Eliminating the Attribute Count from Intrusion Detection System to Reduce the Problem of False Positive in the Network

[1]Vivek Rai, [2]Diamond Jonawal, [3]Pratik Jain

[1]*Department of computer science and engineering, Patel College of Science & Technology, Indore, India*
[2] *Department of computer science and engineering, Patel College of Science & Technology, Indore, India*
[3] *Department of computer science and engineering, IPS Academy, Indore, India*

Abstract: Intrusion Detection System (IDS) has become an integral part of any network. They became easy way to detect anomalies. Today we require an efficient system having high accuracy and detection rate as well as low false alarm rate. Most of the previously proposed methods suffer from the drawback of low detection rate and high false alarm rate. 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. Experimental results show that the class is considered as an anomaly class if it has high number of count. 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.

Keywords - Intrusion detection system, data mining, clustering, k-means, ensemble, detection rate, false alarm rate, false positive

I. Introduction

In the present world, the numbers of attack have been increases exponentially. So, security of the network became an important issue. Now, it is very important to secure our sensitive data stored on any network. In the present world, we are having traditional security such as data encryption, firewall & VPN. They are good within them. Still they are lacking to detect the attacks by crackers. The most challenging threats are intruders. According to Anderson [1] identified three classes of intruders: i) Masquerader: An unauthorized person who penetrates the system's access control, ii) Misfeasor: Legitimate user who accesses unauthorized data & misuse his or her privileges & iii) Clandestine user: An individual who seizes supervisory control of the system. The masquerader is likely to be outsider, the misfeasor generally is an insider & the clandestine user can be either an outsider or an insider.

II. Literature survey

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 [4]. D. Denning, Algorithm utilizes a feature extraction algorithm called symbolic dynamic filtering (SDF)[5]. 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 [6]. 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 [7]. 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 [8]. V. chandola et al, They used Hybrid detection framework that depends on data mining classification and clustering techniques [9]. 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 [10]. 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 [11]. 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 [12]. J. Jonathan, They present a new density-based and grid-based clustering algorithm that is suitable for unsupervised anomaly detection [13]. 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 [14]. A.

Samad, Focuses on detailed comparative study of several anomaly detection schemas for identifying different network intrusion [15]. 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 [16]. 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 [17].

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:

(1)Accuracy = (TP+TN) / (TP+TN+FP+FN)

(2) Detection Rate = (TP) / (TP+FP)

(3) False Alarm Rate = (FP) / (FP+TN)

TABLE 1: General Behavior of Intrusion Detection Data

Actual	Predicted Normal	Predicted Attack
Normal	TN	FP
Intrusions	FN	TP

I. True positive (TP) means attack data detected as attack.

II. True negative (TN) means normal data detected as normal.

III. False positive (FP) means normal data detected as attack.

IV. 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.

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 ,0.00,0.00,0.00,0.08,0.15,0.00,255,1,0.00,0.60,0.88,0.00,0.0 0,0.00,0.00,0.00, normal

Example 2.2

0,tcp,http,SF,232,8153,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,5,5,0.2 0,0.20,0.00,0.00,1.00,0.00,0.00,30,255,1.00,0.00,0.03,0.04,0 .03,0.01,0.00,0.01, normal

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, 1.00,0.00,0.00,0.50,0.08,0.00,255,59,0.23,0.04,0.00,0.00,1.0 0,1.00,0.00,0.00,anomaly

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 problem is related to the false positive. In this case, the definition is given as, if the attribute count is crossing certain threshold value say 20. It will call that class as anomaly. The normal class can be considered as anomaly class in other case. The domain of the problem is related to internet banking. Consider the case of bank account of a person. Now, if person is trying to open in two or three attempts he or she can be login easily. But, if the number of count is increasing they are consider as intrusion in the

network & the person is blocked to access that account for a day. This can be the case of false positive. By taking an example, we can understand the problem of it. Suppose any person "A" is having "50" Bank accounts & they have different password for that. Person "A" remembers the passwords but he or she is not able to map it properly i.e. say password of account 1 is "ABC!@#", password of account 2 is "BCD!@#",password of account 3 is "EFG!@#" & so on up to password of account no. 50. In terms of login, person "A" is not able to map the password & he or she tries it 50 times maximum to open the account. Here, person remembers 50 passwords in his memory but he or she is unable to map it. So, every time Person "A" tries one of its 50 passwords to login. As the person is trying to login every failed password is increasing the counts & making them close to anomaly class. In the above example 2.1 & example 2.2, the value of attribute "count" is 13 & 5 respectively. It is considered as normal class. But in example 2.3 & example 2.4, the value of attribute "count" is 24 & 48 respectively. It is considered as anomaly class. So, because of the number of count it is consider as anomaly. The problem is that if person "A" is trying to login & if he or she exceeds the threshold value of attribute "count". It is considered as intrusion. But according to the scenario person "A" is true person & it is making the case of false positive.

IV. Solution

One solution is to provide OTP message to the customer's mobile. It provides authentication to the user. But it has one drawback. One solution is to generate OTP & send it on the cell phone of a particular person as a text message. The drawback is very unique in its account. In this scenario, we require OTP to be generated by the system & cell phone to receive that OTP. The cell phone consist of registered SIM. The problem arises with the presence of SIM. The user has to register its following Cell phone number & believe that he or she is the only person using that number. But cloning is possible with the SIM. In the past, we have an example of FIR No. 191/10u/s 419/420/468/471 IPC was registered at PS Darya Ganj, Delhi. At the instance, 8 cloned credit cards of City Bank & Chase and 8 SIM cards of Airtel, Vodafone, Reliance, PIP, Hotlink, EWI as well as the recharge coupons have been recovered. So, if the SIM is being cloned by the person then they can break the authentication process.

The solution of the same problem is given in this paper. The problem can be solved with help of messages. The message is the reliable source to authenticate the person. If person is crossing certain value of attribute count. The system has to send one message to his or her email address to confirm his or her authentication & giving more number of chances to them to enter password. Finally, giving them sufficient amount of chances can solve the problem of the user to use that facility. So, when person "A" is trying to enter his or her password & crossing certain value of count, say 10. Then the system automatically generates one message & passes it to the email address of person "A". Person "A" enters that message into the system & authenticates itself as a right person. After authentication system gave him more chances to enter password. We can define that system in such a manner that, after every 10 wrong password entry, person has to authenticate itself with the help of a message. We can make these happen infinite times to the users because the user is authenticating itself with the help of message.

Algorithm to reduce false positive problem - The algorithm is designed to authenticate the person twice. The first authentication is to enter the username & password (which is compulsory). The second authentication is required when the user enters 10 wrong passwords on a particular username. To authenticate that person we use OTP. To overcome the problem of false positive, we are using onetime password (OTP). A onetime password (OTP) is a password that is valid for only one login session or transaction. The algorithm is divided into two parts.

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.

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
7. Else
   Show "Wrong OTP".
8. 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] James P. Anderson Co. Box 42 Fort Washington, Pa. 19034. 215 646-4706. Computer security threat monitoring and surveillance. Contract 79F296400. February 26, 1980. Revised: April 15, 1980.

[2] V.K. Pachghare, Parag Kulkarni, Deven M. Nikam, "Intrusion Detection System Using Self Organizing Maps", In Proceedings of IAMA 2009, IEEE, 2009.

[3] D.E. Denning, "An intrusion detection model," IEEE Transaction on S/W Engineering, 1987.

[4] 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.

[5] Dorothy E. Denning. "An Intrusion- Detection Model" 1986 IEEE Computer Society Symposium on Research in Security and Privacy , pp 118-31.

[6] 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.

[7] Francesco Mercaldo, "Identification of anomalies in processes of database alteration" IEEE 2013.

[8] UgoFiore , Francesco, Aniello "Network anomaly detection with the restricted Boltzmann machine" Neurocomputing 122 (2013) 13–23.

[9] V. Chandola,A.Banerjee,V.Kumar, "Anomaly detection as a survey" ACM Comput. Surv.41(3)(2009)15:1–15:58.

[10] M. Xue , C. Zhu, "Applied Research on Data Mining Algorithm in Network Intrusion Detection," jcai , pp.275-277, 2009 International Joint Conference on Artificial Intelligence, 2009.

[11] 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.

[12] Bharat singh,Nidhi Kushwaha and OP vyas "Exploiting Anomaly Detections for high Dimensional data using Descriptive Approach of Data mining" IEEE(ICCT) 2013.

[13] Jonathan J, Davis , Andrew J. Clark "Data preprocessing for anomaly based network intrusion detection: A review" Elsevier 2011.

[14] Shih-Wei Lina, Kuo-Ching Yingb, Chou-Yuan Leec, Zne-Jung Leed "An intelligent algorithm with feature selection and decision rules applied toanomaly intrusion detection" Elsevier 2011.

[15] Abdul Samad bin Haji Ismail "A Novel Method for Unsupervised Anomaly Detection using Unlabeled Data" IEEE 2008.

[16] Shu Wu, Member, and Shengrui Wang "Information-Theoretic Outlier Detection for Large-Scale Categorical Data" VOL. 25, NO. 3, MARCH 2013.

[17] Bhavani Thuraisingham "Data Mining for Malicious Code Detection and Security Applications" 2009 IEEE/WIC/ACM 2009.

