Hybridizing Clustering and Dissimilarity Based Approach for Outlier Detection in Data Streams

Jasmine Natchial F.*, Elango P.** and Taslina B.***

Abstract: The main aim of Data Stream Outlier Detection is to find the abnormalities effectively and accurately. In practice, outlier detection widely finds its application in telecom fraud detection, credit card fraud detection, network intrusion detection, clinical diagnosis, etc. Due to the need for real-time detection, dynamic adjustment and inapplicability of existing algorithms, there arise a need for a new trend and algorithm. So this paper aims to couple the effectiveness of two different algorithms namely the k-means clustering and outlier mining using Dissimilarity approach.

Key Words: Outlier, Hybridization, Cluster-based, Dissimilarity based.

1. INTRODUCTION

Data mining is a system that reveals the hidden and valuable information from the data and the facts. Outlier mining belongs to the dynamic research area of Data Mining. A data stream is a large series of data elements generated constantly at a faster rate. Data stream has gained importance and has now become an extensively studied field of research area. A unique feature of data stream is that they are time-varying and cannot be scanned multiple times. An outlier is an object or a set of data in an observation or a point that is considerably dissimilar or inconsistent or do not comply with the general behavior or model of the remaining data. 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. According to Hawkins “An outlier is an observation that deviates so much from other observations as to arouse suspicious that it was generated by a different mechanism.” [22]. 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 person’s noise could be another person’s signal. In many cases outliers are more interesting than normal cases.

2. LITERATURE REVIEW

In recent years, many outlier detection algorithms came into existence. However, the real world data sets are concerned; data streams present a range of difficulties that bound the effectiveness of the techniques. Knorr proposed FindAllOutsD based on distance in 1998 [16]. It failed to work effectively when there exist data with various different distributed densities in the data set. Markus in 2000[17] proposed LOF based on outlier degree. But became less effective due to the fact that the computational complexity of the algorithm is high. In 2002, Yu D [18] proposed Find Out algorithm to process high-dimensional data. But the space complexity of the algorithm is high and constantly changing.
Based on LOF, LOCI was proposed in the year 2003 [19]. It guaranteed accuracy but not suitable for the circumstance of the data stream. In 2007 Anguilli [20] proposed exact-STORM and approx-STORM based on distance and index with R-Tree to improve the query efficiency. Yang [21] proposed dynamic grid clustering to implement fast outlier detection in data stream. Presently, the classical technologies of outlier mining can be divided into following categories: statistic-based methods [7], depth-based [15], distance-based methods [8, 6], clustering-based methods [11], deviation-based method [12, 13], density-based [14, 9, 10] methods and dissimilarity-based [4] or similarity-based [3] methods.

In the definition of depth-based, data objects are organized in convex hull layers in the data space, 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. 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 objects that deviate from the given descriptions. 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 points which are having a low density is considered as an outlier.

In dissimilarity-based method, the outlier detection focus on finding the data 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 dissimilarity-based approach.

The rest of this paper is organized as follows. Section III explains the amalgamation process followed by proposed algorithm in Section IV and Conclusion in Section V.

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 segregates clusters as Cluster-Inliers and Cluster-Outlier and B. Constructs dissimilarity matrix and finds the Dissimilarity Degree to drive off the hidden outliers in the cluster-inliers. This Dissimilarity degree reflects the degree of deviation of the data point. The smaller the deviation degree, the greater the possibility of the object or the data point being an anomaly, 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 which are typically the mean of points in the cluster.

There are three 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.

**Third stage:** Segregation of clusters to Cluster-Inliers and Cluster-Outlier based on Average Weight Center (ANC). The segregated Clusters-Inliers are then subjected to the next phase of the algorithm where the hidden outliers are effectively removed using the following algorithm.

### 3.2. Dissimilarity-Matrix Computation Algorithm:

Given the data set belonging to a particular cluster-inlier, the Attribute dissimilarity (ad) can be calculated from which the Object dissimilarity (od) is derived followed by the construction of Dissimilarity matrix (dm). With the help of this matrix, the Synergic Dissimilarity (sd) and the Maximum dissimilarity can be deduced. Then the Dissimilarity degree is found. The smaller the deviation degree is, the greater the possibility of the object being an outlier.

### 4. PROPOSED ALGORITHM

Hybridizing Clustering and Dissimilarity Based Approach (H-CADBA)

1. **First Phase:** (Clustering using k-means)
   
   **Input:** Data set $D = \{d_1, d_2, \ldots, d_n\}$, $n =$ no. of data points. Cluster centre $C = \{c_1, c_2, \ldots, c_k\}$ where $c_i =$ cluster centre and $k =$ no. of clusters. $T =$ the total no. of data points and $t_n =$ total no. of data points in the cluster $c_n$.
   
   **Output:** Cluster-inliers and Cluster-outlier.

   **Step 1:** Compute the distance of each data points $d_n$ and the $k$ cluster centres $c_k$.
   
   **Step 2:** Assign each data object $d_i$ to the cluster with closest centroid $c_i$.
   
   **Step 3:** For each data object $d_i$ 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 6:** Find the Average no. of points in ‘$k$’ clusters, $ANC = \frac{\sum_{n=1}^{k} (t_n \times C_n)}{T}$.
   
   **Step 7:** Segregate clusters as clusters-inliers and clusters-outlier depending on the ANC. Regard the cluster that is not closely related to any neighborhood of data points and which have less than half of ANC as cluster-outlier and remove them. The outliers in the large clusters or cluster-inliers are then detected using the following procedure at the second phase.

2. **Second Phase:** (Dissimilarity based approach)

   **Input:** Data Set $D = (U, A)$, $U$ stands for the object set with $m$ as cardinality, $U = \{u_i \mid i \in L\}$, $L = \{1, 2, \ldots, m\}$, $A$ stands for the attribute set with $n$ as cardinality, $A = \{a_j \mid j \in S\}$, $S = \{1, 2, \ldots, n\}$.
   
   Cluster centre $C = \{c_1, c_2, \ldots, c_k\}$, where $c_i =$ cluster centre, $k =$ no. of clusters.
   
   **Output:** The outlier.

   **Step 8:** Consider each cluster-inliers as a separate Data set and compute the similarity coefficient of each data point of the data set.
Step 9: Compute the Attribute Dissimilarity $AD$. 
\[ AD_{ij} = \| x_{ij} - x_j \| \cdot \| x_{ij} - x_j \| / (x_{\text{max}} - x_{\text{min}}) \]
where $x_{ij}$ is the value of attribute $f$ of $u_i$, $x_j$ is the average value of attribute $f$, $x_{\text{max}}$ is the maximal value of attribute $f$, $x_{\text{min}}$ is the minimal value of attribute $f$.

Step 10: Object Dissimilarity (OD) is calculated using the equation
\[ OD(i, j) = \sum_{k=0}^{n} AD(ak_{ij}) \text{ln} \text{ where } n \text{ is the cardinality of } A. \]

Step 11: Construct Dissimilarity Matrix $dm$.

Step 12: Find the Synergetic Dissimilarity (SD), using the equation,
\[ SD = \sum_{j=1}^{m} OD(i, j) \text{ where } m \text{ is the cardinality of } U. \]

Step 13: Compute maximum dissimilarity, $d_{\text{max}} = \max_{i=1}^{m} SD_i$

Step 14: Dissimilarity Degree, $dd_i = d_{\text{max}} - SD_i / d_{\text{max}}$
Smaller the degree of deviation, greater the possibility of the data point to be an outlier. Thus from each cluster, outlier can be efficiently ruled out.

5. CONCLUSION
There are many methods available for outlier detection in data streams, but their efficiency depends on the nature of data and their distribution in a specific domain. Methods work efficiently if combination of methods is employed. In this paper, the method applied both the clustering method and dissimilarity-based approach for detection of group and individual outliers. The proposed method needs to be implemented on varying datasets. Future work requires approach applicable for categorical and mixed data sets.

References


