Cluster in High Dimensional Data to Detect Outlier

Sonali. A. Patil[1]**, Snehal. S. Thokale**[2]

[1]Professor, Department of Computer Engineering, JSPM's BSIOTR, Pune University, Nagar Road, Wagholi Pune, India

[2] Department of Computer Engineering, JSPM's BSIOTR, Pune University, Nagar Road, Wagholi Pune, India

Abstract: ***In computer world data should be secured. To find deviated data or fraud in data some techniques are introduced and that method is nothing but Outlier Detection., it helps to improve find out frauds and intruders. Outlier detection techniques are used for batch system also but for huge data we have to more careful to find such type of data. So, to overcome this over sampling principle component analysis is used. By using Principal Component Analysis (PCA), it is helpful to find Outlier. In this system ouraim to detect the presence of Outlier from a large amount of data using an online updating method. As we are using Oversampling Principal Component analysis (osPCA), there is no need to store data matrix or covariance matrix each time and thus approach is to find anomalies in online data stream or large amount of data problems. In our proposed system we capture only UDP packets will include. Along with this we are proposing algorithm for clustering***

Keywords: Detection of Outlier, online updating, oversampling, principal components analysis, TCP And UDP packets, DBSCAN cluster algorithm

1. Introduction

Outlier detection is technique is used to find out a deviated data. A well-known definitions of "Outlier" is given in [1]:"an observations which deflect so much from other inspection as to find uncertainties that it was generated by a different mechanism, which gives the general scheme of an influenced data instance and motivates many Outlier detection can be found in applications such as homeland safety, credit card detection in cyber-security, fault detection, or malignant diagnosis. Outlier detection needs to solve an unsupervised and unbalanced data learning problem. When any outlier instance is added or removed to its space principal direction comes into variation, but does not affect when normal data instance added or removed.

So online updating technique for over sampling principal component analysis that is oversampling principal component analysis (osPCA). Along with this we are proposing DBSCAN cluster algorithm for dividing data into small DBSCAN so that accuracy of Outlier detection will increase. In existing system processing of detection is done on TCP packets as we know TCP is connection oriented protocol, to overcome this in proposed system we are using UDP protocol to find Outlier from UDP packets. Now, let's one quick look at TCP and UDP protocols, TCP is one of the important protocols in TCP/IP networks. TCP enables two hosts to create a connection and exchange data. As TCP creates first connection and then exchange data, it will take time for establishing connection between two hosts.UDP (User Datagram Protocol) is an another option for communication protocol to Transmission Control Protocol (TCP) used mainly for set up low-latency and loss tolerating connections between applications on the Internet. Both protocols send small packets of data, called datagrams. In our framework we are using UDP packets for data extracting least square data instance. So our framework will ready to find out outliers in wireless system application. As we are working with high dimensional data we are using cluster algorithm to increase robustness of detection technique. In this system TCP and UDP packets are considered and through cluster we are these packets in different clusters.

2. Liturature Survey

This section presents related work done by the researchers Detection process for Outlier.And also different methods to detect infected data instance that is Outlier or outliers.

2.1 Outlier detection for high dimension data

In this paper [5], new technique introduced to detect outlier which find the influenced data instance by studying on the dataset in which projection of data points are observed. The sparsity of high dimensional data suggests that each point is a similarly good Outlier from the perspective of closeness based definitions. The idea of finding meaningful influenced data instance becomes substantially more difficult and non-observable.

2.2 Angle-based outlier detection in high-dimensional data

To find outlier angle based method uses angle parameter. That means it finds angle between data instances which are considered as target points. If considered targeted data instance is outlier then it shows small angle variation. And if data which is not infected it will show large angle variation. This method has calculation overhead, as each time it need to consider two data instances to find angle and also it requires memory to store data instances, memory cost increases. Sothat limitation comes to solve large-scale problems, as the user will need to keep all data instances to calculate the required angle information, so memory cost increases [3].

2.3 Incremental Local Outlier Detection for Data Streams

In this paper author used LOF algorithm for detecting influenced data points from distributed data stream. incremental LOF algorithm is computationally capable, while at the same time very successful in detecting influenced data instance and changes of distributional pattern in various data stream applications[7].

2.4 Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases

To search arbitrary shape of clusters in large data dimensional .DBSCAN algorithm is used. It requires minimal number of parameter as input. In this algorithm it discovers spatial data cluster. Density Based Spatial Clustering of Applications with noise is designed to discover the spatial data clusters with noise[9].

3. Principal Component Analysis For Detection of Outlier

This section presents concept of finding infected data by using Principal Component Analysis (PCA) method and oversampling principal component analysis.

3.1 Technique of Principal Component Analysis

There are two types of detections methods used supervised and unsupervised. Unsupervised detection technique is used for high dimensional data taken into consideration.PCA is mathematical concept which isused to calculate direction of data flow in unsupervised data method. To find PCA it requires maintaining covariance matrix and eigenvector. Depending on values of vector PCA determines direction of data flow. Eigenvectors keep information among vectors and it gives principal direction. Dominant vector is calculated, if any data is influenced then that vector shows variation in direction. If any data instance is added or removed in data space then PCA shows variation in direction, but it happens if online updating of data is going on, at that time online Outlier detection method is used.

3.2 Detection of Outlier using Oversampling PCA

In general PCA method it needs to calculate n PCA so computational time. And if data instances are inserted or removed it is difficult to observe change in direction of flow. So to overcome this oversampling is introduced, in this method target data instances are oversampled, that means it creates many copies of target data instances. If that objective instance is infected then it will show large difference between principal directions of data flow.

In oversampling method most dominant vector is calculated. Then it calculates score of outlierness of newly added data instances by comparing it with previously determined threshold. Because of this it needs less memory.

For online detection technique oversampling principal component analysis (osPCA) is used in this there are two stages. In first stage calculates PCA by using osPCA technique, this calculation done offline. In second stage online detection is used to detect infected data. Infected data identified by using comparison of previously determined threshold and score of outlierness of that newly added data instance. As this technique does not require store whole data matrix so memory cost is less and also calculation time is less.

4. Implementation Details

This section gives system overview in detail, architecture of proposed system,

4.1 System Overview

The following figure 1 shows the architectural of the proposed system. Following description gives step by step description of architecture.

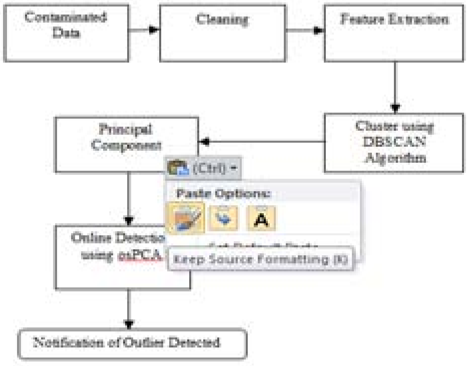
Contaminated Data
Cleaning
Feature Extraction
Cluster using DBSCAN Algorithm
Principal Component
(Ctrl)
Paste Options:
A
Online Detectio using osPCA
Keep Source Formatting (K)
Notification of Outlier Detected

Figure 1: System Architecture

1) **Cleaning**
In this phase we give raw data as input to system, then we clean this data and we consider some data for processing further, and then it will be over sampled using OsPCA. It will calculate then score of influence data instance and set lower value as threshold.

2) **Pattern Extraction**
In this step we extract the pattern form incoming data. For this, system use Least Square Data pattern which will select at runtime by user. Then this selected data pattern will consider as Transaction list. This list is input data for OsPCA calculation. In this first we oversample that data and then calculate PCA by which we detect outlier in next module. As we are working with high dimensional data, it is hard to find infected data instance in large amount of data, so that we are dividing data into different parts for achieving accuracy to find infected data instances. There are four types of attacks are used in existing system and depending on that attack, outliers are detected. Along with proposed algorithm DBSCAN which is used to find DBSCAN [8].

3) **Cluster**
By assuming data is selected for detection of Outlier. So it happens that outliers may assume as normal data in detection technique. So to overcome this problem clusters are created for input dataset. By using DBSCAN algorithm [8] data is divided into multiple clusters. DBSCAN algorithm is basically used for high dimensional data. And detection is performed on threshold.

4) **Principal Component Analysis**
In this phase principal component is calculated on oversampled data instances.

5) **Detection of Outliers**
In this phase Outlier are detected using online detection technique, threshold is used to determine Outlier of received data instances. If threshold value is less than received data instance then, it will be considered as outlier.

4.2 DBSCAN Cluster

Density based algorithm is based on classification grouping which is rely density based notion of cluster, as this system is working with high dimensional data space so it is required to make group of data into different cluster based on classification. DBSCAN algorithm discover cluster of arbitrary shape and also have ability handle noise. It finds meaningful subclass[8].

5. Result and Discussion

As our proposed system used to find outlier, following table showing result fordifferent types of attack over TCP and UDP protocol. Following table 1 shows the time require to osPCA for TCP and osPCA for UDP with the numbers of test of each attack type.

Table 1: Time Comparison

Attack	Online OSPCA for TCP in sec	Online osPCA for UDP in sec
DOS	2.698	2.231
Probe	2.697	2.203
R2L	2.695	2.240
All Attacks	2.734	2.210

6. Conclusion and Future Scope

In this paper, we proposed DBSCAN clustering algorithm with an online detection of Outlier method based on oversample PCA. When oversampling a data instance, proposed method osPCA identify outlier data instances even in large amount of data and also it is more robust for detection as we are using cluster, and take less time for connection because of detection is done on TCP as well as UDP.

7. Acknowledgment

I would like to thanks my guide Prof. S. A. Patil Mam and all professors in our department for their constant support. We are also thankful to the reviewer for their valuable suggestions.

References

[1] D.M. Hawkins, Identification of Outliers. Chapman and Hall, 1980.
[2] V. Chandola, A. Banerjee, and V. Kumar, “Outlier Detection: A Survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 15:1-15:58, 2009.
[3] H.-P. Kriegel, M. Schubert, and A. Zimek, “Angle-Based OutlierDetection in High-Dimensional Data,” Proc. 14th ACM SIGKDDInt’l Conf. Knowledge Discovery and data Mining, 2008.
[4] Y.-R. Yeh, Z.-Y. Lee and Y.-J.Lee, “Outlier Detection viaOversampling Principal Component Analysis,” Proc.First KESInt’l Symp. Intelligent Decision Technologies, pp. 449-458, 2009,
[5] Charu C. Aggarwal, Philip S. Yu,” Outlier Detection for High Dimensional Data”, Proc. ACM SIGMOD Int’l Conf. Managementof Data, 2001.
[6] F. Angiulli, S. Basta, and C. Pizzuti, “Distance-Based Detection and Prediction of Outliers,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 2, pp. 145-160, 2006.
[7] D. Pokrajac, A. Lazarevic, and L. Latecki, “Incremental Local Outlier Detection for Data Streams,” Proc. IEEE Symp. ComputationalIntelligence and Data Mining, 2007.
[8] https://en.wikipedia.org/wiki/DBSCAN
[9] Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases

Author Profile

Prof. Sonali A. Patil is Assistant Professor at Department of Computer Engineering at JSPM’s BSIOTR Wagholi College Pune University, pursuing Phd from BSAU, Chennai. Intrested Domain includes cloud computing, Data Minimg, Software Engineering, Network Security and Grid Computing

Ms. Snehal S. Thokale is M.E. Student in Department of Computer Engineering at JSPM’s BSIOTR Wagholi College Pune University,