2011 International Conference on Asia Agriculture and Animal
IPCBEE vol.13(2011) © (2011)IACSIT Press, Singapore

Adaptive Outlier Detection in Streaming Time Series

Durga Toshniwal [1], Shachi Yadav [2]

[1 & 2] Electronics and Computer Science Department, Indian Institute of Technology Roorkee

Abstract. Most existing outlier detection techniques work on static time sequences and process outliers by working on the entire time sequences to detect global outliers. This would be computationally infeasible in case of data streams as they are ordered sequences of data that arrives continuously. In this work, we aim to develop an adaptive algorithm that detects outliers from streaming time series data. The emphasis is on detecting the local outliers in addition to the global outliers. The HOT SAX algorithm has been extended to detect the local outliers. The algorithm is adaptive as it detects outliers on the basis of a reference set of outliers which keeps on getting updated dynamically depending upon the current data streams instances. The introduction of "outlierness" reflects the extent of abnormal behavior and the "type" refers to the type of deviation from normal behavior i.e. above normal or below normal. The proposed work has been evaluated on real life case data- daily vehicular traffic dataset.

Keywords: Streaming Time Series, HOT SAX, Outlier Detection

1. Introduction

In data mining outlier detection is a type of data analysis technique that seeks to determine and report such data objects which are grossly different from or inconsistent with the remaining set of data. The technique is used for data cleansing, spotting emerging trends and recognizing unusually good or bad performers. Data streams have received considerable attention in current times [1]. The outlier analysis in data streams may lead to detection and prevention of fraudulent activities, criminal activities, failures and improvement in security and safety in diverse applications e.g. in credit card industry. It is a challenging work due to the characteristics of data streams. Most existing outlier detection techniques work on static time sequences. They process outliers by working on the entire time sequences to detect global outliers. This would be computationally infeasible in case of data streams [3]. Therefore the objective of this work is to develop an algorithm that detects outlier subsequences from streaming time series data. The HOT SAX algorithm [2] has been extended to detect the local outlier subsequences in the time series streams. The notion of "outlierness" has also been introduced which is used to capture the extent of abnormal behaviour shown by the outliers. Further a "type" of outlier has been defined to denote the deviation of outliers above or below the normal behaviour.

The rest of the paper is organized as follows. In Section 2, we review related work and discuss some background material related to the proposed algorithm. In Section 3, we discuss the proposed algorithm for detecting the adaptive outliers in time series stream data. Section 4 presents the description of dataset used, results of the experiments performed and its analysis. Finally, Section 5 offers some conclusions and suggestions for future work.

2. Related Work

In this section, a brief summary of some significant, related research works has been presented.

[1] Corresponding Author: durgatoshniwal@gmail.com. This work has been supported by the funding granted by IBM SUR Award
[2] E-mail address: shachi.yadav@gmail.com

2.1. Streaming Time Series Classification

Classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends. Due to the characteristics of the stream data, the traditional classification methods are inappropriate for them. The most distinguishing characteristic of data streams is that they are time-varying that is only the current state is stored, as opposed to traditional database systems. This change in the nature of the data takes the form of changes in the target classification model over time and is referred to as concept drift. Therefore the classification model should be such that it can adapt itself with the dynamics of the stream data [3].

2.2. Outlier Detection Techniques

Statistics based approaches are mainly developed for detecting outliers in static data. They assume that the dataset follows a statistical model, e.g. a normal or Poisson distribution. In case of data streams, prior knowledge about the data distribution may not be known. Cluster based approaches, such as CLARANS [4], DBSCAN [4], etc have been used for outlier detection in diverse datasets. They detect outliers as by-products by finding the number of clusters. The main problem of this approach is that they merge the detection of outliers with the detection of clusters and therefore do not focus entirely on outlier detection. Density-Based Approaches adopt a Local Outlier Factor (LOF) for outlier detection [5].

3. Proposed Work

In this section we present the entire framework of the proposed algorithm. The following is a broad summary of the proposed technique:

- The time series stream data is being generated by some source, say by monitoring of vehicular traffic on a highway, is captured and stored in a data buffer.
- In each local segment, collected in data buffer from stream data, the local outlier is detected. The HOT SAX algorithm has been extended to find the local outliers in local streams. The details of detection of local outliers are discussed in Section 3.1.
- The local outliers detected are stored in a vector of predetermined capacity. The vector is used for generating reference set. The reference set is used for generating the outlier distribution. The reference set is discussed in detail in Section 3.2.
- The outlier distribution, details of which are discussed in Section 3.3, is generated by measuring the dispersion of local outliers received in the reference set. The measures of dispersion are used to develop conditions for rule formulation which are used for classification of outliers.
- The rule-based classification model, the details of which are discussed in Section 3.4, is used to classify the outliers into local or global classes. The algorithm is adaptive as it detects outliers on the basis of a reference set of outliers which keeps on getting updated dynamically to accommodate newly evolved outliers depending on the current data stream instances.
- Further, the conditions developed are used for identifying the degree of "outlierness" and the "type" of these outliers. The "outlierness" reflects the extent of abnormal behaviour and the "type" refers to the type of deviation from normal behaviour i.e. above normal or below normal. The detailed discussion is given in Section 3.4.

3.1. Local Outlier Detection

The HOT SAX algorithm [2] has been extended to find outlier subsequences in the time series stream data. It is used to find the local outliers in local segments of stream data, collected in data buffer. It requires only one parameter that is the length of the outlier subsequence. It simply takes each possible subsequence, in a given local segment, and finds the distance to the nearest non-self match [2]. This distance is known as non-similar distance [2]. The subsequence that has the greatest value of non-similar distance is the outlier. This is achieved with nested loops: where the outer loop considers each possible candidate subsequence and the inner loop is a linear scan to identify the candidate's nearest non-self match. HOT SAX algorithm creates two data structures to support the nested loop heuristics. The two heuristics are used for optimization of nested loop computation.

3.2. Reference Set Generation

Since we are dealing with streaming time series data we don't have any prior data knowledge so as to build a classifier for describing a predetermined set of data classes or concept for learning, as used in traditional classification method [3]. Therefore, in the proposed algorithm we generate a reference set that can be used for learning the nature of stream data that is being used. The local outliers detected from local segments are stored in the reference set. It is a vector of predetermined capacity, following the sliding window concept. It is used for generation of outlier distribution.

3.3. Outlier Distribution Generation

The outlier distribution is generated for developing conditions which are used for formulating rules in classification model. It is generated by measuring the dispersion of local outliers received in the reference set. The most common measures of data dispersion are quartiles, interquartile range, standard deviation, etc [3]. In the proposed algorithm we calculate the first quartile, the third quartile and the inter quartile range to measure the dispersion of local outliers stored the in reference set. Every time the reference set is updated the value of these three measures of dispersion is also re-evaluated. If the value of the new interquartile range is same as the previous value of the interquartile range then the values of the conditions of the rule used for classification is not re-evaluated.

3.4. The Rule-Based Classifier Model

Two conditions are checked for classifying an unseen outlier as global. First one is the identification of repeated outliers. For identifying repeated outliers we compare the location of the newly detected local outlier in the data buffer, with the location of the previously detected outlier. Since the data buffer follows the concept of sliding window, the two sequences at locations p and p-1 at time t and t+1 are same. Therefore, if the location of a newly detected outlier is different from the location of the previously detected outlier, we consider the newly detected outlier as distinct else repeat. The next condition is avoiding trivial local outliers. To determine the interesting outliers from the rest of the local outliers detected so far a common rule of thumb is used for identifying suspected outliers in a given data distribution. It singles out values falling at least 1.5 * IQR above the third quartile or below the first quartile [3].

The proposed algorithm also establishes the "type" of the global outlier. If the non- similar distance of the unseen outlier is below the value of (1.5 * IQR – first quartile) then it is considered as "below normal" type of outlier. Otherwise, if the non-similar distance is above the (1.5 * IQR + third quartile) it is considered as "above normal" type. The degree of "outlierness" is also identified for both types of outliers. The length of reference set is used as the set of degrees that will be mapped to outliers. We first sort the two vectors containing "above normal" and "below normal" outliers in the increasing order. The largest value of non-similar distance is considered severe in "above normal" type and the as mild degree. In the "below normal" type the severe degree is mapped to the smallest value of the non similar distance, and rest are as mild.

4. Results and Discussions

The proposed work has been applied to real world case data - St. Gotthard tunnel dataset. The St. Gotthard Tunnel is in Switzerland and is the third longest road tunnel in the world. This tunnel forms part of the shortest road link from Hamburg, Germany to Sicily, Italy. The dataset consists of daily transportation data. It measures the frequency of motorcycles passing in one direction through Gotthard tunnel each day in a week. It contains 365 records from Jan 2005 to Dec 2005 [6]. The data buffer size is taken 30. The outlier subsequence length is taken 8. The reference set size is also taken as 30. The Fig. 1 shows the general traffic condition in Gotthard tunnel all the year round including all types of vehicles one particular direction during the year 2002 -03 [6] and this data has been taken as the case data in the present study. Table 1 presents the analysis. The evidence of the facts presented in Table. 1 is given in [7]. Fig. 2 shows the outliers and the Table 2 gives a description of the outliers detected.

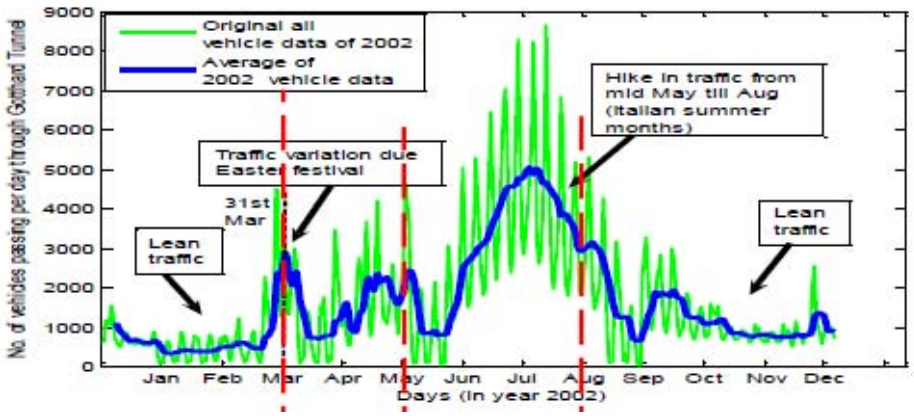

Fig. 1: The Gotthard tunnel traffic of all types of vehicles in both directions in year 2002 and its analysis.

Table 1: The explanation of the two graphs of Gotthard tunnel traffic in year 2002 showing the general traffic all the year round.

Section (Fig.2) explained	**Traffic condition in year 2002**
Section 1 (Jan to Feb)	Lean traffic
Section 2 (Mar to Apr)	Traffic is high in last week of March due to Easter falling on 31st of March'02
Section 3 (May to Aug)	Hike in traffic
Section 4(Sept to Dec)	Lean traffic

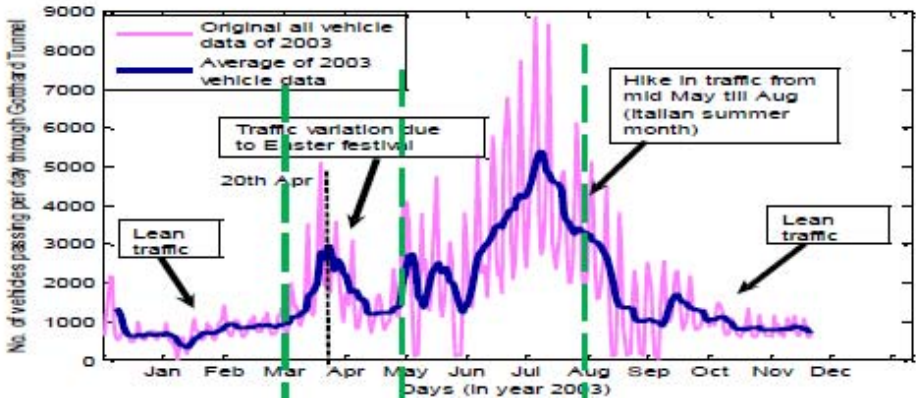

Fig. 2: The Gotthard tunnel traffic (03) of all types of vehicles in both directions and its outlier analysis.

Table 2: The outliers visualized in Gotthard tunnel dataset with the number of motorcycles indicating traffic in one particular direction.

Section of Fig. 2 explained	Traffic condition in year 2003
Section 1 (Jan to Feb)	Lean traffic
Section 2 (Mar to Apr)	*Outlier visualized*: Traffic is high in last week of March due to Easter falling on 20th of April'03
Section 3 (May to Aug)	*Outlier visualized:* Hike in traffic
Section 4(Sept to Dec)	Lean traffic

4.1. Results with Gotthard Tunnel Dataset: Number of Motorcycles in One Direction (in Year 2005)

In this section, we will present some of the results obtained. The complete results cannot be shown due to space constraints

Outlier detected: May 3-14, 2005 UEFA European Under-17 Football Championship in Italy. The Fig. 3 shows the outliers and Table. 3 presents the analysis of the detected outliers, and the evidence is as per [8].

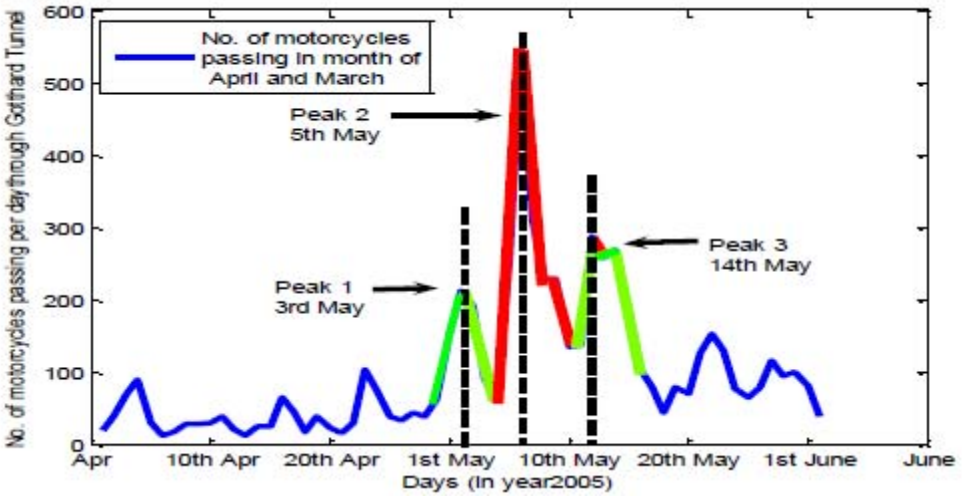

Fig. 3: Outliers detected in month of May for motorcycle passing in one direction through Gotthard tunnel.

Table 3: Explanation for the three peaks on 3rd, 5th, and 14th of May 2005 (Refer to [8]).

S.No.	Name	Analysis
1	Peak 1	The UEFA under -17 football championship started on 3rd May. Group matches were scheduled, between: Belarus & England, Italy & Turkey, Israel & Switzerland, Croatia & Netherlands
2	Peak 2	On 5th May, the group matches were between - Italy & Belarus, Turkey & England, Switzerland & Netherland, Israel & Croatia.
3	Peak 3	Final was on 14th May, between Netherland & Turkey.

5. Conclusion and Future Work

In this work, we considered the problem of detecting adaptive outliers in streaming time series data. The HOT SAX algorithm has been extended successfully to detect outliers in the streaming time series data. An outlier has been classified successfully into global or local by adaptive rule-based classifier. The type of outlier has also been established A degree of "outlierness" has been identified. In future, the proposed algorithm can be extended to detect the outlier from multidimensional or categorical data. Various other existing outlier detection techniques can be extended, just like HOT SAX algorithm, for detecting local outliers.

6. References

[1] Feng Han, Yan-Ming Wang, Hua-Peng Wang, ODABK: An Effective Approach to Detecting Outlier in Data Stream. In: 5th International Conference on Machine Learning and Cybernetics, Dalian (2006)

[2] Eamonn Keogh, Jessica Lin, Ada Fu, HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequences. In: 5th IEEE International Conference on Data Mining, pp 226-233 (2005)

[3] Jiawei Han, Micheline Kamber: Data Mining Concepts and Techniques. Elsevier (2008)

[4] Kozue Ishida,Hiroyuki Kitagawa.: Detecting Current Outliers: Continuous Outlier Detection over Time-Series Data Streams, vol. 5181,pp. 255-268, Springer-Verlag Berlin Heidelberg (2008)

[5] Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying Density- Based Local Outliers. In:

SIGMOD, pp. 93–104 (2000)

[6] http://www.kdd.org/

[7] http:// en.all.experts.com/q/Switzerland-157/St-Gotthard-Road-Tunnel

[8] http://www.wikipedia.com 11.