An Efficient Outlier Detection Using Amalgamation of Clustering and Attribute-Entropy Based Approach

Jasmine Natchial .F\textsuperscript{a},* Elango Parasuraman\textsuperscript{b} and Taslina. B\textsuperscript{c}

\textsuperscript{a}PhD. Scholar, Bharathiar University, Coimbatore, India. \\
\textsuperscript{b}Assistant Professor, Department of IT, PKIET, Karaikal, India. \\
\textsuperscript{c}M.E Scholar, Dept of CSE, BCET, Karaikal, India.

Abstract

Many organizations today have more than very large databases that grow without limit at a rate of several million records per day. Mining these continuous data streams brings unique opportunities but also new challenges. Thus data stream has become a dynamic research area of Data mining. Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. The data stream is motivated by emerging applications like Consumer click streams and Telephone records, Bulky set of web pages, Multimedia data and so on. Outlier detection in streaming data is a very challenging problem because of the reality that data streams cannot be scanned multiple times. The outliers may exert undue influence on the results of statistical analysis. So they should be identified using reliable detection methods prior to performing data analysis. The main objective of this proposed research work is to detect outliers using amalgamation technique where one or more techniques are combined for efficient anomaly detection.

Keywords: Outlier, Amalgamation, Cluster-based, Attribute-Similarity based.

2010 MSC: 34G20.

1 Introduction

A Data Stream is an enormous sequence of data elements continuously generated at a fast rate. In data streams, huge amount of data continuously inserted and queried, such data has very large database. It is motivated by emerging applications with massive data sets. In recent years, we have observed that enormous research activity motivated by the explosion of data collected and transferred in the format of data streams. So data stream has gained importance and has now become an extensively studied field of research area.

An outlier is an object or a set of data in an observation or a point that is found to be dissimilar or inconsistent or do not comply with the general behavior or model of the remaining data. Depending upon different application domains these abnormal patterns are often referred to as outliers, anomalies, discordant, observations, faults, exceptions, defects, aberrations, errors, noise, damage, surprise, novelty, peculiarities or impurity. Attempts to eliminate them altogether result in the loss of important hidden information as one persons noise could be another persons signal. In many cases outliers are more interesting than normal cases for example network intrusion detection, credit card fraud detection, weather prediction, detecting outlying cases in medical data, marketing and customer segmentation, etc.,

Outliers may be erroneous or real in the following sense. Real outliers are observations where actual values are very different than those observed for the rest of the data and violate plausible relationships among variables. Erroneous outliers are observations that are distorted due to misreporting errors in the

*Corresponding author.
data collection process. Outliers are produced by following three causes: (i) Data caused by inherent changes, (ii) Erroneous data as a result of execute error such as manual entry errors, hackers brake, and equipment failure, (iii) Data that fall into wrong classes.

2 Related Study

Over the years, a large number of techniques have been developed for building models for outlier and anomaly detection. However, the real world data sets, data streams present a range of difficulties that bound the effectiveness of the techniques. The Classical technologies of outlier detection can be categorized as following: Statistic-based methods [5], Depth-based [16], Distance-based methods [6, 7], Clustering-based methods [10], Deviation-based method [11, 12], Density-based [3, 8, 9] methods and dissimilarity-based [13] or Similarity-based [4] methods.

The statistical outlier detection techniques are essentially model-based techniques and are suited to quantitative real valued data sets or ordinal data distributions. A data instance is declared as an outlier if the probability of the data instance to be generated by this model is very low. They are based on statistical estimates of unknown distribution parameters [14, 15] and here lies their limitation.

In the definition of depth-based, data objects are organized in convex hull layers in the data space according to peeling depth, and outliers are expected with shallow depth values. As the dimensionality increases, the data points are spread through a larger volume and become less dense. This makes the convex hull harder to discern and is known as the “Curse of Dimensionality”. The distance-based methods rely on the measure of full dimensional distance between a point and its nearest neighbor in the data set. The careful selection of suitable parameters for is the major drawback of this method.

Cluster analysis is popular unsupervised techniques to group similar data instances into clusters. This involves a clustering step which partitions the data into groups of similar objects. Here the outlier either belong to very small clusters or do not belong to any cluster or are forced to belong to a cluster where they are very different from other members. A major limitation of this approach is that they require multiple passes to process the data set. The deviation-based methods identify outliers by examining the main characteristic of objects in a group instead of by applying statistical tests or distance based measurement. Objects that deviate from the given descriptions are considered outliers. This method has perfect performance but the hypothesis of exception is too idealization.

Density-based detection estimate density distribution of a data point within data set and compares the density around a point with the density around its local neighbor. The relative density of a point compared to its neighbors is computed as an outlier score and points which are having a low density is considered as an outlier.

In dissimilarity-based method, the outlier is detected on the basis of the percentage with which the data objects vary from the other or objects which are very dissimilar to the other data objects in some data set.

In Similarity-based approach the outlier detection focuses on finding the similarity coefficient and object deviation degree. Above classical methods have respective advantages in application, but they all have some limitation in certain aspects. To solve this problem, recently, amalgamation of techniques for outlier detection is proposed and has gained more attention in recent years. These hybrid approaches combine or merge two or more techniques for efficient anomaly detection. This paper suggests amalgamation of clustering based method with the similarity-based approach.

The rest of this paper is organized as follows. Section 3 explains the amalgamation procedure followed by proposed algorithm in Section 4 and Conclusion in Section 5.

3 Formal Definition

This paper proposes a two phase method to detect outliers. (A). The first phase groups the data into clusters using the Euclidean distance and (B). The Second phase constructs similarity matrix and finds the Object Deviation Degree. All the attributes are taken into account in the Object Deviation Degree, the larger the value, the greater the possibility of the object being outlier, and vice versa.
3.1 Clustering Algorithm

A prototype based, simple partition clustering technique called K-Means clustering is used here. This algorithm attempts to find a user specified k number of clusters represented by their centroids. A cluster centroid is typically the mean of points in the cluster.

There are two stages in this algorithm:

**First stage:** Selection of k centres randomly where k is fixed in advance.

**Second stage:** Assignment of data objects to the nearest centre. Euclidean distance is used to determine the distance between each data object and the cluster centres.

When all the data objects are included in some clusters, recalculation is done on the average of the clusters. This iterative process continues repeatedly until the criterion function becomes minimum.

3.2 Similarity - Matrix Computation Algorithm

Given the input data set belonging to a particular cluster, first Attribute similarity Coefficient, Attribute Entropy can be calculated. Next step is to construction of Attribute Entropy Matrix. Finally with the help of this computed matrix, the object entropy can be easily calculated in turn the Maximal Attribute Entropy and the Object Deviation Degree can be deduced. The larger the deviation degree is, the greater the possibility of the object being an outlier.

4 Proposed Algorithm

Input: Data Set \( D = (U, A) \), \( U \) stands for the object set, \( U = u_i | i \in L, L = 1, 2, \ldots, m \), \( A \) stands for the attribute set, \( A = a_j | j \in S, S = 1, 2, \ldots, n \).

Cluster centre \( C = c_1, c_2 \ldots c_k \), where \( c_i \) =cluster centre, \( k \) = no. of clusters.

Output: The outlier.

Step 1: Compute the distance of each data points \( u_m \) and the cluster centres \( c_k \).

Step 2: Assign each data object \( u_i \) to the cluster with closest centroid \( c_i \).

Step 3: For each data object \( u_i \) compute its distance from the centroid \( c_i \) of the nearest cluster.

Step 4: If the calculated distance is less than or equal to previous calculated distance then the data objects stay in that cluster itself or else calculate the distance to each of the new cluster centres and assign the data object to the nearest cluster.

Step 5: Repeat the above two steps until no new centroids are found or a convergence criteria is met.

Step 7: Consider each cluster as a separate Data set and compute the similarity coefficient of each data point of the data set.

Step 8: Compute the attribute entropy and construct the attribute entropy matrix from which the object entropy can be deduced.

Step 9: Compute Object deviation degree and compare with pre-set threshold.

Step 10: Larger the object deviation degree, greater the possibility of the data point to be an outlier.
Thus from each cluster, outlier can be efficiently ruled out.

5 Conclusion

The research is pursued taking into study the different existing conditions found in practice which is represented. The key objective of the proposed work is to merge the clustering based outlier detection method for the streaming data with attribute entropy based approach. The method applied both the clustering method and attribute entropy method for detection of group and individual outliers. The proposed method needs to be implemented on varying datasets and is found that this proposed method is well suited for Outlier Detection in Data Stream Mining. Future work requires approach applicable for categorical and mixed data sets. In order to achieve accurate results and faster computation, it might have overlooked or ignored some superior points which could have made the model more sophisticated but it would have led to computational difficulty.

References


Received: August 18, 2015; Accepted: September 10, 2015

UNIVERSITY PRESS

Website: http://www.malayajournal.org/