Volume 2, Issue 5, May 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

IDS Improved with K-Means Algorithms, Self Organizing Map and Auto Class

Vivek Vashishtha*
GCET Greater Noida

Durgesh kumar
GCET Greater Noida
durgesh.durge@gmail.com

***Abstract*— Intrusion detection system is use to detect suspicious activities is one form defence. This paper aim to build an Intrusion detection system that can detect known and unknown network intrusions automatically. Under a data mining framework ,the ids are trained with unsupervised leaning algorithms, namely the k-means algorithms , self organizing map and auto class. Based on these unsupervised learning algorithms, these novel ids methods are proposed and tested . Much higher detection rates are obtained with reasonable true positive rates, when compared to the best result obtained on the KDD1999 data set. Moreover, this ids is modularized so as to simplify the incorporation of new algorithms when necessary .We perform experiments on the tcpdump data and extract appropriate feature . Clear distinction between normal and abnormal data is observed when data mining techniques are applied on these features. A series of experiments have been conducted on KDD1999 dataset .Different detection methods are tried to see how they perform in our ids. Bearing KDD 1999 winner's result a very high detection rate has been obtained, although with a reasonable true positive rate.**

***Keywords*— KK1999,weka, NET lab, K-mean algorithm,**

I. INTRODUCTION

With the fast development in information technology. It is cheaper and easier to develop and deploy computer networks of all shapes and sizes. Unfortunately ,it is also cheaper and easier to probe and attack our networks. Therefore keeping our network secure becomes vitally important . Data is very vital to an organization. Organizations usually wish to preserve the confidentiality of their data. With the widespread use of the internet , it has become a key challenge to maintain the secrecy and integrity of organization vital data . Conventional techniques for network security mechanisms like user authentication , cryptography and intrusion prevention systems like firewalls. Intrusion Detection System address problems that are not solved by these techniques . For instance , firewalls simply act as a fence around a network .It is incapable of recognizing attacks against a network . An IDS is capable of recognizing this attack which firewalls is not able to prevent . Also , newer attack are being developed that are able to penetrate through firewalls. We need newer approaches to defend against these newer kind of attacks IDS provide a solution to this problem. Current ids are far from intelligent in that they solely rely on human intervention to operate effectively. Therefore ,a more advanced log analysis tool is highly desired .it should be capable of detecting known and unknown intrusion intelligently and automatically, distinguishing normal network activates from those abnormal(very possibly malicious)ones with minimum human inputs. Intrusion detection system can be classified in two broad categories Misuse detection : the system learns patter from already know attacks. There learned patterns are searched through the incoming data to find intrusion of the already know types . This method is not capable in detecting new attack that do not follow pre-defined patterns.

Anomaly detection : here patterns are learned from normal data. The unseen data is checked and searched to find deviation from these learned patterns. These deviations are anomalies or possible intrusions . this method is not capable of identifying the type of attack

II.BACKGROUND RELATED WORK

The related work will be separated into three sections

1. Traditional log analysers :-

a)snort- Snort is a libpcap-based packet sniffer and logger that can be used as a lightweight[4] network intrusion detection system. It features rules based logging to perform content pattern matching and detect a variety of attacks and probes, such as CGI attacks, SMB probes, and much more. Snort has real-time alerting capability, with alerts being sent to separate alert file. The detection engine is programmed using a simple language that describes per packet tests and actions. Ease of use simplifies and expedites the development of new exploit detection rules. Snort is cosmetically similar to tcpdump but is more

focused on the security applications of packet sniffing. The major feature that Snort has which tcpdump does not is packet payload inspection. Snort decodes the application layer of a packet and can be given rules to collect traffic that has specific data contained within its application layer. This allows Snort to detect many types of hostile activity. Another advantage is that its decoded output display is somewhat more user friendly than tcpdump output. One powerful feature that Snort and tcpdump share, is the capability to filter traffic with Berkeley Packet Filter(BPF) commands. This allows traffic to be collected based upon a variety of specific packet fields.

2 Data mining approaches to Intrusion Detection There is often the need to update an installed IDS[3] due to new attack methods or upgraded computing environments. Since many IDSs are constructed by putting rule base of attack signatures. Hence updating these rule bases of IDSs is expensive and slow. Hence we use data mining framework for generating attack signatures. The central idea is to utilize auditing programs to extract an extensive set of features that describe each network connection or host session, and apply data mining programs to learn rules that accurately capture the behaviour of intrusions and normal activities . These rules can then be used for misuse detection. Data mining generally refers to the process of extracting descriptive models from large stores of data. Recent rapid development in data mining has made available a wide variety of algorithms particularly useful in mining audit data.

FEATURE SELECTION

2. Dataset used in feature extraction can be very huge. Hence employing a efficient data structure in very important. Hence I have used DBMS to store all the connection records. At first feature important with respect to all classes are to be found out. Here we are using KDD dataset. For this, all the attacks are distributed into five categories. Before applying, feature analysis steps, we have to convert continuous attributes into discrete attributes. A lot of algorithms are available like decision tree, Error, Entropy, Equal frequency method. Here I have used equal frequency interval method.

3. **K-means algorithms:-**

K-Means is an simple learning algorithm for clustering analysis. The goal of K-Means algorithm[11] is to find the best division of *n* entities in *k* groups, so that the total distance between the group's members and its corresponding centric, representative of the group, is minimized. Formally, the goal is to partition the *n* entities into *k* sets S_i, *i=1, 2, ..., k* in order to minimize the within-cluster sum of squares (WCSS), defined as:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^j - c_j \|^2$$

where term $\| x_i^j - c_j \|$ provides the distance between an entity point and the cluster's centroid.

The most common algorithm, described below, uses an iterative refinement approach, following these steps:

1.Define the initial groups' cancroids. This step can be done using different strategies. A very common one is to assign random values for the cancroids of all groups. Another approach is to use the values of *K* different entities as being the cancroids.

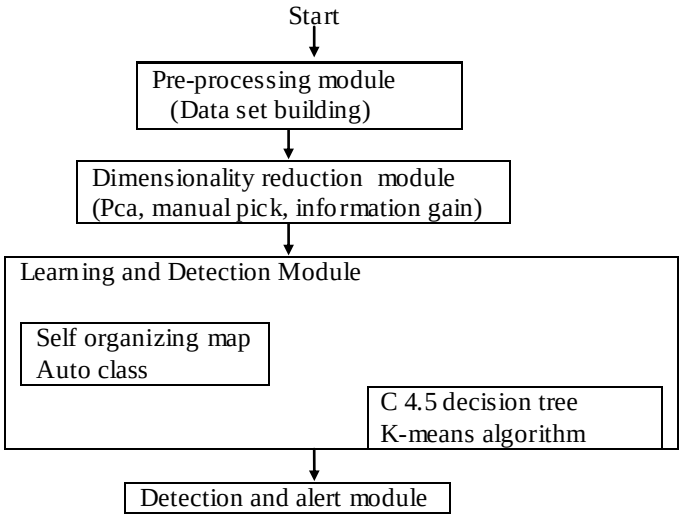
2.Assign each entity to the cluster that has the closest centroid. In order to find the cluster with the most similar centroid, the algorithm must calculate the distance between all the entities and each cancroids.

3.Recalculate the values of the cancroids. The values of the centurion's fields are updated, taken as the average of the values of the entities' attributes that are part of the cluster.

4.Repeat steps 2 and 3 iteratively until entities can no longer change groups.

The K-Means is a greedy, computationally efficient technique, being the most popular representative-based clustering algorithm

III.SYSTEM FRAMEWORK



Auto Class

Auto class [14],[5] is an unsupervised classification system that seeks a maximum posterior probability classification. In is based on classical mixture model and implemented by a Bayesian method to determine the optimal class numbers. Traditional clustering algorithm do this automatically discovery of data classes by partitioning the cases from up to down or reversely conglomerating the cases unlike their auto class attempt to fine the best class description in a model space and avoid over filtering data by enforcing a

trade off between the fit to the data and complexity of the class description.

Weka tools

Weka is a collection of machine learning algorithms for data mining tasks[7]. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

NET lab

The NET Lab Tool kit [7]is a system for integrating tangible interaction and media. Designed for project sketching and production, the toolkit enables novices and experts to integrate hardware, media and interactive behaviours for products, installations, and research.

Experiment Design

Detection Rate =(number of network attack detection) /(number of network attack)

True Positive Rate=(True Positive)/(True Positive + False Positive)

Data Set Choice

This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad" connections, called intrusions or attacks, and ``good" normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

Connection type	Training Sets	Test Set
Normal	19.60%	19.48%
Probe	0.83%	1.34%
Dos	79.24%	73.90%
U2r	0.01%	0.07%
R21	0.23%	5.20%

KDD 1999Winner results

Predicted / Actual	0	1	2	3	4	% correct
0	60262	243	78	4	6	99.5
1	511	3451	184	0	0	83.3
2	5299	1328	223226	6	0	97.1
3	168	20	0	30	10	13.2
4	14527	294	0	8	1360	8.4
% correct	73.6	64.8	99.9	71.4	98.8	

Table for KDD1999 Winner

0	Normal
1	Probe
2	Dos(Denial of services)
3	User of roots
4	Remote to local

Categories of network attacks

EXPERIMENT RESULTS

Construct a new data set with all the u2r attacks and assume it as the abnormal dataset. Train k-mean algorithm on the normal dataset to get a model with 33 clusters (here it should be noted the auto class fail to provide meaning full decision boundary and therefore not trained in this experiment). Mark those points on the inter clusters with decision boundary as network intrusion.

Calculate the true positive rate

Run	True positive	Positive	True Positive Rate%
1	219	59146	0.3703
2	219	58976	0.3713
3	220	54980	0.4001
4	217	59396	0.3953
5	219	54706	0.4003
6	220	54392	0.4045
7	218	59742	0.3649
8	215	59295	0.3626
9	221	54786	0.4034
10	219	59448	0.3684
Mean	218.70	57496.7	0.3811

I. CONCLUSIONS AND FUTURE WORK

This paper initially aimed to build a ids especially , finding out proper detection methods for known and unknown network intrusion. Experiments based on the KDD1999 data set show some encouraging results. The achievement of this project will be evaluated firstly. The next subsection will discuss some of the limitations in our IDS . finally, some possible extensions are discussed.

Due to the time limit , several interesting ideas have not been implemented yet , However , it will be worthwhile trying them in the future .

Real time detection

Two approaches may provide real time capability in our log analyzer. One is to extract the parameters of mixture models learned so far to a real time detector , which uses the mixture model to classify the incoming network records . As calculating the class membership probability needs only those model parameters, it is computational very in signification.

During clustering, auto class [6]. Has generated an interesting byproduct, the weight of features. In the influence value reports given by auto class, an ordered list of normalized feature influence value summed overall classes printed. This gives a rough heuristic measure of relevant influence of each feature in differentiating the classes from the overall dataset.

REFERENCES

[1] Wikipedia Article of IDS. Available online at http://en.wikipedia.org/ids.html.

[2]Amor, N.B., Benferhat, S., Elouedi, Z., 2004. Naive Bayes vs decision trees in intrusion detection systems. In: SAC '04: Proceedings of the 2004 ACM Symposium on Applied Computing. ACM Press, New York,NY,USA,pp.420–424.

[3] Parel Berkhin, Survey of clustering data mining techniques, Technical report , accrue Software , San jose , CV , 2002

[4] C.Bishop , Novelty detection and neural network validation 1994.

[5] CERT coordination center (CERT/CC), (CERT/CC statistics 1988-2003).

[6]B. S. Everitt and DJ hand , finite miture distributions . Chapman and hall , London NewYork 1981.

[7]weka application tols in NET lab weka.html

[8]Wenke Lee and Salvatore stolfo. Date mining approaches for intrusion detection

[9]P.Mitra, C.A. Murthy, and S.K.pal unsuperuised feature selection using feature .

Similarity , IEEE transactions on pattern analysis and machine intelligences 24(4), 2002

[10]The self – organizing Map KOHONEN

[11] G.D. Ramkumar,S.Ranka, and S. Tsur, "Weighted Association Rules: Model and Algorithm," Proc. Fourth ACM Int'l Conf. Knowledge Discovery and Data Mining, 1998.

[12] M. Roesch, "SNORT—Lightweight Intrusion Detection for Networks," Proc. USENIX 13th Systems Administration Conf. (LISA '99), pp. 229-238, 1999.

[13] F. Tao, F. Murtagh, and M. Farid, "Weighted Association Rule Mining Using Weighted Support and Significance Framework," Proc. Ninth ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD), pp. 661-666, 2003.

[14] G.B. White, E.A. Fisch, and U.W. Pooch, "Cooperating Security Managers: A Peer-Based Intrusion Detection System," IEEE Network, pp. 20-23, Jan. 1996.

[15] Y. Xie, H. Kim, D.R. O'Hallaron, M.K. Reiter, and H. Zhang, "Seurat: A Pointillist Approach to Anomaly Detection," Proc. Seventh Int'l Symp. Recent Advances in Intrusion Detection (RAID)

[16] M. Celenk; T. Colony; J. Willies; J. Graham; , "Predictive Network Anomaly Detection and Visualization," Information Forensics and Security, IEEE Transactions on , vol.5, no.2, pp.288-299, June 2010.

[17] W. Yunwu, "Using Fuzzy Expert System Based on Genetic Algorithms for Intrusion Detection System", Information Technology and Applications, IFITA, pages 221-224, 2009.

