Intrusion Detection Systems Using K-Means And Random Forest Algorithms

Authors: Marwa Ali Khaddor

Marwa_79337@svuonline.org. Marrosh90@gmail.com

Dr.Basel Al-Khattib

t_balkhatib@svuonline.org

Syrian virtual university

Syria-Damascus

Abstract :

Many researches nowadays working on increase the ability of the intrusion detection systems IDS, which depend on data mining techniques in their mechanism.

Whereas cyber-attacks of all kinds have been evolving , that require more efforts to fill in the gaps in intrusion detection system, increase their accuracy or the their ability of detect any intrusion and block or stop it, and reduce false alert rate..

In this research .. we review related works about development IDSs , and we propose an idea about merge two algorithms (k-means & Random forest) intention to fill in gaps detect the U2R breakthroughs, represented by reduce false alert rate.

Keywords: intrusion detection system , data mining algorithms, random forest , k-means , NLS KDD

International Journal of Scientific & Engineering Research Volume 11, Issue 9, September-2020
ISSN 2229-5518

I. Introduction:

Internet is the best way to storage and transform data across the world, it is fast and low cost, But it is not completely secure, because of the constant attacks.

It has been develop many intrusion detection techniques, which depend on data mining algorithms (clustering, association, classification).

As well Cyber-attacks has developed, so there are many research papers about improve IDS, Like this paper.

II. Related work:

I. Anomaly Detection using Data Mining Techniques:

Review data mining techniques which used for IDS, and compare them to gather, this paper refers that hybrid techniques is better than the classic, particularly when researcher has merged classification algorithm with clustering one. **[1]**

II. IDS using SVM:

Intrusion detection depend on SVM [Support vector machine], this technique achieved a high accuracy rate and high detection rate also, but it give a high false alarm.

Accuracy rate (ACC) **:** it is the system ability to defined activity as attack or normal activity

Detection rate: it is the number of attacks which has defined it correctly.**[2]**

III. Anomaly Detection Using K-Means Clustering:

Anomaly behavior detection by measure the distance between center and data point in the data set, based on D-max threshold. **[3]**

IV. IDS using clustering:

this paper found that the clustering method better than classification one, because it doesn't need learning, subsequently it is able to detect the new attacks in network. **[4]**

International Journal of Scientific & Engineering Research Volume 11, Issue 9, September-2020
ISSN 2229-5518

V. Intrusion detection metrics:

Metrics help evaluate the performance of an intrusion detection system. Some of the commonly used evaluation metrics used with respect to intrusion detection are False Alarm Rate (FAR), Detection Rate (DR), Accuracy, Precision, Specificity, F-score13. All these evaluation metrics are basically derived from the four basic attributes of the confusion matrix depicting the actual and predicted classes. These elements of the confusion matrix are:

True Negative (TN): Number of instances correctly predicted as non-attacks.

False Negative (FN): Number of instances wrongly predicted as non-attacks.

False Positive (FP): Number of instances wrongly predicted as attacks.

True Positive (TP): Number of instances correctly predicted as attacks.**[5]**

Matrix :

Actual	Predicted Normal	Predicted Attacks
Normal	(TN)	(FP)
Attacks	(FN)	(TP)

III. Hybrid approach K-Means and Random Forest (RF):

Anomaly learning approaches are able to detect attacks with high accuracy and to achieve high detection rates. However, the rate of false alarm using anomaly approach is equally high. In order to maintain the high accuracy and detection rate while at the same time to lower down the false alarm rate, we proposed a combination of two learning techniques. For the first stage in the proposed hybrid learning approach, we grouped similar data instances based on their behaviors by utilizing a K-Means clustering as a pre-classification component. Next, using Random Forest we classified the resulting clusters into attack classes as a final classification task. We found that data that has been misclassified during the earlier stage may be correctly classified in the subsequent classification stage.

IV. Weka:

Waikato Environment for Knowledge Analysis (Weka) is a data mining tool available free of cost under the GNU General Public License. The version used in this study is 3.7.11 that has many state of the art machine learning tools and algorithms for data analysis and predictive modeling. This tool accepts the data file either in comma separated value (csv) or attribute-relation file format (arff) file format. For the simulation, arff files is already available with 42 attributes whereas arff files with lesser attributes as discussed in research methodology section are created through the pre-processing tab of the tool **.[6]**

V. Dataset Description:

In our experiments The NSL-KDD data set with 42 attributes is used in this empirical study. This data set is an improvement over KDD'99 data set .This data set has number of versions available, out of which 20% of the training data is used which is identified as KDDTrain+_20Percent with a total number of 25192 instances. The test data set is identified by the name KDDTest+ and has a total of 22544 instances.

The training dataset contains 24 types of attack, while the testing data contains more than 14 types of additional attack. Further description for the available features and intrusion instances can be found in . KDD dataset covered four major categories of attacks which is Probe, DoS, R2L and U2R.**[7]**

VI. Intrusion Detection System (IDS)

IDS is defined as a malicious, externally induced operation a fault. IDS plays an important role in detecting various types of attacks. The main goal of IDS is to find intrusions and can be considered as classification problem. IDS can be classified into various attacks such as DOS, probe, U2R, R2L.**[8]**

VII. k-means :

International Journal of Scientific & Engineering Research Volume 11, Issue 9, September-2020
ISSN 2229-5518

K−Means Clustering Network intrusion class labels are divided into four main classes, which are DoS, Probe, U2R, and R2L. The main goal to utilize K−Means clustering approach is to split and to group data into normal and attack instances. K−Means clustering methods partition the input dataset into k− clusters according to an initial value known as the seed points into each cluster's centroids or cluster centers. The mean value of numerical data contained within each cluster is called centroids. In our case, we choose k = 2 in order to cluster the data into three clusters (C1, C2).

The K−Means algorithm works as follows:

- Select initial centers of the K clusters.
- Repeat step 2 through 3 until the cluster membership stabilizes.
- Generate a new partition by assigning each data to its closest cluster centres.
- Compute new clusters as the centroids of the clusters.**[9]**

VIII. Random Forest (RF)

Random forest(RF) is an ensemble classifier used to improve the accuracy .Random forest consists of many decision trees. Random forest has low classification error compared to other traditional classification algorithms. Number of trees, minimum node size and number of features used for splitting each node. Advantages of RF are listed below

. 1) Generated forests can be saved for future.

2) Random forest over comes the problem over fitting.

3) In RF accuracy and variable importance is automatically generated

When constructing individual trees in random forest, randomization is applied to select the best node to split on. This value is equal to A, where A is no. of attributes in the data set. However ,RF will generate many noisy trees, which affect accuracy and wrong decision for new sample.

We applied RF on the data sets which resulted from k−means clustering in the previous step, the matrixes results as follow:

table 1: RF with clustered anomaly data set

Actual	Predicated normal	Predicated attacks
Normal	(TN) 2851	(FP)7
Attacks	(FN)10	(TP)8982

table 2: RF with clustered DoS data set

Actual	Predicated normal	Predicated attacks
Normal	(TN) 5502	0 (FP)
Attacks	4 (FN)	986 (TP)

table 3: RF with probe data set

Actual	Predicated normal	Predicated attacks
Normal	(TN) 1674	3 (FP)
Attacks	3 (FN)	2874 (TP)

table 4: RF with R2L data set

Actual	Predicated normal	Predicated attacks
Normal	(TN) 1986	17 (FP)
Attacks	3 (FN)	2900 (TP)

table 5: RF with U2R data set

Actual	Predicated normal	Predicated attacks
Normal	(TN) 961	7 (FP)

Attacks	3 (FN)	1383 (TP)

IX. Result and Discussion:

We see results From tables(1,2,3,4,5), and by applying these rules :

- DR = (TP) / (TP+FP)
- ACC = (TP+TN) / (TP+TN+FP+FN)
- FA =(FP) / (FP+TN)

Finally we had reached to these values as a follow table:

Attack	DR	ACC	FA
anomaly	99.9%	99.9%	0.2%
DOS	100.0%	99.9%	0.0%
U2R	99.5%	99.6%	0.7%
R2L	99.4%	99.6%	0.8%
Probe	99.9%	99.9%	0.2%

X. Conclusions

This paper we merge two types algorithm (classification & clustering). first , k−means has split the data set into two cluster that help classifier to defined the attack data.

the Random Forest (RF) algorithm to detect four types of attack like DOS, probe, U2R and R2L.We adopted 10 cross validation applied for classification. Feature selection is applied on the data set to reduce dimensionality and to remove redundant and irrelevant features. We applied symmetrical uncertainty of attributes which overcomes the problems of information gain. The proposed approach is evaluated using NSL KDD data set. We compared our hybrid approach with previous approach, this approach improved ACC and DR, and it relatively increased the FA.

We hope the next studies can achieve more accurate results.

References:

1. *Dipankar Dasgupta, Robert Kozma Fabio Gonzalez .(2002) .*Combining Negative Selection and Classification .*
2. *Levent Ertoz*, Vipin Kumar*, Aysel Ozgur*, Jaideep Srivastava* Aleksandar Lazarevic .* .(2019)A Comparative Study of Anomaly Detection Schemes in Network Intrusion*
3. *Jitendra Agrawal Shikha Agrawal .(2015) .Survey on Anomaly Detection using Data Mining Techniques .www.sciencedirect.com.*
4. *Prof(Dr.)Laxman Sahoo Sarita Tripathy .(2015) .A Survey of different methods of clustering for anomaly detection .*
5. *W. Yassin, M.N. Sulaiman, N.I. Udzir Z. Muda .(2011) .Intrusion Detection based on K-Means Clustering and.*
6. *(Shweta Srivastava Assistant Professor (CSE Department) ABES Engineering College, 2014)*
7. *Preeti Aggarwala, S. K. (2015). Analysis of KDD Dataset Attributes - Class wise For Intrusion Detection . ScienceDirect, 10.*
8. *Prof. Nilesh Marathe, Prof. Puja Padiya Prof. Ujwala Ravale .(2015) .Feature Selection Based Hybrid Anomaly Intrusion Detection .www.sciencedirect.com.*
9. *Jiaqu Yi, S. L. (n.d.). Apriori Algorithm And K-means Clustering Algorithm.*
10. *Heafield, K. (n.d.). Hadoop Design And K-Means Clustering.*