIES DATA ALGORITHMS

Vikas Kumar[1], G C Sathish[2]
[1]M.Tech (AIT) Student, [2]Senior Associate Professor
C & IT, Reva University, Bangalore, India

Abstract: The Large Scale Time Series Data Mining process is crucial because it can't fit in the traditional mining algorithm. Each Data Mining algorithm is pertained with the approximate and exact approach. Approximate approach has limited degree of the accuracy whereas exact approach has high computational time. A middle approach can be followed between these two exact & approximate approaches and algorithm should me made based on this approach. First, it will follow the approximate algorithm and if it's pattern result is sufficient for interpretation then end this approach here only but if interpretation result is not sufficient then we need to go for the exact approach where outliers and missing value can be applied at that point where approximate approach ends as agile steps. This Hybrid Algorithm will reduce the computational time and give the high degree of accuracy for Large Scale Time Series Data.
Keywords: Exact, approximate, accuracy, computation time,

I. INTRODUCTION

There is a great requirement of the Data Mining Algorithm everywhere to analyze and predict the future of any probable situation in all area and it will be very important when it's about the Medical Data. Generally the Medical data are delicate in nature and most available in the form of Large Scale Time Series data. [1] These Large Scale Time Series data can't be analyzed with the available Data Mining algorithm. There is no question mark or any obsolete condition on the accuracy and efficiency about many Data Mining algorithms, but each & every available algorithms are work under some certain and specified condition and it can be understood from their processing nature. If we talk about the Large Scale Time Series Data which can't be stored, cleansed, aligned and analyzed efficiently through available algorithm because of the two reasons; first is Storage and second is Computing speed. [2] The size of the Large Scale Time Series Data can't be stored properly before pre-processing; It is required to divide into smaller chunks and then pass through the processing environment, but divided chunks may lost the originality during the segmentation. Computing speed is depend upon the environment and the Size of the aligned data but these pre-processed aligned data can be analyzed with the help of certain defined condition and predefined algorithms. [3] Large Scale Time Series Data have re-occurence motif which will generally repeat after each certain interval and then single motif is enough to predict the complete behavior so no need to worry about loosening the originality and algorithm's work to just detect at which certain interval of time Data is re-occuring. All Data Mining Algorithms are based on the exact and approximate methods. The algorithm, which is based on the exact approach it will take more computation time to process because in search of exactness and accuracy, it has to process all data parts with various different situational aspect. The algorithms which are based on the approximate approach it will take less computation time rather than exact approach. By observing both the approach, need to compromise if algorithms based on exact approach it has compromise with time and when algorithms is based approximate then it has to compromise with the accuracy. To solve these problems several researchers had proposed several algorithms. The major goal of this survey paper is to focus on existing Data Mining algorithms, which is not suitable for processing the Large Scale Time Series Data. It requires a Hybrid algorithm which will work for both exact and approximate methods and with the taste of agility.

II. LARGE SCALE TIME SERIES DATA

From name, Large Scale Time Series Data could be understood that which would be large in size. Large Scale Time Series Data doesn't bound with the time limitation and it has repeated sequence. It is very difficult to store, process & analyze. [4] Generally, Large Scale Time Series Data will be pertained with medical data, weather statistic information, and web browsing history. The ECG (Electrocardiogram) is the example of the Large Scale Time Series Data in Medical domain. The ECG data will give the brief information about the electrical changes in our heart system. These data are repeating after certain interval of time and each interval of time most probably data will be same or with very small ignorant values. These data are very delicate in nature because it interprets several meanings. Finding the specific intervals where same sequence will be going to repeat and then analyze the sequence by sequence cycle data and if meaning is not clear then analyze from the second cycle but correlated sequence or cycle can be ignored. [5] The Correlated sequence should be ignored because it will give the same meaning and increase the processing time as well resources consumption. The effective interpretation with exact or approximate approach is required but time is the major objectives. Less time with the effective interpretation can be achieved only through Hybrid Data Mining Algorithm. Hybrid Data Mining Algorithm is the best option for the Large Scale Time Series Data where a middle solution can be drawn in between exact and accurate approach.

III. EXISTING DATA MINING ALGORITHMS

Many techniques had been proposed for Large Scale Time

Series Data for interpreting & extracting the pattern. All are examined with benefits and drawbacks. The major drawback of all the algorithms are, all of them made to work and analyze in predefined situations and in other environment, it will give less accuracy in their result patterns.

A. Anonymization Algorithm:
It is important algorithm because it will have the sensing capacity which will work out for the large volume of data. In this, data can be sliced and later sent for the examination but because of it's annoymity behaviour it will slice the data from any where so that the originality may surpassed. Continuity will carry the originality but if continuity will be lost then quality of the data get compromised. It will take less computation time but quality degraded data can't predict the better accurate pattern or result.

B. Fuzzy Algorithm:
This algorithm is best suited for statistical computation for the large volume set of data with maintaining the degree of originality because it will consider numerical with several decimal value to compute and achieve the best degree for transformation of the data set. If it will consider several decimal value then it will occupy much more space and latter it will take much more time to process the data. Though it's accuracy is very high but for real time scenario where instant result is required then this algorithm doesn't stand for the Large Scale Time Series Data.

C. Perturbation Algorithms:
This algorithm preferably chosen by the distributive computing for large set of data, because multiplicative projection matrices will be formed after the transformation from the target data and it will process these data very easily. This method has one major drawback for the missing data, while splitting the data few data will get lost and it will affect on the accuracy for the pattern recognition. If the accuracy could be compromised then it would not work well for the Large Scale Time Series Data.

D. Zero R Algorithm:
This Algorithm works on voting strength or can say it is based on democracy. It depends upon the target values but ignores the predictors. If it ignores the predictors then definitely the computation time will be more but the accuracy will be in average. Here, one dimensional formed matrix which is the outcome of the transformed data, the accuracy will be high but for all other situations we need to compromise with the timing. So, this algorithm approach is not suitable for Large Scale Time Series Data.

E. One R Algorithm:
It is based on one rule and it will consider all the target data and predictors. It process such that all the target will examined with each available predictor one by one and made a frequency table for interpreting the result. It's best part is, it will give very low total error for the complete process. So, It's accuracy is very high but computation time will be more than other and then it couldn't be ideal for Large Scale Time Series Data.

IV. PROPOSED SYSTEM

Generally, Data mining process will work; the data set given for the selection and it will become the target data, target data will be preprocessed and preprocessed data will be transformed and transformed data will be given for the data mining. Here, preferably data set will be chosen for exact approach in data mining because the degree of accuracy will be high but time consumption is much more compare than others. So, this paper express such that first we will start with approximate approach and if pattern result will interpret sufficient details about that data then we will stop this process but if it won't give the sufficient pattern result then from that step onwards itself as agile step process, now data will follow the exact approach where it accommodates the outliers and missing value. If this data will be processed then definitely it will give the high degree accuracy pattern result. So, it can be observed that in Approximate, first it will save the time and if predicted result is not up to as standard then only it will go for the exact approach and this process will not again start from beginning it will start right from where approximate approach ends.

V. CONCLUSION

This hybrid approach which follows the process between the approximate approach and exact approach will work efficiently for the Large Scale Time Series Data. Using this approach, if algorithm made to work then it need not to compromise with either degree of accuracy or computation time. Large Scale Time Series Data will work efficiently only when time bound is not limited and computation process will have less time. So, It can be said this Hybrid Algorithm is suitable for this Large Scale Time Series Data.

REFERENCES

[1] E. Keogh et al. (2011). The UCR Time Series Classification/Clustering Homepage [Online]. Available: www.cs.ucr.edu/~eamonn/time_series_ Data.
[2] B. Chiu, E. Keogh, and S. Lonardi, “Probabilistic discovery of time series motifs,” in Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2003, pp. 493–498.
[3] X. Luo et al., “An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems,” IEEE Trans. Ind. Informat., vol. 10, no. 2, pp. 1273–1284, May 2014.
[4] F. Shifeng et al., “An integrated system for regional environmental monitoring and management based on Internet of Things,” IEEE Trans. Ind. Informat., vol. 10, no. 2, pp. 1596–1605, May 2014.
[5] D. Wegener et al., “Toolkit-based high-performance data mining of large data on mapreduce clusters,” in Proc. Int. Conf. Data Mining Workshops (ICDMW’09), 2009, pp. 296–30.

