Improving the Performance of Intrusion Detection System by Removing the Count Attribute from KDD Cup 1999 Data

Yash Jain[1] Pratik Jain[2]
[1,2]Department of Computer Science & Engineering
[1,2]IPS Academy, Indore, India

Abstract— Intrusion detection encompasses a range of security techniques designed to detect (and report) malicious system and network activity or to record evidence of intrusion. To understand intrusion detection one must fully understand what intrusion is. Webster's dictionary defines an intrusion as "the act of thrusting in, or of entering into a place or state without invitation or welcome". For the purpose of this article, we will define intrusion as any unauthorized system or network activity on one (or more) computer(s) or network(s). This could be an instance of a legitimate user of a system trying to escalate his privileges so that he can gain greater access to the system that he is currently assigned, or a legitimate user trying to connect to a remote port of a server to which he is not authorized. These intrusions can originate from the outside world, a disgruntled ex-employee who was fired recently, or from your trusted staff. In this paper, one scenario of false positive is considered. The false positive is the case in which the normal data is detected as attack. We are focusing on this problem with the help of an example & proposing one solution for the same problem. The KDD CUP 1999 data set is used. The result of experiment shows that if a class has higher number of counts then this class is considered as an anomaly class. But if the true person is crossing the threshold value of count it will be count as anomaly. To detect the true person & to remove false positive, one solution is proposed.

Key words: Data Mining, Anomaly Detection System (ADS), K-Means, Ensemble, Detection Rate, False Alarm Rate, False Positive, Clustering

I. INTRODUCTION

In the last two decades, computer technology have been utilized by many people all over the world in several areas. With the development of computer technology, security of network system is become an important issue, as network attack have been increased day by day over the past few years. It is very essential to find an effective way to protect our data as it contain highly sensitive information. In the present world, we are having very traditional security such as data encryption, fire wall and VPN. They are good within them. Still they are lacking to detect the attacks by a crackers. However, intrusion detection is a dynamic one which can give dynamic protection to the network security in monitoring attacks and counter attacks.

Network intrusion detection systems (NIDS) generally follow one of three design models. Each design model has its own strengths and weaknesses and many devices are a combination of the three models. These general IDS design categories are signature-based, anomaly-based and protocol modelling.

A. Signature-Based NIDS

This technology analyses packets for specific patterns related to known attacks. This is the most common design: almost all NIDS devices have a strong dependence on signature-based detection at some level. Signature-based detection is relatively easy to understand, deploy, and update, and is good at positively identifying known attacks. However, one drawback to signature-based systems is that they may not detect unknown or modified attacks.

B. Categories of Intrusion Detection Systems

1) Signature Based Detection Systems

Signature based intrusion detection system (SBIDS) based on the known signature. This type of detection is more effective against known attacks, and it depends on the continuance updating signature. The main drawback of SBIDS is, it is unable to detect the unknown attacks and novel attacks, but the detection rate is higher than anomaly intrusion detection rates [9].

2) Anomaly Based Detection System

Anomaly based intrusion detection system (ABIDS) has attracted many researchers due to its capability of detecting novel attack. Novel attack detection is technique for identification of unidentified attack that the machine learning system is not aware during training [10]. ABIDS has two main advantages over SBIDS, First is the ability to detect unknown and "zero day" attack. This is done by comparing the normal activity with that of deviation from them. Second one is the normal activity profile are customized for system, network and therefore making it very difficult for an attacker to know with certainty what activities it can carry out without getting detected [11]. The efficiency of the system depends on how well it is implemented and tested on all protocols. The major drawback of anomaly detection is defining its rule set.

3) Protocol Modeling

Protocol modeling is performed by analyzing network traffic for abnormal protocol activity and alerting on traffic with certain designated protocols or protocols that are unknown to the system. Protocol modeling relies on several different data sources to determine what normal protocol activity is. Common sources for this data may include protocol specification RFCs, popular applications that use that protocol, and thorough analysis of normal network traffic.

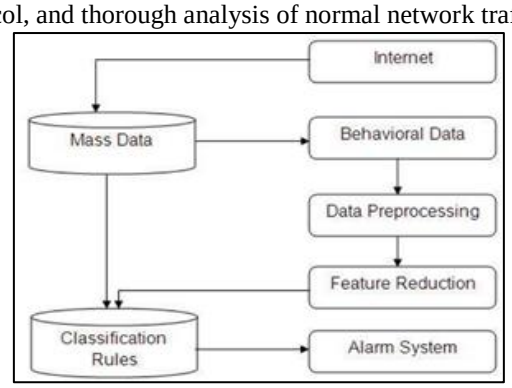

Fig. 1: Process of Intrusion Detection Systems[8]

II. LITERATURE SURVEY

V. Chandola et al, They used Hybrid detection framework that depends on data mining classification and clustering techniques [1]. Francesco Mercaldo, in his work there aim is to use data mining techniques including classification tree and support vector machines for anomaly detection. The result of experiments shows that the algorithm C4.5 has greater capability than SVM in detecting network anomaly and false alarm rate by using 1999 KDD cup data [2]. D. Denning, Algorithm utilizes a feature extraction algorithm called symbolic dynamic filtering (SDF)[3]. In SDF, time-series data are partitioned for generating symbol sequences that then construct probabilistic finite state automata (PFSA) to serve as features for pattern classification [4]. Ugo Fiore et al, in this paper, it is firstly understand the behavior of the leaning method when noise increases because it could alter the capability of extracting correct rules. Effectiveness is evaluated with 3 metrics: Max rule confidence, Precision and Recall [5]. T. Bhavani et al, they uses Cluster Analysis for Anomaly Detection. We used a simple K-mean clustering procedure2. K-mean clustering is a simple, well-known algorithm. It is less computer-intensive than many other algorithms, and therefore it is a preferable choice when the dataset is large [6].

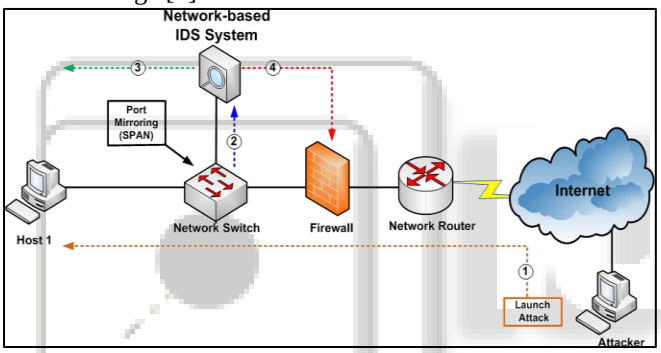

Fig. 2: Architecture of network base IDS system

T. Bhavani et al, they uses Cluster Analysis for Anomaly Detection. We used a simple K-mean clustering procedure2. K-mean clustering is a simple, well-known algorithm. It is less computer-intensive than many other algorithms, and therefore it is a preferable choice when the dataset is large [6]. S. Lina et al, for High dimensional dataset these fixed number of cluster given by user are not good estimation, because it leads to inefficient data distribution or its leads to various outlier [7]. B. Thuraisingham, Network intrusion detection systems employ signature-based methods or data mining-based methods which rely on labeled training data. Anomaly network intrusion detection method based on Principal Component Analysis (PCA) for data reduction and Fuzzy Adaptive Resonance Theory (Fuzzy ART) for classifier is presented [8]. S. wu et al, New hybrid intrusion detection system using intelligent dynamic swarm based rough set (IDS-IR) for feature selection and simplified swarm optimization for intrusion data classification [9]. B. Singh et al, The approach is studied through simulation and applied to an industrial case study. The results suggest potential use for decision making in production management. It uses Algorithm for the creation of a dynamic network based on work order data [10]. M. Xue et al, they used hybrid approach for IDS based on data mining. The main method is clustering analysis with aims of improve detection rate and decreases false alarm rate [11]. K.Wankhade et al, in this paper, Anomaly traffic detection system based on the Entropy of network features and Support Vector Machine (SVM) are compared. Further, a hybrid technique that is combination of both entropy of network features and support vector machine is compared with individual methods [12]. A.Samad, Focuses on detailed comparative study of several anomaly detection schemas for identifying different network intrusion [13]. J. Jonathan, They present a new density-based and grid-based clustering algorithm that is suitable for unsupervised anomaly detection [14].

III. PROBLEM IDENTIFICATION

Intrusions are controlled by intrusion detection systems. An Intrusion Detection System IDS secure the network & protect it. It has the ability to detect anomalous activity automatically. The techniques for the detection of the anomalous activity are classified into two groups:-

A. Predefined intrusion behavior

It first stores the pattern of malicious behavior which is related to intrusion & then judge the intrusion according to the obtained pattern. It has the higher detection accuracy & having low false alarm rate. The main disadvantage of it is that it can only find predefined patterns intrusions.

B. Predefined normal behavior

It first stores the pattern of user's normal behavior into the database & then judges the normal behavior according to the stored pattern. If the deviation is huge enough, we can say that there is anomalous activity[2], [3], [4].

An Intrusion Detection System (IDS) requires high accuracy and detection rate as well as low false alarm rate. In general, the performance of IDS is evaluated in term of accuracy (AC), detection rate (DR), and false alarm rate (FAR) as in the following formula:

- Accuracy = (TP+TN) / (TP+TN+FP+FN)
- Detection Rate = (TP) / (TP+FP)
- False Alarm Rate = (FP) / (FP+TN)

Actual	Predicted Normal	Predicted Attack
Normal	TN	FP
Intrusions	FN	TP

Table 1: General Behavior of Intrusion Detection Data

- True positive (TP) means attack data detected as attack.
- True negative (TN) means normal data detected as normal.
- False positive (FP) means normal data detected as attack.
- False negative (FN) means attack data detected as normal.

Now, the problem is related to false positive. In which the normal data is detected as intrusion. For that we have to understand how data is considered as normal or anomaly. For that we will take data of KDD Cup 1999 data. The 1998 DARPA Intrusion Detection Evaluation Program was prepared and managed by MIT Lincoln Labs. The objective was to survey and evaluate research in intrusion detection. A standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment, was provided. The 1999 KDD intrusion detection contest uses a version of this dataset. When I observe this data & compare the normal classes & anomaly classes. I found that it takes 41 attributes to check whether the input is of normal class or of anomaly class. The attributes are

(duration, protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, urgent, hot, num_failed_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creations, num_shells, num_access_files, num_outbound_cmds, is_host_login, is_guest_login, 'count', srv_count, serror_rate, srv_serror_rate, rerror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, srv_diff_ host_rate, dst_host_count, dst_host_srv_count, dst_host_ same_srv_rate, dst_host_diff_srv_rate, dst_host_same _src_port_rate, dst_host_srv_diff_host_rate, dst_host _serror_rate, dst_host_srv_serror_rate, dst_host_ rerror_rate, dst_host_srv_rerror_rate) 'class' {'normal', 'anomaly'}

Now, these 41 attributes are the one who decides whether the data in normal or anomaly. For example: Let us consider four data of KDD Cup 1999 dataset.

1) Example 2.1

0,udp,other,SF,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,13,1,0.00,0.00,0.00,0.00,0.08,0.15,0.00,255,1,0.00,0.60,0.88,0.00,0.00,0.00,0.00,0.00, normal

2) Example 2.2

0,tcp,http,SF,232,8153,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,5,5,0.20,0.20,0.00,0.00,1.00,0.00,0.00,30,255,1.00,0.00,0.03,0.04,0.03,0.01,0.00,0.01, normal

3) Example 2.3

0,tcp,finger,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,24,12,1.00,1.00,0.00,0.00,0.50,0.08,0.00,255,59,0.23,0.04,0.00,0.00,1.00,1.00,0.00,0.00,anomaly

4) Example 2.4

0,tcp,private,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,48,16,1.00,1.00,0.00,0.00,0.14,0.06,0.00,255,15,0.06,0.07,0.00,0.00,1.00,1.00,0.00,0.00, anomaly

The Intrusion detection system is working on 41 attributes to calculate the anomalous behavior. There are particular values of each attribute & if any of the entry is deviating from the mean value then it is consider as anomaly. We have the problem of false positive in the intrusion detection system. We have to remove count attribute to solve the false positive in IDS. The problem is that, we have to find out the attack before time. By this we can take action against the attack. 41 attributes takes time to find anomalous behavior. We have to decrease the number of attributes if possible. If we can decrease the attribute we can improve the efficiency. The main factors must be kept under measurement while removing attribute. So, rather than improving algorithm, we have to work on attributes.

IV. SOLUTION

Due to the count attribute there is increase in the rate of false positive. To evaluate our system we were interested in two major indications of performance: the detection rate and the false positive rate. The detection rate is defined as the number of intrusion instance detected by the system divided by the total number of intrusion instances present in the test set. The false positive rate is defined as the total number of the normal instance that was (incorrectly) classified as intrusions divided by the total number of normal instances. These are good indicators of performance since they measure what percentage of intrusions the system is able to detect and how many incorrect classifications it make is the process. We calculate these values over the labels data to measure performance.

So, the problem is of authentication. If we can prove authenticity to the system. By this, the count attribute is of no use so we can remove the count attribute by providing an OTP or one-time password to the user Email address or to the contact number. Hence by using this our problem is able to solve easily.

A. Algorithm 1: Registration

1) Start
2) Fill all the fields of the registration form. Including username, email id & passwords.
3) If there are incomplete information in the field or fields then
   Show "error message" in dialog box
4) Else
   Register successful.
5) Exit.

B. Algorithm 2: Login

1) Start
2) Input username & password.
3) If username & password are correct then Login successfully
4) Else (for i=1 to i= 10)
   // (Where i is the number of attempts)
   Repeat 1 to 2.
5) Generate onetime password (OTP) & send it to the email id of the user.
6) If OTP is correct Repeat 1 to 4
   Else
   Show "Wrong OTP".
7) Exit.

V. CONCLUSION

In the current scenario, many people suffer from these when they have to open account with the help of internet banking & because of having more accounts they have more password in their memory. In case of encountering with three wrong attempts they are blocked by that bank's website for next 24 hours. In this paper, the solution is given for the particular problem. So if this solution is followed by system the problem of false positive can be reduced.

REFERENCES

[1] Chandola,A.Banerjee,V.Kumar, "Anomaly detection as a survey" ACM Comput. Surv.41(3)(2009)15:1–15:58.
[2] Francesco Mercaldo, "Identification of anomalies in processes of database alteration" IEEE 2013.
[3] Dorothy E. Denning. "An Intrusion- Detection Model" 1986 IEEE Computer Society Symposium on Research in Security and Privacy , pp 118-31.
[4] S. K. Chaturvedi1 , Prof. Vineet R. , Prof. Nirupama T. "Anomaly Detection in Network using Data mining Techniques" International Journal ISSN 2250-2459 Volume 2, Issue 5, May 2012.
[5] UgoFiore, Francesco, Aniello "Network anomaly detection with the restricted Boltzmann machine" Neurocomputing 122 (2013) 13–23.
[6] T. Bhavani et al., "Data Mining for Security Applications," Proceedings of the 2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing Volume 02, IEEE Computer Society, 2008.

[7] Shih-Wei Lina, Kuo-Ching Yingb, Chou-Yuan Leec, Zne-Jung Leed "An intelligent algorithm with feature selection and decision rules applied toanomaly detection" Elsevier 2011.
[8] Bhavani Thuraisingham "Data Mining for Malicious Code Detection and Security Applications" 2009 IEEE/WIC/ACM 2009.
[9] Shu Wu, Member, and Shengrui Wang "Information-Theoretic Outlier Detection for Large-Scale Categorical Data" VOL. 25, NO. 3, MARCH 2013.
[10] Bharat singh,Nidhi Kushwaha and OP vyas "Exploiting Anomaly Detections for high Dimensional data using Descriptive Approach of Data mining" IEEE(ICCT) 2013.
[11] M. Xue, C. Zhu, "Applied Research on Data Mining Algorithm in Network Intrusion Detection," jcai , pp.275-277, 2009 International Joint Conference Artificial Intelligence, 2009.
[12] Kapil Wankhade, Mrudula Gudadhe, Prakash Prasad, "A New Data Mining Based network Intrusion Detection Model", In Proceedings of ICCCT 2010, IEEE, 2010, pp.731-735.
[13] Abdul Samad bin Haji Ismail "A Novel Method for Unsupervised Anomaly Detection using Unlabeled Data" IEEE 2008.
[14] Jonathan J, Davis, Andrew J. Clark "Data preprocessing for anomaly based network intrusion detection: A review" Elsevier 2011.

