ISSN:0975-9646

Sonali Rathore et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3073-3076

An Efficient Grid Partition based Classification Algorithm for Intrusion in KDDCup 99

Miss. Sonali Rathore
Dept .of Computer Science & Engg.
Truba Institute of Engg & I.T
Bhopal (M.P)

Prof. Amit Saxena
Dept. of Computer Science & Engg.
Truba Institute of Engg. & I.T.
Bhopal (M.P)

Dr. Manish Manoria
Director
Truba Institute of Engg. & I.T
Bhopal (M.P)

Abstract — Data mining is a technique of analyzing the dataset so that some meaningful information can be extracted so that it can be used for various applications. KDDCup 99 is one of the network based dataset which contains a set of attributes of packets and instances which classifies the packets contains anomalous behavior or not. The existing data mining based classification algorithms are used for the classification of packets but the algorithm implemented contain less correctly classified instances and more error rate which needs to be minimized and accuracy is improved, Hence in the paper an efficient technique for the classification of intrusion using improved form of the classification is implemented which is more efficient as compared to the existing classification algorithm. The proposed technique not only improves the classification accuracy but also minimizes the computational time.

The proposed algorithm is based on the concept of applying clustering on the KDDCup 99 and then these clusteres values are classified using Grid partition based decision tree.

I. INTRODUCTION

The internet is rapidly improving as a platform for deploying sophisticated interactive applications, as people start to use the internet to share information with others. The web can be viewed as a large, transparent database that its information can be retrieved and updated from time to time [1]. Although the shift from desktop-centric applications to web-based computing and cloud computing brings many benefits, such as efficient communication with ubiquitous access and availability, the way that internet users share and exchange information also opens their own information to new web-related security and privacy problems. Today, attackers routinely track the identities of internet-connected users, steal privacy data, abuse users' personal information, and expose users' unsolicited data or programs using malware. Although these attackers can also accomplish these goals by other means, the web has made it much easier for attackers to locate victims, search private information and initiate unsolicited communication with the victims. Therefore, internet users have raised concerns on attacks that can cause billions of dollars in loss [2–5].

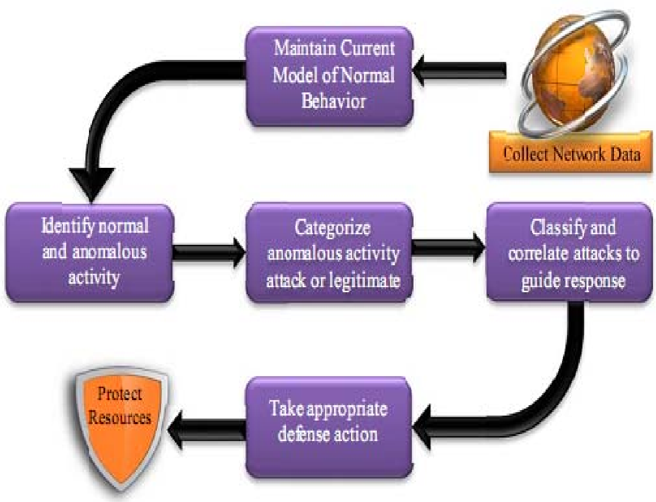

Figure1: Basic Elements of Anomaly Detection Process

Figure:1 identifies the basic steps that must be accomplished in order to use an anomaly detection system for network defense. Existing research in anomaly based network intrusion detection is not evenly distributed among each of the steps The majority of existing work focuses on identifying attacks by detecting anomalies [6-9]. Very little work is dedicated to other elements of the process. While a body of work has been done on distinguishing between legitimate anomalies and attack anomalies, most of this work involves reducing false positive rates examines the act of sanitizing training data used to develop base traffic models. [10-11] discuss approaches to classifying attacks in anomaly based detection systems. There are also proposals for active systems that automatically respond to attacks.

II. LITERATURE SURVEY

S. No.	Paper	Title	Technique Used	Issues
1	Intrusion Detection based on K-Means Clustering and OneR Classification.	Z. Muda, W. Yassin, M.N. Sulaiman, N.I.Udzir.	A new technique of detecting intrusions using hybrid combinatorial method of K-mean clustering and OneR Classification.	Doesn't provides efficient results for large datasets.
2	Intrusion Detection Using Data Mining Techniques	Mohammadreza Ektefa, Sara Memar and Fatimah Sidi	The methodology includes combination of classification tree and Support Vector Machine. After implementing the proposed methodology it proofed that the classification decision tree C4.5 is better than SVM learning algorithm.	Can be applied for other domains such as warehousing.
3	GA-NIDS: A Genetic Algorithm based Network Intrusion Detection System	Anup Goyal and Chetan Kumar	Here implemented intrusion detection using genetic algorithm. This technique also includes a machine learning approach called Genetic Algorithm for the identification of harmful or unwanted attacks in the network.	Boosting algorithm can be applied for the betterment of algorithm.
4	Network Intrusion Detection System Using Fuzzy Logic	R. Shanmugavadivu and Dr.N.Nagarajan	Here in this paper a fuzzy logic based system is developed for the generation of set of rules and from these set of rules intrusions are detected and classified in a better way.	Can't be applied for missing attributes.
5	Intrusion Detection System Using Data Mining Technique: Support Vector Machine	Yogita B. Bhavsar, Kalyani C.Waghmare	A new and efficient way of detecting intrusions using support vector machine. Support Vector Machine is a learning algorithm which is used to classify intrusions on NSL-KDDCup 99 dataset.	Learning approach and hence error rate needs to be reduced for better results.
6	Performance Comparison For Intrusion Detection System Using Neural Network With KDD Dataset.	S. Devaraju and S. Ramakrishnan.	Here in this paper various neural network based classifiers are implemented for the detection of intrusions in the network.	Less efficient and contains more false alarm rate.
7	A Real-time Intrusion Detection System Based on PSO-SVM	Jun Wang, Xu Hong, Rong-rong Ren, Tai-hang Li	Here proposed hybrid combination of PSO-SVM. Here support vector machine is applied on the KDDion of Cup 99 dataset for the classification of intrusions and then particle swarm optimization technique is applied for the optimization.	Provides complex system for the detection of intrusioms.

III. PROPOSED METHODOLOGY

In this work, we are first applying Clustering algorithm on the original data set to form clustered data set. The clustered data set is then partitioned horizontally and vertically into two parties say P1 and P2. After this partition of the dataset into two sub sets Grid based ID3 Decision Tree Algorithm is applied and a decision tree is formed.

This Method has two phases.

1. Cluster the dataset using supervised learning support vector machine.
2. Classified the cluster data Using Enhanced ID3 algorithm with Grid partitioning.

SUPPORT VECTOR MACHINE

Consider training sample, where is the input pattern, is the desired output:

$$W_0^T X_i + b_0 \geq +1, for\ d_i = +1$$
$$W_0^T X_i + b_0 \leq -1, for\ d_i = -1$$

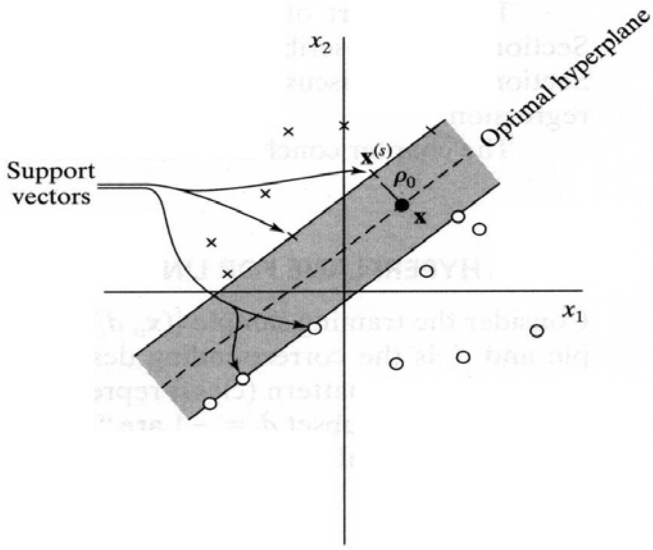

Figure 1 Basic Architecture of SVM

The data point which is very near is called the margin of separation

The main aim of using the SVM is to find the particular hyperplane of which the margin is maximized
Optimal hyperplane
For example, if we are choosing our model from the set of hyperplanes in *Rn*, then we have:

$$f(x; \{w; b\}) = \text{sign}(w \cdot x + b)$$

We can try to learn *f*(*x;* _) by choosing a function that performs well on training data:

$$R_{emp}(\alpha) = \frac{1}{m}\sum_{i=1}^{m} l(f(x_i, \alpha), y_i)$$

Grid Partitioning ID3 Decision Tree
Require: R, a set of attributes.
Require: C, the class attribute.
Require: S, data set of tuples.
1: if R is empty then
2: Return the leaf having the most frequent value in data set S.
3: else if all tuples in S have the same class value then
4: Return a leaf with that specific class value.
5: else
6: Determine attribute A with the highest information gain in S.
7: Partition S in m parts S(a1), ..., S(am) such that a1, ..., am are the different values of A.
8: Return a tree with root A and m branches labeled a1...am, such that branch i contains ID3(R − {A}, C, S (ai)).
9: end if

- Define P1, P2… Pn Parties. (Grid partitioned).
- Each Party contains R set of attributes A1, A2, …., AR.
- C the class attributes contains c class values C1, C2, …., Cc.
- For party Pi where i = 1 to n do
- If R is Empty Then
- Return a leaf node with class value
- Else If all transaction in T(Pi) have the same class Then
- Return a leaf node with the class value
- Else
- Calculate Expected Information classify the given sample for each party Pi individually.
- Calculate Entropy for each attribute (A1, A2, …., AR) of each party Pi.
- Calculate Information Gain for each attribute (A1, A2,…., AR) of each party Pi.
- Calculate Total Information Gain for each attribute of all parties (TotalInformationGain( )).
- ABestAttribute ← MaxInformationGain( )
- Let V1, V2, …., Vm be the value of attributes. ABestAttribute partitioned P1, P2,…., Pn parties into m parties
- P1(V1), P1(V2), …., P1(Vm)
- P2(V1), P2(V2), …., P2(Vm)
- .
- .
- Pn(V1), Pn(V2), …., Pn(Vm)
- Return the Tree whose Root is labelled ABestAttribute and has m edges labelled V1, V2, …., Vm. Such that for every i the edge Vi goes to the Tree
- NPPID3(R – ABestAttribute, C, (P1(Vi), P2(Vi), …., Pn(Vi)))
- End.

IV. Result Analysis

As shown in the below Table is the time complexity comparison between existing id3 based decision tree and vertical partition based decision tree and was found that the proposed algorithm has less complexity when experimented on different values of dataset. The comparison is done between existing ID3 algorithm and the proposed algorithm. The proposed algorithm takes less time when tested on various instances of the dataset as compared to the ID3 algorithm.

number_of_instances	id3_time(ms)	HP_time(ms)
10	15.3	5
20	18	7
30	20.1	8.4
40	25.7	9
50	28	10.2

Table 1: Time Comparison between existing id3 and horizontal id3

As shown in the below Table is the mean absolute error rate of the proposed rate which is less as compared to the existing id3 decision tree. The comparison is done between existing ID3 algorithm and the proposed algorithm. The proposed algorithm has less error rate when tested on various instances of the dataset as compared to the ID3 algorithm.

number_of_instances	ID3_Mean absolute error	HP_Mean absolute error
10	0.2860	0.034
20	0.280	0.083
30	0.310	0.184
40	0.350	0.19
50	0.380	0.2

Table 2: Evaluation of Mean Absolute Error (MAE)

The proposed methodology implemented provides better classification of instances as compared to other existing classification algorithms such as id3, J48, Random Forest and CART. The methodology provides classification accuracy of 96.44%. Also the methodology provides less error rate and high root relative squared error and less time taken to build the decision tree. The result analysis shows the performance of various classification algorithms implemented for the detection of intrusions on the packets. The proposed algorithm implemented here takes less computational time and has less error rate and more number of correctly classified instances as compared to other existing algorithms.

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Time taken to build model
Id3	70.8333 %	29.1667 %	0.4381	0.1944	0.441	51.4706 %	100.965 %	0.01 sec
J48	83.3333 %	16.6667 %	0.71	0.15	0.3249	39.7059 %	74.3898 %	0.02 sec
Random Forest	70.8333 %	29.1667 %	0.4381	0.1826	0.338	48.3333 %	77.3981 %	0.02 sec
SimpleCART	79.1667 %	20.8333 %	0.625	0.1574	0.3368	41.6667 %	77.1238 %	0.02 sec
Proposed	96.4492%	3.5508%	0.3657	0.0861	0.3529	46.2710%	99.6723%	0.007 sec

Table 3 Comparison of various algorithms

V. CONCLUSION

The proposed methodology implemented here for the classification of intrusions using hybrid combinatorial method of clustering and decision tree based classification provides efficient results as compared to the existing techniques implemented for the classification of intrusions in KDDCup99 dataset.

The proposed work implemented here for the detection of intrusions in KDDCup 99 dataset is efficient and provide more accuracy of detection intrusions in the packets. The algorithm provides better classification of intrusion as compared to the existing techniques. The methodology also provides less computational time for the detection as well as provides high rate of correctly classified instances as compared to the existing ID3 algorithm. The experimental results shows that the proposed methodology improvement over other classification algorithms.

REFERENCES

[1]. J. F. Kurose and K. Ross, Computer Networking: A Top-Down Approach Featuring the Internet, 6th ed. Addison-Wesley Longman Publishing Co., Inc., 2012.

[2]. J. Scambray, S. McClure, and G. Kurtz, Hacking Exposed, 6th ed. McGraw-Hill Professional, 2009.

[3]. S. Garfinkel, G. Spafford, and A. Schwartz, Practical Unix & Internet Security, 3rd Edition. O'Reilly Media, Inc., 2003.

[4]. E. Skoudis and L. Zeltser, Malware: Fighting Malicious Code. Prentice Hall PTR, 2003.

[5]. D. Pollino, T. B. Bill Pennington, and H. Dwiyedi, Hacker's Challenge. McGraw Hill Professional, 2006.

[6]. S.E. Smaha, Haystack: An intrusion detection system, in: Proceedings of the IEEE Fourth Aerospace Computer Security Applications Conference, Orlando, FL, 1988, pp. 37–44.

[7]. S. Forrest, S.A. Hofmeyr, A. Somayaji, T.A. Longstaff, A sense of self for unix processes, in: Proceedings of the IEEE Symposium on Research in Security and Privacy, Oakland, CA, USA, 1996, pp. 120–128.

[8]. A. Valdes, K. Skinner, Adaptive model-based monitoring for cyber attack detection, in: Recent Advances in Intrusion Detection Toulouse, France, 2000, pp. 80–92.

[9]. Shon, T., & Moon, J. (2007). A hybrid machine learning approach to network anomaly detection. Information Sciences, 177, 3799–3821.

[10]. G. F. Cretu-Ciocarlie, A. Stavrou, M.E. Locasto, S. Stolfo, Adaptive anomaly detection via self-calibration and dynamic updating. In: RAID '09: Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection, Berlin, Heidelberg, Springer-Verlag (2009) 41-60.

[11]. D. Bolzoni, S. Etalle, and P. Hartel. (2009). Panacea: Automating Attack Classification for Anomaly-Based Network Intrusion Detection Systems. In Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection (RAID '09).

[12]. Robertson, W., Vigna, G., Kruegel, C., Kemmerer, R.: Using generalization and characterization techniques in the anomaly-based detection of web attacks. In: NDSS '06: Proc. 13th ISOC Symposium on Network and Distributed Systems Security. (2006).

[13]. Wang F, Qian Y, Dai Y, Wang Z (2010) A model based on hybrid support vector machine and self-organizing map for anomaly detection. In: International conference on communications and mobile computing, cmc 2010, vol-1. Shenzhen, China, pp 97–101.

[14]. Z. Muda, W. Yassin, M.N. Sulaiman," Intrusion Detection based on K-Means Clustering and OneR Classification", IEEE 2011.

[15]. Mohammadreza Ektefa, Sara Memar," Intrusion Detection Using Data Mining Techniques", IEEE 2010.

[16]. Anup Goyal, Chetan Kumar", GA-NIDS: A Genetic Algorithm based Network Intrusion Detection System", 2005.

[17]. R. Shanmugavadivu, Dr.N.Nagarajan," Network Intrusion Detection System Using Fuzzy Logic", IJCSE 2011.

[18]. Yogita B. Bhavsar, Kalyani C.Waghmare," Intrusion Detection System Using Data Mining Technique: Support Vector Machine", IJETAE 2013.

[19]. S. Devaraju and S. Ramakrishnan," Performance Comparison for Intrusion Detection System Using Neural Network With KDD Dataset", ICTACT 2014.

[20]. Jun Wang, Xu Hong, Rong-rong Ren, Tai-hang Li," A Real-time Intrusion Detection System Based on PSO-SVM", IWISA 2009.