Online Oversampling Principle Component Analysis for Streaming Data Problems

P.Indira Priya, A.Annapoorani

Abstract— In a network unauthorized access to a computer is more prevalent that involves a choice of malicious activities. Hence it is essential for a system to be aware of normal system activities. Data mining is the process of summarizing information from various perspectives. Confessing of sensitive information poses another security risk when large amount of interrelated data is processed. This is where intrusion detection plays a vital role in identifying the suspect. Several data mining approaches for intrusion detection have been proposed as a countermeasure and anomaly detection is one such technique. However , most anomaly detection methods are typically implemented in batch mode, and thus cannot be easily extended to large scale problems without sacrificing computation and memory requirements. In this paper , the system propose an online oversampling principle component analysis (osPCA) algorithm is used to detect the presence of outliers from a large amount of data via an online updating technique. In terms of accuracy and efficiency, the osPCA is preferable for online large scale or streaming data problems.

Index Terms— Anomaly, distance-based, density-based, principal component analysis, Intrusion detection, machine learning, outlier detection.

1 INTRODUCTION

In the modern world, computer has become an inevitable resource. To get connected with one another and to share the information, network has become an emerging technology. As the users are geographically distributed, resources and information sharing are done only via internet. The users are not aware that the people with whom they are connected are authorized users. So that many hackers pretend to be the authorized users and try to hack the private information. To overcome this problem Intrusion Detection System (IDS) is developed [2]. The idea of Intrusion Detection was started in 1980. Any action that is performed by a user illegally is called Intrusion. Intrusion Detection is detecting such anomalous activity at computing and networking resources [1].

The techniques for intrusion detection are of two major categories as shown in Fig. 1: Misuse Detection and Anomaly Detection. Misuse detection- Catch the intrusions in terms of the characteristics of known attacks or system vulnerabilities [3]. Anomaly detection- Detect any action that significantly deviates from the normal behavior. Anomaly detection is robust against new types of attacks and it can learn by examples no need to write rules by hand. So, it has become hot topic in the field of computer security [1], [2].

Data Mining is the growing technology in the field of Computer Science. The trend of research and development on data mining is expected to be flourishing because the huge amounts of data have been collected in databases and the necessity of understanding and making good use of such data in decision making has served as the driving force in data mining.

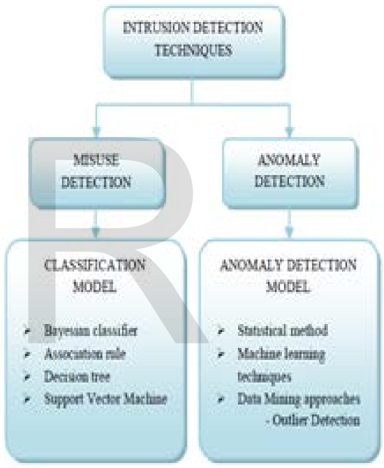


Fig. 1 Classification of Intrusion Detection

Moreover, with the fast computerization of the society, the social impact of data mining should not be under-estimated. When a large amount of interrelated data is effectively analyzed from different perspectives, it can pose threats to the goal of protecting data security and guarding against the incursion of privacy. It is a tricky task to develop effective techniques for preventing the disclosure of sensitive information in data mining, especially as the use of data mining is rapidly increasing in various domains.

Data mining involves six familiar classes of tasks:

1. Anomaly detection
2. Association rule learning
3. Clustering
4. Classification
5. Regression
6. Sequential pattern mining

The presence of errors, missing values and outliers may affect the tasks performed on normal data. So the outliers are predicted and removed from normal data. The removal of outliers introduced the emerging task called Anomaly Detection. It has used in many applications like NIDS (Network Intrusion Detection System), credit card fraud detection, medical application, industrial damage etc.

In the Intrusion Detection System, anomaly detection has been performed by many techniques. Among which outlier detection technique is applicable for many online applications and it also involves many approaches. This paper is based on the survey made on intrusion detection system that includes outlier detection techniques [1]. It includes both supervised as well as unsupervised learning approaches.

2 Intrusion Detection System

2.1 What Is An Intrusion?

An intrusion can be defined as a rebellion of security to gain access to a system. This intrusion can use several attack methods and can span long period of time. These unauthorized accesses to computer or network systems are often planned to learn the system’s weaknesses for upcoming attacks. Other forms of intrusions are aimed at off-putting access or preventing the access to computer systems or networks[1].

The methods used by intruders can often enclose any one or even combination of the subsequent intrusion types:

- Denial of Service
- Trojan horse
- Viruses and Worms
- Spoofing
- Network Port Scans
- Buffer Overflow

Attack on the test systems were classified into four categories:

- Denial-of-service attacks (DOS)
- Probing surveillance attacks e.g., port scanning
- Remote-to-local attacks (R2L)e.g. guessing password
- User-to-root attacks (U2R)e.g., various “buffer overflow” attacks.

The denial of service attacks attempts to make a system or service unusable to genuine users. Probing/surveillance attacks attempt to map out system vulnerabilities and usually serve as a beginning point for upcoming attacks. Remote to local attacks try to gain local account privilege from remote and an unauthorized report or system. User to root attacks attempt to promote the privilege of a local user to root (or super user) privilege.

2.2 What Is Intrusion Detection System (IDS)?

Intrusion Detection System helps information systems prepare for, and deal with the above given attacks.
can provide the following:

- Can add a superior degree of integrity
- Can sketch user activity from point of entry to point of impact
- Can identify and report alterations to data
- Can computerize a task of monitoring the internet searching for the latest attacks
- Can sense when the system is under attack
- Can discover errors in the system configuration
- Can formulate the security management of your system achievable by non-expert staff

3 Intrusion Detection Techniques

In the Intrusion Detection system, the anomaly detection can be performed by various methods. This paper contains the survey about the outlier detection techniques that are involved in Anomaly detection [2]-[7]. Some of the outlier detection approaches that are used in Anomaly detection are as follows,

- Distance based approaches
- Density based approaches
- Ranking outliers using symmetric neighborhood relationship
- Principal Component Analysis

3.1 Distance Based Approaches

Distance based outlier detection will detect outliers by measuring the distance of an instance with the other instances in a given dataset.

An object O in a dataset T is a DB (p, D) outlier if at least p fraction of the objects in T are >= distance D from O [2], [3].

Some algorithms were designed based on distance based approaches,

- Index-based algorithm
- Nested-loop algorithm
- Cell-based algorithm

Index based algorithm is mainly designed for indexing structures such as R-tree, K-D tree that are built for the multi-dimensional database. The index is used to search for neighbours of each object O within radius D around that object. Once K (K = N (1-p)) neighbours of object O are found, O is not an outlier. Worst-case computation complexity is O(K*n2), K is the dimensionality and n is the number of objects in the dataset. Nested loop algorithm divides the buffer size into two halves. Break data into blocks and then feed two blocks into the arrays. Directly computes the distance between each pair of objects, inside the array or between arrays then decide the outlier. This method also suffers with the same computational complexity [3].

In cell based algorithm divide the dataset into cells with length , L=D/2√k where, K is the dimensionality, D is the distance Define Layer-1 neighbours – all the intermediate neighbour cells. The maximum distance between a cell and its neighbour cells is D Define Layer-2 neighbours – the cells within 3 cell of a definite cell. The bare minimum distance between a cell and the cells outside of Layer-2 neighbours is D To check for outliers, if there are M objects inside, all the objects in the cell are not outlier. If there are M objects inside the cell and its layer-1 neighbours, all the objects in this cell are not outlier. If there are less than M objects inside a cell and neighbour cell in both layers, all the objects in this cell are outliers. But distance based approaches has some drawback that it is not applicable for datasets with different densities [3].

3.2 Density Based Approaches

In this approach, the outliers are defined by comparing the density around a point with the density around its local neighbours. The relative density of a point compared to its neighbours is computed as an outlier score. It has the assumption that density around a normal point is similar to the density around its neighbours and in case of outliers density varies when compared to its neighbours.

The previous approaches are not applicable to detect local outliers. In this case a score called LOF (Local Outlier Factor) is computed for each object in the given dataset. Based on the outlier score the instances are classified as either normal data or outliers. The outlier factor is local in the sense that only a restricted neighbourhood of each object is taken into account.

After computing it defines that a point with LOF>1, is a point in a cluster (a region with homogeneous density around the point and its neighbour) and a point with LOF >> 1 is an outlier [4].

For example,

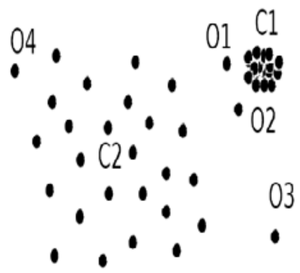


Fig. 2 Outliers

In the above figure o1 and o2 are local outliers to C1, o3 is a global outlier, but o4 is not an outlier. The way in the LOF is computed is given below [4],

- For each data point q calculate the distance to the kth nearest neighbour (k-distance)
- Compute reachability distance (reach-dist) for each data example q with respect to data example p as (1):

$$\text{reach-dist}(q,p) = \max\{k\text{-distance}(p), d(q,p)\} \quad (1)$$

- Compute local reachability density (lrd) of data example q as inverse of the average reachability distance based on the MinPts nearest neighbours of data example q as (2)

$$lrd(q) = \frac{MinPts}{\sum_{p} reach\_dist_{MinPts}(q,p)} \qquad (2)$$

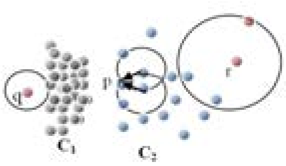
- Compute LOF(q) as ratio of average local reachability density of q's k-nearest neighbours and local reachability density of the data record q as (3)

$$LOF(q) = \frac{1}{MinPts} \cdot \sum_{p} \frac{lrd(p)}{lrd(q)} \qquad (3)$$

3.3 Ranking Outliers Using Symmetric Neighbourhood Relationship

The outlier score measured in the previous approaches is not applicable to complex situation in which the dataset contains multiple clusters with very different density distribution. To avoid this problem outlier scores are measured based on symmetric neighbourhood relationship.

It considers both nearest neighbours (NN) and reverse nearest neighbours (RNN) when estimating the density distribution for each object. The RNNs of an object are essentially the objects that have that object as one of their k nearest neighbours [5]. By considering the symmetric neighbourhood relationship of both NN and RNN, the space of an object influenced by other objects is well determined, the densities of its neighbourhood will be practically estimated, and thus the outliers found will be more meaningful .



- p has two RNNs {s, t}
- q has no RNNS {}
- r has only 1

Each object is assigned with the degree of being Influenced Outlierness (INFLO). The object with higher INFLO is more likely to be an outlier. The object with lower INFLO is the member of cluster. Specifically, the object with INFLO≈1 is the object located at the core part of cluster and the objects with INFLO>1 is an outlier [5].

3.4 Principal Component Analysis (PCA)

The previous approaches are not applicable for high dimensional data. To overcome the „curse of dimensionality in this approach outliers are detected using Principal Component Analysis (PCA). PCA is an unsupervised dimension reduction method. It can retain those characteristics of the data set by keeping the principal components. Those few components often contain the most important aspects of the data. With PCA outliers are detected by means of "Leave One Out" procedure to check each individual point the "with or without" effect on the variation of principal directions [7].

The principal directions for the given dataset are obtained by constructing the data covariance matrix and calculate its dominant eigenvectors. These eigenvectors are the principal directions. The eigenvectors are most informative that holds most of the information about the whole dataset by which the reconstruction error gets reduced [6], [7].

When the PCA concept for anomaly detection is applied for large datasets, adding and removing a single point didn't create much deviation in the principal directions. So the oversampling concept is used. The instance that is added is oversampled (duplicated multiple times). If the added instance is an outlier the principal direction gets deviated and if it is a normal point there won't be deviations. Though oversampling scheme is efficient it causes some computation issues, the principal direction has to be recomputed multiple times [6], [7]. To overcome this drawback this approach proposed two strategies,

- The first one is the fast updating for the covariance matrix
- The another one is the solving the eigen value problem via the power method

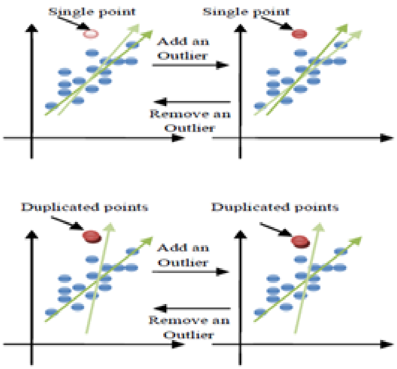


Fig. 4 The effect of adding and removing an outlier as a single point/duplicated points .

The above figure illustrates the deviation in principal direction as an event of adding and removing an outlier. In case of large dataset the deviation can't be identified clearly. So the instance is oversampled (duplicated multiple times). After which the deviation of principal direction in adding an outlier is clearly identified [7].

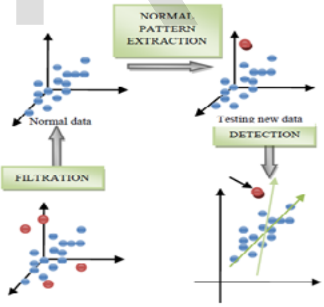


Fig. 5 The Framework of online osPCA

This approach has 2 phases it its framework .
They are,

- Data Cleaning Phase
- On-line Anomaly Detection Phase

In the data cleaning phase, the goal is to spot the distrustful outliers. The outlier score are defined as "one minus the absolute value of cosine similarity". The instances are rank accordingly. Then the outliers in the given data are filtered according to the ranking. In the on-line anomaly detection phase, the goal is to identify the new arriving abnormal instance. The quick updating of the principal directions given in this approach can satisfy the on-line detecting demand. A new arriving instance will be marked if its suspicious score is higher than the mean plus a specified multiple of the standard deviation [6], [7].This online osPCA (oversampling Principal Component Analysis) approach is well suited for online applications as it overcomes the computation issues and utilizes the memory efficiently.

4 Conclusion

This paper has attempted to establish the significance of anomaly detection using outlier techniques. The comparative study of distance-based, density-based and PCA approaches for anomaly detection is given in this paper. In addition we analysed the experimental work of the given 4 techniques on KDD cup dataset. This paper concludes that among the techniques surveyed in this paper PCA has the high AUC score, so that it is found to be more accurate in detecting anomalies/intrusions and it is also efficient for online applications.

References

[1] W. Lee and S. Stolfo, "Data mining approaches for intrusion detection" Proc. of the 7th USENIX security symposium,1998.

[2] E. Bloedorn, et al., Data Mining for Network Intrusion Detection: How to Get Started, MITRE Technical Report, August 2001.

[3] A.K. Jones and R.S. Sielken, Computer System Intrusion Detection: A Survey. Technical report, University of Virginia Computer Science Department, 1999.

[4] 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.

[5] E.M. Knox and R.T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. In"tl Conf. Very Large Data Bases, 1998.

[6] M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," Proc. ACM SIGMOD Int"l Conf. Management of Data, 2000.

[7] W. Jin, A.K.H. Tung, J. Han, and W. Wang, "Ranking Outliers Using Symmetric Neighborhood Relationship," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2006.

[8] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang, "Anomaly Detection via Online Oversampling Principal Component Analysis," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 7, July 2013.

[9] Y.-R. Yeh, Z.-Y. Lee, and Y.-J. Lee, "Anomaly Detection via Oversampling Principal Component Analysis," Proc. First KES Int"l Symp. Intelligent Decision Technologies, pp. 449-458, 2009.

[10] KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[11] A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying Poisson processes," in Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), Philadelphia, PA, Aug. 2006.

[12] H. Teng, K. Chen, and S. Lu, "Adaptive real-time anomaly detection using inductively generated sequential patterns," in Proc. IEEE Comp. Soc. Symp. Research in Security and Privacy, Oakland, CA, May 1990.

[13] T. Singliar and M. Hauskrecht, "Towards a learning traffic incident detection system," in Proc. Workshop on Machine learning Algorithms for Surveillance and Event Detection, Pittsburgh, PA, Jun. 2006.

[14] A. Mu˜noz and J. Moguerza, "Estimation of high-density regions using one-class neighbor machines," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 3, pp. 476–480, Mar. 2006.

[15] T. Ahmed, M. Coates, and A. Lakhina, "Multivariate online anomaly detection using kernel recursive least squares," in Proc. IEEE Infocom, Anchorage, AK, May 2007, to appear.

[16] M. Davenport, R. Baraniuk, and C. Scott, "Learning minimum volume sets with support vector machines," in Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP), Maynooth, Ireland, Sep. 2006.

[17] C. Scott and R. Nowak, "Learning minimum volume sets," J. Machine Learning Research (JMLR), vol. 7, pp. 665–704, Apr. 2006.

[18] Transports Quebec. Organization webpage. [Online]. Available: http://www.mtq.gouv.qc.ca/en/information/cameras/montreal/index.asp

[19] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least squares algorithm," IEEE Trans. Signal Proc., vol. 52, no. 8, pp. 2275–2285, Aug. 2004.