Vol. 5, No. 8 August 2014 ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences

http://www.cisjournal.org

Efficient Intrusion Detection using Weighted K-means Clustering and Naïve Bayes Classification

[1]Yousef Emami, Marzieh [2]Ahmadzadeh, [3]Mohammad Salehi, [4]Sajad Homayoun
Department of Information Technology, Shiraz University of Technology, Shiraz, Iran

[1]Y.emami@sutech.ac.ir, [2]ahmadzadeh@sutech.ac.ir,[3]m.salehi@sutech.ac.ir,[4]s.homayoun@sutech.ac.ir

ABSTRACT

Intrusion detection system (IDS) is becoming a vital component to secure the network. A successful intrusion detection system requires high accuracy and detection rate. In this paper a hybrid approach for intrusion detection system based on data mining techniques is proposed. The principal ingredients of the approach are weighted k-means clustering and naive bayes classification. The C5.0 algorithm is used for ranking attributes, so the attributes receive a weight which is used in K-means clustering therefore accuracy of clustering is increased.

Keywords: *Intrusion Detection System, K-means Clustering, Naïve Bayes Classification*

1. INTRODUCTION

An intrusion detection system (IDS) is a defense system that plays an important role to protect or secure a network system and its prime goal is to monitor network activities automatically to detect malicious attacks. IDS is becoming an absolutely vital component to secure the network. IDS are divided into two types: misuse detection and anomaly detection. Misuse detection first builds pattern for malicious behavior and then identifies intrusion based on this known pattern. The great merit of misuse detection is its higher detection accuracy to all known attacks. Anomaly detection defines the expected behavior of the network or profile in advance. Any significant deviations from such defined expected behavior are reported as possible attacks. The outstanding merit of this approach is that it can examine unknown and more convoluted intrusions [1].

MINDS and EBayes are examples for data mining based anomaly detection model for IDS. IIDS (Intelligent Intrusion Detection System) and RIDS-100(Rising Intrusion Detection System) are examples for data mining based both anomaly and misuse detection model for IDS.

Host data, network log data and alarm messages are examples of diverse sources of information in IDS. Since the variety of different data sources is too complex, the complexity of the operating system also increases. Also network traffic is huge, so the data analysis is very hard [2]. Data mining techniques can help to detect new vulnerabilities as well as intrusions and provide decision support for intrusion management.

Data mining techniques such as classification and clustering are valuable and can be utilized to acquire information about intrusions by observing network data. Various classifiers can also be used to form a hybrid learning approaches such as combination of clustering and classification technique.

Two major achievement of hybrid learning approaches are high detection rate and low false alarm rate. Different classifiers such as combination of clustering and classification technique are used to form a hybrid learning approaches [3].The utilized hybrid learning approach in this paper is a combination of weighted K-means clustering and naïve bayes classification. The weighted K-means clustering algorithm make cluster based on the new Euclidean distance function.

The proposed method executes on the kddcup.data_10_percent Data set, this data set is used in international level for evaluating/calculating the performance of various intrusion detection systems (IDS)[4].

The rest of this paper organized as follows: Section 2 presents literature review .Section 3 discusses the proposed method .Section 4 presents results. Finally section 5 concludes the paper.

2. LITERATURE REVIEW

ADAM (Audit Data Analysis and Mining) is a testbed for using data mining techniques to detect intrusions. ADAM uses a combination of association rule mining and classification to discover attacks in a TCPdump audit trail. First, ADAM builds a repository of normal frequent itemsets Secondly ADAM runs a sliding-window, on-line algorithm that find frequent itenmsets in the last D

Vol. 5, No. 8 August 2014 ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences

http://www.cisjournal.org

connections and compares them with those stored in the normal itemset repository [5].

MADAM ID (Mining Audit Data for Automated Models for intrusion detection) uses data mining algorithm to compute activity patterns from system audit data and extracts predictive features from the patterns. It then applies machine learning algorithm to the audit records that are processed according to feature definition to generate intrusion detection rules [6].

In [7] the authors propose a hybrid learning approach based on combination of k-means clustering and naïve bayes classification to improve current anomaly–based detection capabilities in the term of accuracy, detection rate as well as false alarm rate. The proposed approach is evaluated using KDD CUP 99.

In [8] a comparative study of k-means clustering via naïve bayes classification and naïve bayes classification for identifying novel network intrusion detections is given. The experiments are done on KDDCup 99 data set. Results have demonstrated that k-means clustering via naïve bayes classifier model is much more efficient in the detection of network intrusion, compared to the naïve classification based classification techniques.

Using a parallel clustering ensemble algorithm the high speed, high detection rate and low false alarm rate can be achieved. This algorithm keeps the advantage of the evidence accumulation that combines the results of multiple clustering into a single data partition and then detects abnormal network behavioral patterns with related algorithm [3].

The proposed hybrid intrusion detection system in [9] combines the merits of anomaly and misuse detection. Anomaly detection has high false alarm rate, in order to reduce it k-means algorithm for clustering has been applied for clustering followed by hybrid classifier, combining k-nearest neighbor and naïve bayes classifier for detecting intrusion.

The general structure of the proposed method is rather similar to that of [7],[8] and [9], but deployed K-means algorithms in [7] , [8] and [9] are simple one and leave K-means intact without any preprocessing on it while in the proposed k-means a weighted coefficient is assigned to each attribute then these coefficients are incorporated in updated distance function so that accuracy to be improved.

3. THE PROPOSED METHOD

In figure 1 upon completing normalization, C5.0 algorithm is used to assign a weight to each attribute. The importance of each attribute is determined using related weight. In the next step, the weighted k-means clustering is applied to the dataset so that the clusters to be shaped then naïve Bayes classification is run and classifier model is created. In the end based on the shaped model the performance is reviewed. Weighted K-means and naïve bayes classification are further elaborated in the following sections.

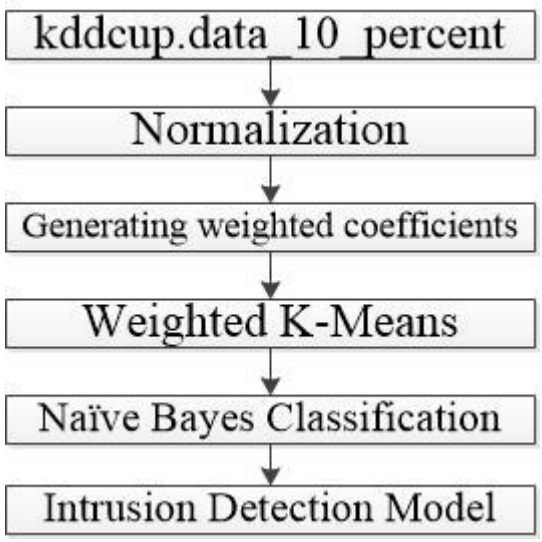

Figure 1: Weighted K-Means and Naïve Bayes

A. Clustering

Clustering technique is a good candidate for detecting intrusion from network data, because clustering methods can unearths complex intrusions over a different time period. In the clustering process, similar objects are assigned to a group and each group is called a cluster. Each group consists of members from the same cluster that are similar and members from different clusters are different from each other [1]. The k-means clustering is a clustering analysis algorithm that group objects based on their feature values into k disjoint clusters. K is a positive integer number specifying the number of clusters and has to be given in advance. Here the Basic k-means algorithm is presented.

Select k point as initial centroid
Repeat
Form k clusters by assigning each point to its closest centroid
Recomputed the centroid of each cluster
Until centroids do not change

The distance function is required in order to compute proximity between two objects.
Euclidean is the most commonly used distance function and is defined as[10] :

$$d(x,y)=\sqrt{\sum}(x^i - y^i)^2$$

The utilized distance function is

$$d(x,y)=\sqrt{\sum} \boldsymbol{w_i} (x^i - y^i)^2$$

This function introduces a new coefficient called Wi, Wi shows the importance of each field, the values for Wi are generated using the C5.0 algorithm. Table 1 demonstrates the assigned weight for each attribute of KDD dataset .The assigned weight to the attributes not mentioned in table 1 is zero hence not taken into consideration.

Table 1-Attribute and weighted coefficient

dst_host_srv_serror_rate	0.0001
Flag	0.0001
dst_host_srv_rerror_rate	0.0003
dst_host_rerror_rate	0.0004
dst_host_count	0.0005
srv_serror_rate	0.0011
Service	0.0021
dst_host_srv_count	0.0025
dst_host_serror_rate	0.004
same_srv_rate	0.0044
dst_bytes	0.0054
Duration	0.0062
dst_host_srv_diff_host_rate	0.0082
Count	0.0106
dst_host_same_src_port_rate	0.0227
root_shell	0.0306
Hot	0.0316
num_compromised	0.0665
protocol_type	0.1233
num_failed_logins	0.1354
dst_host_diff_srv_rate	0.1363
wrong_fragment	0.166
src_bytes	0.2418

The C5.0 algorithm is used for ranking attributes, so the Attributes receive a weight which is used in K-means clustering therefore accuracy of clustering is increased.

B. Classification

Classification is a data mining technique which takes each instance of a data set and assigns it to a particular class. It extracts the models for defining important data classes. Such type of models are called as classifiers. A classification based IDS will classify all the network traffic into either normal or intrusion. Data classification consists of two steps, first step is learning and second step is classification. In the learning step A classifier is formed and in the classification step that model is used to predict the class labels for a given data. Classification is a supervised machine learning mechanism. It can handle only labeled data. So, the major disadvantage of classification technique is that, it is less efficient in the field of intrusion detection as compared to clustering because classification cannot handle unlabeled data, which degrades the performance of intrusion detection system [3].

Naïve Bayesian classification has been successfully used in many fields. It has a solid theoretical foundation and enjoys from smaller error rate than the other classification methods. Naïve bayes is based on very strong independence assumption and the construction of naive bayes is very simple [11].

4. RESULTS

Finding number of clusters (K value) always is a challenge in clustering task. To find appropriate K value, two Steps algorithm is employed (because it does not need to know number of clusters). Afterwards, to take the advantage of K-means, the K value is set to the clusters count output by Two Step algorithm. After clustering dataset into 4 clusters 'a', 'b', 'c' and 'd', the clusters are investigated and similar attacks are found into same clusters. The ‘a’ cluster is the normal vector, the ‘b’ cluster is smurf’ attack, the ‘d’ cluster is ‘neptune’ attack and the ‘c’ cluster is related to other attacks. On the next step, a field is appended to dataset which shows the desirable cluster label from 'a' to 'd' for each instance, while the original field class label was removed. After running Naïve Bayes as a classification algorithm, a classifier model for prediction of 'a' to 'd' created and the confusion matrix depicted as table 1. For example, table 2 shows that our approach can predict 'a' attacks group by accuracy of 100 percent. It means that we can predict a new instance belongs to a group of attack 'a' and doesn't belong to other groups. It helps the network administrators to easily identify an attack type by

denying other clusters attack group and only concentrate on potential attacks to handle.

Table 2: Confusion Matrix

a	b	c	d	
6750	0	0	0	a
25	2643	18	0	b
0	8	61	0	c
58	0	0	1515	d

5. CONCLUSION

In this paper we have used two learning algorithm of data mining for intrusion detection: k-means and naïve bayes classifier .k-means group data sample on the basis of their similarities and dissimilarities by considering related importance weight(achieved by C5.0 algorithms) for each field. K-means output 4 clusters which consist of similar TCP attacks. Naïve Bayes makes a classification model which predict attack group type and help the administrator to identify the attack type earlier and quickly negate the its effects.

REFERENCES

[1] Kapil Wankhade, Sadia Patka ,Ravindra Thool," An efficient approach for Intrusion Detection using data mining methods", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013

[2] Deepthy K Denatious, Anita John," Survey on data mining techniques to enhance intrusion detection", International Conference on Computer Communication and Informatics (ICCCI), 2012

[3] Kapil Wankade,Sadia Patka,Ravindra Thool," An Overview of Intrusion Detection Based on Data Mining Techniques", International Conference on Communication Systems and Network Technologies (CSNT), 2013

[4] kddcup.data_10_percent Dataset, 2014,http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html,

[5] D.Barbara, J. Couto,S.Jajodia and N.Wu,"ADAM : A test bed for exploring the use of data mining in intrusion detection ",SIGMOID,vol30,no.4,pp 15-24,2001

[6] Wenke Lee,Salvatore J.Stolfo ,"A framework for constructing features and models for intrusion detection systems", ACM transactions on information and system security(TISSEC)",vol.3,no.4,2000

[7] Z.Muda,W.Yassin ,M.N.Sulaiman, N.I.Udzir ," Intrusion detection based on K-Means clustering and Naïve Bayes classification" International Conference on Information Technology in Asia (CITA 11), 2011

[8] Sanjay Kumar Sharma, Pankaj Pandey, Susheel Kumar Tiwari, " An improved network intrusion detection technique based on k-means clustering via naïve bayes classification"., International Conference on Advances in Engineering, Science and Management (ICAESM), 2012

[9] Hri Om,Aritra Kundu, " A hybrid system for reducing the false alarm rate of anomaly intrusion detection system", 1st International Conference on Recent Advances in Information Technology (RAIT), 2012

[10] Pang-ning, micheal Steinbach,vipin kumar "Introduction to data mining" ,pearson publication ,2006

[11] Z.muda ,w.yassing ,m.n sulaiman," A k-means and naïve bayes learning approach for better intrusion detection", information technology journal ,vol 10,pp.648-655,2011