ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Principal Component Analysis to Detect Anomaly in High Dimensional Data using Cluster

Snehal Thokale[1], Sonali A Patil[2]

P.G. Student, Department of Computer Engineering, JSPM's BSIOTR, Wagholi Pune, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, JSPM's BSIOTR, Wagholi Pune, Maharashtra, India[2]

ABSTRACT: In Data analysis large amount of records or variables are processed. There are different types of anomaly detection techniques. These techniques are used for different application such as credit card fraud, voting irregularity analysis etc. To secure data detection methods are used. Anomaly is nothing but outliers, it helps to improve find out frauds and intruders. Anomaly detection techniques are used for batch system also but for huge data are will work. So, to overcome this over sampling principle component analysis is used. By using Principal Component Analysis (PCA), it is helpful to find anomaly. In this system ouraim to detect the presence of anomaly 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 approachis to find anomalies in online data stream or large amount of data problems. In our proposed system we capture UDP and TCP packets . Along with this we are proposing algorithm for clustering.

KEYWORDS: Detection of anomaly, online updating, oversampling, principal components analysis, TCP And UDP packets, Subspace cluster algorithm.

I. INTRODUCTION

Anomaly detection technique is used to identify deviated data instances. 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 anomaly detection can be found in applications such as homeland safety, credit card detection in cyber-security, fault detection, or malignant diagnosis. However ,since only a small amount of considered data are available in the above real world applications, how to determine influenced data point of hidden data (or events) draws attention from the researchers in data mining and machine learning communities [2][3].In spite of the small number of the deflected data, its presence might have an effect on the solution model such as the distribution or directions of the data flow. For example, the calculation of the data mean or the least squares solution of the connected linear regression model is both susceptible to infected data instance. As a result, anomaly detection needs to solve an unsupervised and unbalanced data learning problem. When any infected data 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 subspace cluster algorithm for dividing data into small subspace so that accuracy of anomaly 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 anomaly 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 data grams. In our framework we are using UDP as well as TCP packets for data extracting least square data instance. As we are working with high dimensional data we are using cluster algorithm to increase robustness of detection technique.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

II.LITURATURE SURVEY

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

II.I Distance-based detection Algorithm

A distance-based outlier detection technique uses subset of data to find infected data instances. It finds top influenced data instance for that first it creates group of data instances, which are not labeled. That group is called as solving set for outlier detection. It consider distance between instances in subset or group, and depending on that instance it finds infected data instances which are influenced. If all data space taken to solve it creates complexity, to overcome this it uses method to calculate false positive rate of new data instances. It requires quadratic time to solve this set. It uses ROC analysis to get accuracy[11].

II.II Outlier detection for high dimension data

In this paper [10], new technique for anomaly discovery is introduced which find the influenced data instance by studying on the behaviour of projections from the data set. However, in large space, the data is insufficient and the opinion of nearness fails to recollects its meaningfulness. Truth is told, the sparsity of high dimensional data suggests that each point is a similarly good anomaly from the perspective of closeness based definitions. The idea of finding meaningful influenced data instance becomes substantially more difficult and non observable.

II.III Angle-based outlier detection in high-dimensional data

To find influence data instances angle based method uses angle parameter. That means it finds angle between objective instances. If considered objective data instance is influence then it shows small angle variation. And if data which is not infected it will show large angle variation. This method has disadvantage of 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. So that 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.

II.IV 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 [12].

III. PRINCIPAL COMPONENT ANALYSIS FOR DETECTION OF ANOMALY

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

III.I Technique of Principal Component Analysis

There are two types of detections methods are used supervised
and unsupervised. When high dimensional data taken into consideration unsupervised detection technique is used.PCA is used to calculate direction of data flow in unsupervised data method. To find PCA it requires to maintain 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.

PCA calculates data matrix and presence of infected data instance are susceptible for data matrix.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 anomaly detection method is used. Some time exception instances are far away from its data space but still there with its principal direction, at this time oversampling method is useful.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

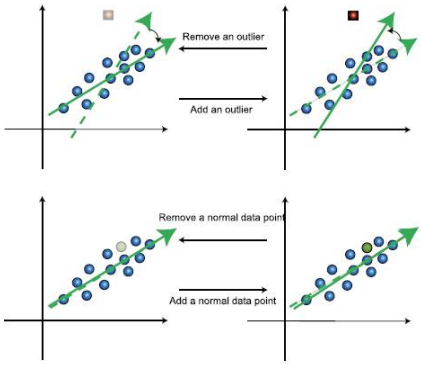

Fig. 1. Variation in Principal Direction while adding/removing anomaly point

Fig.1 shows how PCA used to detect anomaly circles in fig.1 shows normal data points and square indicated anomaly point and arrow shows principal direction and dotted arrow shows difference in direction when adding anomaly point and normal data point.

III.II Detection of Anomaly using Oversampling PCA

In previous PCA method it calculates component n times and also required memory to store data matrix, and eigen vector, because of that its time complexity increases, also memory cost. but in oversampling method it uses only most dominant eigen vector. After that it calculates score of outlierness of newly added data instance, then it compares that score with previously determined threshold. If score is above previously determined threshold then newly added data instance is considered as infected data instance.

In PCA method which is used to find infected data, calculates n PCA analysis for vast scale data, it has computational overhead. 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 objective data instances are oversampled, that means it creates many copies of objective data instances. If that objective instance is infected then it will show large difference between principal directions of data flow.

As oversampling concept is nothing but duplicating normal data point and infected data point, so that it will shows difference in principal direction. Fig. 2 shows that when normal data point is oversampled that means duplicated there is no change in principal direction.

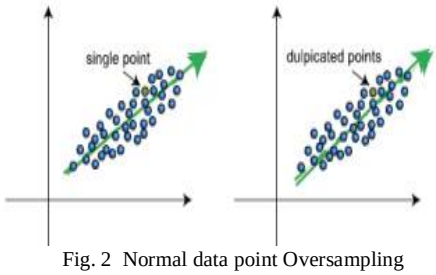

Fig. 2 Normal data point Oversampling

But fig 3 shows that when outlier is oversampled principal direction shows more difference. 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 outliernessof 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.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

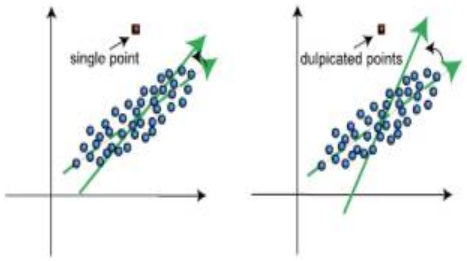

Fig. 3 Anomaly oversampling

IV. CLUSTER ALGORITHM SUBCLU

Algorithm 1: SUBCLU cluster Algorithm
Input: Data Set
Process:
Step 1:Generate k-Dimensional Cluster is generated.
Step 2: Generate (k+1) Dimensional cluster.
k=1
while $C_k \neq \Phi$
-Generate (k+1) D subspace cluster.
-Test Candidate and generate (k+1)
Dimensional cluster.
Output: S k-Dimensional cluster.
C set of cluster in k dimension subspace
Where, C_k is set of clusters

Algorithm 2: Anomaly Detection via Online Oversampling PCA
Input: The data matrix and the set of outliers .
Process:
Step 1: Compute the first principal direction u
Step 2: for I =1 to n do
Step 3: principal component found that is u
Step 4: Score of Outlier.
Step 6: end for
Output: Outlier

V. IMPLEMENTATION DETAILS

This section gives system overview in detail, architecture of proposed system,
The following figure 1 shows the architectural of the proposed system. Following description gives step by step description of architecture.

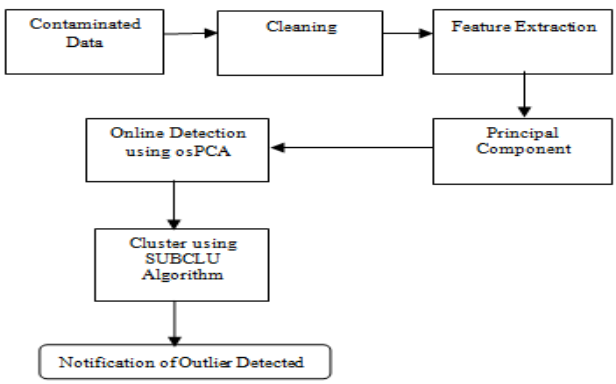

Fig 4 : System Architecture

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

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 SUBCLU which is used to find subspace [9].

3. Principal Component Analysis
   In this phase principal component is calculated on oversampled data instances.\

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

5. Cluster
   By assumption data is selected for detection of anomaly. So it happens that outliers may assumed as normal data in detection technique. So to overcome this problem clusters are created for input dataset. By using SUBCLU algorithm[11] data is divided into multiple clusters. SUBCLU algorithm is basically used for high dimensional data. And detection is performed on threshold.

6. Notifications
   In existing system detection is done on TCP packets. In proposed system detection is done on UDP packets. After detecting outliers in UDP packets, system will give notification.

VI. RESULT AND DISCUSSION

As our proposed system used for detection of anomaly , attack over TCP and UDP protocol. Following fig. 5 shows less memory required by using osPCA for TCP and UPD packets. As we are working with osPCA that is oversampling PCA so there is no need to store whole data matrix and covariance matrix. Fig 5 shows that memory required is less because of this method.

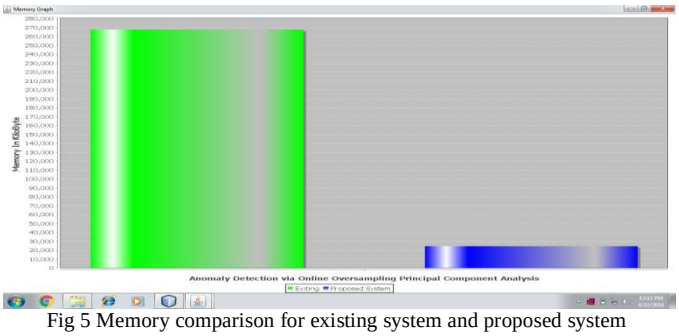

Fig 5 Memory comparison for existing system and proposed system

osPCA for UDP and TCP with the numbers of test of each attack type and time comparisons is shown in fig 6 There are four attacks are found, This fig shows that TCP requires more time than UDP on UDP packets.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

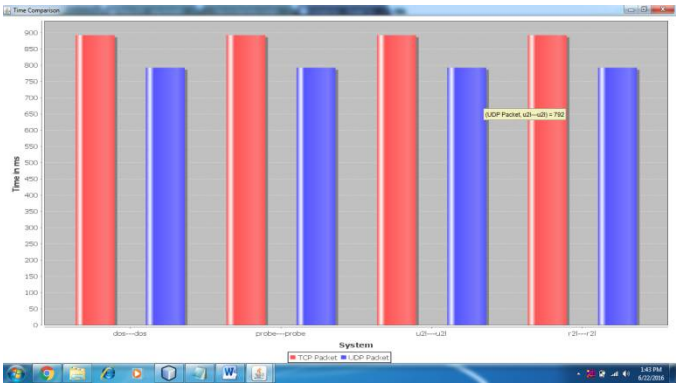

Fig 6 Time comparison four attacks on TCP and UDP

VII. CONCLUSION AND FUTURE SCOPE

In this paper, we proposed subspace clustering algorithm with an online detection of anomaly method based on oversample PCA. When oversampling a data instance, proposed method osPCA identify infected data instances even in large amount of data and also it is more robust for detection as we are using cluster. Also it requires less memory space.

ACKNOWLEDGMENT

I would like to thanks my guide Prof. S. A. Patil Mam and We are also thankful to the reviewer for their valuable suggestions.

REFERENCES

[1] D.M. Hawkins, "Identification of Outliers", Chapman and Hall, 1980.
[2] Hans-Peter Kriegel, Peer Kroger "Outlier detection in Axis Parallel Subspaces of High Dimensional Data" page No.831-838, 2009
[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 15:1-15:58, 2009.
[4] Snehal Thokale, Sonali A Patil "Survey for Different Techniques for Anomaly Detection" International Conference on Science, Technology, Engineering and Management 2016.
[5] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-Based Outlier Detection in High-Dimensional Data," Proc. 14th ACM SIGKDDInt'l Conf. Knowledge Discovery and data Mining, page No 444-452 2008.
[6] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A Comparative Study of Anomaly Detection Schemes in NetworkIntrusion Detection," Proc. Third SIAM Int'l Conf. Data Mining,2003.
[7] Y.-R. Yeh, Z.-Y. Lee and Y.-J. Lee, "Anomaly Detection viaOversampling Principal Component Analysis," Proc. First KESInt'l Symp. Intelligent Decision Technologies, pp. 449-458, 2009,
[8] 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.
[9] https://en.wikipedia.org/wiki/SUBCLU
[10] Sonali. A. Patil, Snehal Thokale "Cluster in High Dimensional Data to Detect Outlier" International Journal of Science and Research Volume 5 Issue 6 | 01 July 2016 2016.

BIOGRAPHY

1. **Ms. Snehal S. Thokale** is a PG student in the Computer Engineering Department, JSPM's BSIOTR, Pune University. Her research interests are Data Mining, Computer Networks (wireless Networks) etc.

2. **Prof. Sonali A Patil is** Assistant Professor in the Computer Engineering Department, JSPM's BSIOTR, Pune University, Pursuing Phd from BSAU, Chennai Interested Domain Cloud computing, Data Mining, Software Engineering, Network Security, Grid Computing