ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 496-501(2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

An Analysis of K-means Algorithm Based Network Intrusion Detection System

Yi Yi Aung*, Myat Myat Min

Faculty of Computer Sciences, University of Computer Studies, Mandalay, UCSM, 0000, Myanmar

ARTICLE INFO

Article history:
Received: 30 November, 2017
Accepted: 30 January, 2018
Online: 10 February, 2018

Keywords:
Network Intrusion Detection System
K-means
Random Forest
KDDCup 99

ABSTRACT

In this modern age, information technology (IT) plays a role in a number of different fields. And therefore, the role of security is very important to control and assist the flow of activities over the network. Intrusion detection (ID) is a kind of security management system for computers and networks. There are many approaches and methods used in ID. Each approach has merits and demerits. Therefore this paper highlights the similar distribution of attacks nature by using K-means and also the effective accuracy of Random Forest algorithm in detecting intrusions. This paper describes full pattern recognition and machine learning algorithm performance for the four attack categories, such as Denial-of-Service (DoS) attacks (deny legitimate request to a system), Probing attacks (information gathering attacks), user-to-root (U2R) attacks (unauthorized access to local super-user), and remote-to-local (R2L) attacks (unauthorized local access from a remote machine) shown in the KDD 99 Cup intrusion detection dataset.

1. Introduction

On the Internet, users share valuable information around the world. The internet has created various ways to threaten the stability and security of interrelated systems. Both of these mechanisms are static and dynamic. Static mechanisms like firewalls and software updates provide dynamic security and mechanisms such as intrusion detection systems. Today, security is the most serious problem for getting valuable information. Therefore, static mechanisms or dynamic mechanisms are needed to protect individual information despite the precautionary technology. The intrusion detection system detects not only successful aggression, but also helps monitor and prevent timely action.

The intrusion Detection System (IDS) is a standard component of a security infrastructure that allows network administrators to detect policy violations. Check all incoming and outgoing network activity and determine suspicious patterns that indicate network or system attacks from people trying to break or compromise the system.

A secure network must provide the following:

- Data confidentiality: Data transferred over the network must be accessible only to data that has been approved accordingly.
- Data integrity: Data must maintain integrity from when it is sent when it is received. No damage or loss of data from random events or malicious activities is accepted.
- Data availability: The network must be resistant to service attack denial.

IDS technology based on tracking process can be categorized into two approaches:

Abuse/Signature detection: This technology searches for signature attacks and known signatures in network traffic and are used as a reference to detect future attacks. Regularly updated databases are usually used to store signatures of known attacks. The way this technology controls intrusion detection is similar to antivirus software. The advantage of this type of detection is that it can accurately and efficiently detect known attacks. Anomaly detection: This technology is based on tracking traffic anomalies. The gap between traffic is monitored and regular profiles are measured. Different implementations of this technology have been reserved based on metrics used to measure the deviation of traffic profiles. The advantage of this detection type is that it is well suited to detect unknown attacks.

IDS are divided into two parts based on analysis and retention of audit data:

* Yi Yi Aung, Email: yiyiaungresearch@gmail.com

Host-based IDS (HIDS): HIDS is a home based tracking method that allows the system to collect data in the form of multiple host activity records, such as event logs and system logs. Since everything is in the host, there's no need to install additional hardware or software [1]. The advantages of hosted IDS are to check the success or failure of attacks, monitor system activity, and detect attacks that IDS networks cannot detect, close tracking and real-time responses, are not required.

Network-based IDS (NIDS): NIDS is a network approach that collects data directly from a network monitored as a packet instead of collecting data from a particular host / agent. Most NIDS are a free and easy-to-use operating system [2]. Network-based IDS offers advantages such as low cost of ownership, easier, placement, network attack detection, evidence preservation, real-time tracking and rapid response, and detection of failed attacks.

2. Literature Review

Most of intrusion detection system focused on four major attack categories such as denial of service, probe, user-to-root, and remote-to-local but this author specially emphasized on User-to-Root (U2R) attacks in NSL-KDD dataset by using Weka tool. This paper focused on a comparative study analysis of user-to-root attack, which the attacker tries to access normal user account and gains root access information of the system based on several machine learning techniques such as navie bayes, random forest, J48, etc [3].

This paper analyzed anomaly intrusions detection system by using Random Forest classifier with Principal Component Analysis. The author got experimental results by using simulation connection dataset of NSL-KDD. The performance of the system was measured by using Precision, Recall and F-Measure. And also this paper was specially focused on to detect various attacks present in Denial of Service (DoS) such as Neptune, Smurf, Pod, Teardrop, Land, Back, Apache2, Processtable, Mailbomb [4].

This paper used C4.5, CART (Classification and Regression Trees), Random Forest, and REP (Reduced Error Pruning) Tree to investigate the detection of intrusions contained in KDDCUP 1999 DARPA dataset. And compared the performance of the above algorithms based on the measures such as Accuracy, Learning Time (in seconds) and Size of the Tree. According to the experimental results, Random Forest was better as it correctly identifies more number of instances than other. And the accuracy of the REP Tree was very less than other algorithms but the learning time of REP Tress is very less than other [5].

They used Support Vector Machine with Principal Component Analysis (PCA) to choose the optimum feature subset that was useful in applying for intrusion detection system. To determine the effectiveness and feasibility of the proposed IDS system, they choosed NSL-KDD dataset for simulation their system. They found that PCA algorithm is good to select a best subset of features for classification of intrusions. It can help to speed up the training and testing process of intrusions detection which is important for high-speed network applications [6].

In this proposed paper, several classification techniques and machine learning algorithms have been considered to categorize the network traffic. Out of the classification techniques, they have found nine suitable classifiers like BayesNet, J48, PART, JRip, Random Tree, Random Forest and REPTree. The comparison of these algorithms has been performed using WEKA tool [7].

Security has become a crucial issue for computer systems. IDS can protect to our computer network. Different classification and clustering algorithms have been proposed in recent year for IDS. In this paper, multiple algorithms were analyzed to find the optimal algorithm. At last the optimal algorithms Random Forest and DB Scan were occurred for IDS [8].

The purpose of this survey paper was to describe the methods/ techniques which are being used for Intrusion Detection based on Data mining concepts and the designed frame works. This survey paper stated the methods and techniques of data mining to aid the process of Intrusion Detection and the frameworks [9]. The concept of intercepting these two different fields, gives more scope for the research community to work in this area. New approaches enhanced the existing interference detecting system and it was a stepping stone to build effective and efficient IDS to detect different types of attacks [10].

This paper proposed a novel hybrid model for intrusion detection. The proposed framework in this paper may be expected as another step towards advancement of IDS. The Hybrid framework led to effective, adaptive and intelligent intrusion detection [11].

This paper drew the conclusions on the basis of implementations performed using various data mining algorithms. Combining more than one data mining algorithms had be used to remove disadvantages of one another and lead to a better performance than any single classifier. Different classifiers had different knowledge regarding the problem [12].

3. Methodology

This section consists of the conversation of the two algorithms of data mining classification approaches. These are K-means and Random Forest.

3.1. K-means Clustering Algorithm

Clustering, based on distance measurements performed on objects, and classifying objects (invasions) into clusters. Unlike classification, classification because there is no information about the label of learning data is an unattended learning process. For anomalous detection, we can use welding and in-depth analysis to guide the ID model. Measurement of distance or similarity plays an important role in collecting observations into homogeneous groups. Jacquard affinity measurement, the longest common order scale (LCS), is important that the event is to awaken the size to determine if normal or abnormal. Euclidean distance is approximately two vectors X and Y in space Euclidean n-dimensions, the size of the distance widely used for vector space. Euclidean distance can be defined as the square root of the total difference of the same vector dimension. Finally, grouping and classification algorithms need to be channeled effectively, massively, it possible to handle dimension of network data and heterogeneity [13].

In this paper, we use K-means algorithm to cluster dataset connections. The K-means algorithm is one of the widely recognized clustering tools. K-means groups the data in accordance with their characteristic values into a user-specified number of K distinct clusters. Data categorized into the same cluster have identical feature values. K, the positive integer denoting the number of clusters, needs to be provided in advance. The steps involved in a K-means algorithm are given consequently: [14]

1. K points denoting the data to be clustered are placed into the space. These points denote the primary group centroids.
2. The data are assigned to the group that is adjacent to the centroid.
3. The positions of all the K centroids are recalculated as soon as all the data are assigned.
4. Repeat steps 2 and 3 until the centroid unchanged.

This results in the partition of data into groups. The preprocessed dataset partition is performed using the K-means algorithm with K value as 5. Because we have the dataset that contains normal and 4 attack categories such as DoS, Probe, U2R, R2L.

3.2. Random Forest Algorithm

One of the most popular methods or frameworks used by scientists in the science of data is Random Forest. It is a supervised classification algorithm. It can be seen from its name, which is to create a forest by some way and make it random. There is a direct relationship between the numbers of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach [15].

Random Forests grows many classification trees. Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random – but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on this m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

There are many of top benefits of Random Forest algorithm. Some of these benefits are as follows:

- Accuracy
- Runs efficiently on large data bases
- Handles thousands of input variables without variable deletion
- Provides effective methods for estimating missing data
- Maintains accuracy when a large proportion of the data are missing

4. KDDCup 99 Dataset

The evaluation of any intrusion detection algorithm on real network data is extremely difficult mainly due to the high cost of obtaining proper labeling of network connections. Due to the real sample cannot be gotten for intrusion detection, the KDDCup'99 dataset is used as the sample to verity the performance of the misuse detection model. The KDDCup'99 dataset, referred by Columbia University, was arranged from intrusions simulated in a military network environment at the DARPA in 1998. It contains network connections obtained from a sniffer that records all network traffic using the TCP dump format. The total simulated period is seven weeks. It was performed in the MIT Lincoln Labs, and then announced on the UCI KDD Cup 1999 Archive [16].

KDDCup'99 dataset have two variations of training dataset; one is a full training set having 5 million connections and the other is 10% of this training set having 494021 connections. Since the whole dataset is huge, the experiment has been performed on its smaller amount of dataset that is 10% of KDD. Additionally, the KDDCup'99 dataset includes many attack behaviors, classified into four groups: Probe, Denial of Service (DoS), User to Root (U2R), and Remote to Local (R2L) [17]. These can be seen in table I. Normal connections are created to profile that expected in a military network. The detailed information of the two variations of training dataset can be seen in table II.

Table I: Various Attacks and Categories

Categories	Attacks Subclass
DoS	back, land, Neptune, pod, smurf, teardrop
Probe	ipsweep, nmap, portsweep, satan
U2R	buffer_overflow, loadmodule, perl, rootkit
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster

Table II: Number of Instances in KDD and 10% KDD

Class	Whole KDD	10 % KDD
DoS	3883370	391458
Probe	41102	4107
U2R	52	52
R2L	1126	1126
Normal	972780	97278
Total	4898430	494021

The data set includes 41 features classifying the data records into normal or a type of attacks. The features consist of 34 types of numeric features and 7 types of symbolic features, according to different properties of attack. The nature of features can be divided into the following groups [18].

- Basic Features: Basic functions can be obtained from the packet header without checking the load.

- Content Features: Domain knowledge is used to assess the original TCP packet load. This includes features such as the number of unsuccessful login attempts.
- Time-based Traffic Features: This function is designed to capture properties in the 2-second window. Examples of such functions are the number of connections to the same host every two seconds.
- Host-based Traffic Features: Use the history window to estimate the number of connections (in the case 100) and not the time. Therefore, host-based functionality is designed to assess attacks that include two or more intervals.

4.1. Pre-Processing

KDDCUP 99 data set is pre-processed in order to make it suitable for the data mining learning algorithm. Pre-processing is performed for the following reasons.

Each record in the dataset consists of categorical as well as numeric features. Textual (plain) data is used for categorical features. K-means algorithm needs numeric data (either discrete or continuous). The first step in pre-processing is to covert this categorical feature attributes to numeric attributes. For converting symbols into numerical form, an integer code is assigned to each symbol. For instance, in the case of protocol type feature, 0 is assigned to tcp, 1 to udp, and 2 to the icmp symbol and so on. The dataset contains three categorical attributes while the rest of the thirty eight attributes are numeric. Every category of an attribute is assigned a specific number.

We have used K-means and Random Forest to define normal and attacks in the system. They need specific format so we have converted the dataset to K-means and Random Forest compatible format.

5. Experimental Results and Discussion

To facilitate the experiments, we used eclipse java and weka tool to implement the algorithms on a PC with 64-bit window 7 operating system, 4GB RAM and a CPU of Intel core i3-4010U CPU with 1.70GHz. Data come from MIT Lincoln laboratory of KDDCup99 data set. The table lists the number of instances available in the whole dataset, 10% of KDDCup'99 dataset.

The analysis is performed by using K-means and Random Forest algorithms. We use K-means algorithm to generate heterogeneous dataset to nearly homogeneous dataset. The clustering results of K-means algorithm are described from table III to table VIII.

Table III: Detailed Information of Attack Categories in Cluster-1

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	107219	107217	2
Probe	1610	1605	5
U2R	0	0	0
R2L	6	3	3
Normal	10	3	7
Total	108845	108828	17

Table IV: Detailed Information of Attack Categories in Cluster-2

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	1067	1065	2
Probe	1221	1207	14
U2R	4	0	4
R2L	1	0	1
Normal	21235	21230	5
Total	23528	23502	26

Table V: Detailed Information of Attack Categories in Cluster-3

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	280782	280782	0
Probe	0	0	0
U2R	0	0	0
R2L	0	0	0
Normal	16	14	2
Total	280798	280796	2

Table VI: Detailed Information of Attack Categories in Cluster-4

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	2203	2202	1
Probe	12	1	11
U2R	46	29	17
R2L	1087	1068	19
Normal	75409	75398	11
Total	78757	78698	59

Table VII: Detailed Information of Attack Categories in Cluster-5

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	187	186	1
Probe	1264	1255	9
U2R	2	0	2
R2L	32	22	10
Normal	608	604	4
Total	2093	2067	26

Table VIII: Detailed Information of Attack Categories with Clustering

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	391458	391452	6
Probe	4107	4068	39
U2R	52	29	23
R2L	1126	1093	33
Normal	97278	97249	29
Total	494021	493891	130

By analyzing the clustering results, the characteristics of Denial of Service (DoS) attacks are mostly related to themselves in cluster-3. And then, it is closely similar to the nature of Probe attacks in cluster-1. Probe attacks are also mostly related to DoS attacks in cluster-1. And then, it is nearly same with the nature of

Normal by looking in cluster-5. Normal is mostly similar nature with User-to-Root attacks and Remote-to-Local attacks by studying in cluster-4. And then, Normal is related to Probe by studying cluster-2 and cluster-5. Normal is related to all attacks by looking in all 5 clusters because attacks mimic to normal behavior in intrusions.

Then we apply Random Forest algorithm to know the intrusions and normal traffic. The performance of attacks categories with Random Forest algorithm in 5 clusters of K-means can be seen from table IX to table XIV. The Precision and Recall of the normal and attacks detection are good and the false positive rate is nearly zero.

Table IX: Performance Analysis of Attack Categories in Cluster-1

Attacks	False Positive Rate	Precision	Recall
DoS	0.00738	0.999888	0.999981
Probe	0.000009	0.999377	0.996894
U2R	0	0	0
R2L	0.000018	0.6	0.5
Normal	0.000018	0.6	0.3

Table X: Performance Analysis of Attack Categories in Cluster-2

Attacks	False Positive Rate	Precision	Recall
DoS	0	1	0.998125
Probe	0.000224	0.995874	0.988533
U2R	0.000042	0	0
R2L	0	0	0
Normal	0.008722	0.999058	0.999764

Table XI: Performance Analysis of Attack Categories in Cluster-3

Attacks	False Positive Rate	Precision	Recall
DoS	0.125	0.999992	1
Probe	0	0	0
U2R	0	0	0
R2L	0	0	0
Normal	0.875	0	1

Table XII: Performance Analysis of Attack Categories in Cluster-4

Attacks	False Positive Rate	Precision	Recall
DoS	0	1	0.999546
Probe	0	0	0.083333
U2R	0.000165	0.690476	0.630434
R2L	0.000077	0.994413	0.98252
Normal	0.000495	0.999483	0.999854

Table XIII: Performance Analysis of Attack Categories in Cluster-5

Attacks	False Positive Rate	Precision	Recall
DoS	0.001049	0.989361	0.994652
Probe	0.00965	0.993665	0.992879
U2R	0.000478	0	0
R2L	0.002911	0.785714	0.6875
Normal	0.00606	0.985318	0.993421

Table XIV: Performance Analysis of Attack Categories with K-means Clustering

Attacks	False Positive Rate	Precision	Recall
DoS	0.000156	0.999959	0.999984
Probe	0.000028	0.99657	0.990504
U2R	0.00003	0.65909	0.557692
R2L	0.000028	0.987353	0.969831
Normal	0.000148	0.99928	0.999701

6. Conclusion

This paper presents a comparative analysis hybrid machine learning technique to detect Denial of Service (DoS) attacks, Probing (Probe) attacks, User-to-Root (U2R) attacks and Remote-to-Local (R2L) attacks. We can know the similar nature of attack group by using K-means algorithm. And then we use Random Forest algorithm to classify normal and attack connections. The experiments show that, KDDCup 99 dataset can be applied as an effective benchmark dataset to help researchers compare different intrusion detection models. Future work includes analyzing with other data mining algorithms to classify attack categories and how it can detect on other real time environment dataset.

References

[1] Aung Yi Yi and Myat Myat Min, "An Analysis of Random Forest Algorithm Based Network Intrusion Detection System", Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNDP), 2017 IEEE/ACIS 18 th International Conference, 2017.

[2] Yan. K.Q., S. C. Wang and C. W. Liu, "A Hybrid Intrusion Detection System of Cluster-based Wireless Sensor Networks", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, March 18-20, 2009, Hong Kong.

[3] S. Revathi and Dr. A. Malathi, "Detecting User-To-Root (U2R) Attacks Based on Various Machine Learning Techniques", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 4, ISSN (Online): 2278-1021, ISSN (Print): 2319-5940, April 2014.

[4] S. Revathi and Dr. A. Malathi, "Detecting Denial of Service Attack Using Principal Component Analysis with Random Forest Classifier", International Journal of Computer Science & Engineering Technology (IJCSET), Vol.5, No. 03, ISSN: 2229-3345, March 2014.

[5] Jayshri R. Patel, "Performance Evaluation of Decision Tree Classifiers for Ranked Features of Intrusion Detection", Journal of Information, Knowledge and Research in Information Technology, Vol. 02, Issue -02, ISSN: 0975-6698, Nov 12 to Oct 13.

[6] Heba F. Eid et al., "Principal Components Analysis and Support Vector Machine based Intrusion Detection System", 10 th International Conference on Intelligent Systems Design and Applications, (IEEE, 2010).

[7] Manzoor, Muhammad Asif, and Yasser Morgan, "Network Intrusion Detection System using Apache Storm" , Special Issue on Recent Advances in Engineering Systems, Advances in Science, Technology and Engineering System Journal (ASTES), Vol. 2, No. 3, 812-818 (2017).

[8] S. Choudhury and A. Bhowal, "Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection", International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015, pp 89-95.

[9] R. Venkatesan, R. Ganesan and A.A.L. Selvakumar, "A Comprehensive Study in Data Mining Frameworks for Intrusion Detection", International Journal of Advanced Computer Research, December-2012, Volume-2 Number-4 Issue-7, ISSN (print): 2249-7277 ISSN (online): 2277-7970.

[10] Somani Manish and Roshni Dubey, "Hybrid Intrusion Detection Model Based on Clustering and Association", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.3, Issue 3, ISSN (Print): 2320-3765, ISSN(Online):2278-8875 March 2014.

[11] M. Dhakar and A. Tiwari, "A Novel Data Mining based Hybrid Intrusion Detection Framework", Journal of Information and Computing Science, 2014, Vol-9 No-1 pp. 037-048, ISSN 1746-7659, England, UK.

[12] TR. Patel, A. Thakkar and A. Ganatra, "A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems", International Journal of Soft Computing and Engineering (IUSCE), March-2012, Vol-2, Issue-1, ISSN: 2231-2307.

[13] Youssef Ahmed and Ahmed Emam, "Network Intrusion Detection Using Data Mining and Network Behavior Analysis", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011.

[14] X. Wu, V.Kumar, Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg, "Top 10 algorithms in data mining", Survey Paper(2008).

[15] S. Devaraju and S. Ramakrishnam, "Performance Comparison for Intrusion Detection System using Neural Network with KDD Dataset", ICTACT Journal on soft Computing , Vol:04, Issue:03, ISSN: 2229-6956, April 2014.

[16] P. S. Rath, M. Hohanty, S. Acharya and M. Aich, "Optimization of IDS Algorithms Using Data Mining Technique", Proceeding of 53rd IRF International Conference, Pune, India,ISBN 978-93-86083-01-2, 2016.

[17] L.S. Parihar and A. Tiwari, "Survey on Intrusion Detection Usingn Data Mining Methods",IJSART, January-2016, Volume-2 Issue-1 ISSN (online): 2395-1052.

[18] Md.E. Haque and T.M. Alkharobi, "Adaptive Hybrid Model for Network Intrusion Detection and Comparison among Machine Learning Algorithms", International Journal of Machine Learning and Computing, February 2015, Vol-5, No-1.