IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-6, Issue-3 E-ISSN: 2347-2693

# A Comparative Review on the Performance of Intrusion Detection Algorithms and Datasets in Networks Using Data Mining Techniques

**Ramakant Soni[1*], Pradeep Singh Shekhawat[2]**

[1]Department of Computer Science, B. K. Birla Institute of Engineering & Technology, RTU, Pilani, Rajasthan, India
[2]Department of Computer Science, B. K. Birla Institute of Engineering & Technology, RTU, Pilani, Rajasthan, India

[*]*Corresponding Author: ramakant.soni1988@gmail.com*

**Available online at: www.ijcseonline.org**

Received: 20/Feb//2018, Revised: 26/Feb2018, Accepted: 19/Mar/2018, Published: 30/Mar/2018

***Abstract***— In today's world where everything relies on the networks, the data in transfer may be susceptible to outside attacks. And these attacks are vulnerable because the data is huge in size and critical or may be confidential in nature. Due to this it becomes the prime activity to protect the information and the system processing this huge amount of information from the unauthorized access and theft. And this makes the role of Intrusion detection system very important as this helps in the protection of Confidentiality and maintenance of the integrity and reliability of the information. A number of methods are present and being used to their limits for the protection. Data mining techniques are used for the purpose of pattern extraction and analysis of the attack patterns helps in developing better system for the network. After the review of a number of data mining algorithms for clustering, classifications and classification via clustering (CvC) the conclusion is that CvC algorithm shows the best performance in intrusion detection. In the review datasets like KDDcup 99, NSL_KDD, GureKDD and Kyoto 2006+ is discussed with their performance and results for analysis.

***Keywords***— *Intrusion, IDS, ID3, C4.5, Classification, Decision Tree, Clustering, Pruning, Classification via Clustering*

## I. INTRODUCTION

In past few years the internet usage has rapidly increased due to emergence of internet based facilities. These facilities has captured the human race in a world of internet technologies and encouraged the massive use of internet. This has resulted in the huge data transfer and information generation. With the times attackers have also become intelligent and developed advanced devices and ways to attack the network and steal the information in an anonymous way.

This reason has encouraged the development of intrusion detection systems which can prevent the attackers from data theft. Intrusion is a set of activities malicious in nature which harms the network and the information in transit and its detection is a process of identification of the patterns and analysis of the extracted information for the purpose of prevention of theft.

There are two categories of intrusion detection. First one is the Anomaly based and another is the Misuse based [17].

### A. *Anomaly based detection:*

It works on the system behavior and whenever it identifies any sort of abnormality it signals the system about it.

### B. *Misuse based detection:*

It signals when it finds the pattern stored in the database matches with the attack pattern.

The Intrusion Detection System (IDS) is classified into three types.

### A. *Host Intrusion Detection System (HIDS):*

This is a single system analysis process for the data packets.

### B. *Network Intrusion Detection System (NIDS):*

It analyzes the whole network and its constituent devices.

### C. *Distributed Intrusion detection System (DIDS):*

It analyzes multiple attacks from different sources.

IDS protects the network from the attackers but attackers keep on finding the new ways to attack which generates the new patterns. That is why the concept of data mining will help in identifying the new attack methods by analysing the previous attack patterns [3]. Existing techniques have been discussed in this paper with their advantages as well as their shortcomings.

For the analysis learning is used. Learning methods are classified into two: Supervised and Unsupervised. Assumption is that any sort of activity: legal or illegal will have entry in the audit logger data [19][26]. Under supervised learning Classification method is used. It constructs the algorithm based on input and output. Pre classified dataset generates the classification model for the detection [5].

Classification method constitutes of two stages:

**A. *Training Stage:***

In the training stage, the training data set is loaded and pre-processed (cleaned & reduced) for generating the classifier using classification algorithm.

**B. *Testing Stage:***

When a data tuple is tested against the generated classifier then the system can decide it to be an attack or a normal process. For the decision, Decision tree is used as a technique for classification. It selects the most promising attributes and then classifies the values in the classes.

In most instances Decision Tree gives the maximum detection rate with respect to other techniques like: Artificial neural network, Biological neural networks and Support Vector Machines [7] with quick response and maximum comparisons too [5].

**Paper Organization:** The paper is organized as follows, Section I contains the Introduction of the Intrusion detection system evolution, Section II contains the related work done in the field of intrusion detection and data mining applications, Section III contains the discussion of Classification algorithms, Section IV contains the discussion of Hybrid Methods, and Section V concludes the paper with review.

## II. RELATED WORK

Over the time the security requirement in the network has increased and in this case data mining technique faces large datasets. From these data sets interesting patters emerge. Pattern extraction involves the usage of different types of methods.

The main aim of these methods is high attack detection rate. In this area lot of research has been done. The work includes the classification algorithms, clustering and the hybrid methods.

The comparative analysis of these methods is helpful to find the best suitable intrusion detection method.

Though many factors are there based on which detection can be done but important are the processing speed, scalability and the dependency [11].

A decision tree algorithm like ID3 works well when single valued attribute is there. For the multivalued attribute creates problem in the classification. Sometimes incorrect classification happens. An improved ID3 algorithm does better analysis for detection [8].

C4.5 algorithm also a decision tree algorithm doesn't handle data sets with unique values but it handles the multi valued attributes. The problem arises due to unbalanced split and larger tree size. The optimization technique helps to overcome this [9].

Pruning increases the detection rate and speed using data sets like KDDcup 99 and NSL_KDD over the C4.5 method [3, 25]. Pruning can be performed as pre and post. Pre pruning do not help as it terminates before the time and similarly post pruning branches are deleted after the complete growth of tree. Optimization of the tree improves the method and the final results by implementing the multi strategy pruning algorithm. [6]

Using Radial based functions with SVM (Support Vector Machine) algorithm the problem of ability of handling heterogeneous datasets is solved [15].

Classification methods are helpful in the detection but the detection rate is not high. To achieve the high rate of detection cascading is used with the classification. Supervised and unsupervised technique is used [16].

High Detection rate can be achieved by cascading two methods: K_Means clustering with C4.5 method [10]

K-Means clustering of Naïve Bayes Classification method uses the cascading techniques [18].

Different Classification algorithms are used together to get high detection rate [13].

The problem of single layer intrusion detection can be solved by Genetic Algorithm. It is a multilayer approach and detects the R2L attack but unable to detect Dos [20].

Genetic Algorithm and SVM is applied together to form the set of optimal features. It increases the detection rate. Genetic Algorithm generates the optimal attribute subset. In KDDcup 99 dataset only 10 attributes form the optimal subset out of 45 which gives the best result in terms of accuracy. When compared to other algorithms it gives the highest detection rate of 97.3% [21].

## III. CLASSIFICATION ALGORITHM

### *A. ID3:*

ID3 stands for Iterative Dichotomiser 3 and it is a decision tree learning algorithm developed by J. Ross Quinlan in year 1980.

It classifies the test cases into the classes. Using greedy approach it selects the attribute with highest information gain and splits on each iteration. Information gain or entropy has to be a high value for splitting and it is not always optimal that is why ID3 is considered as reliable method because it sometimes terminate wit suboptimal results. Some improvements have been done to give desired results [8, 11].

### *B. C4.5:*

C4.5 is the extended algorithm over ID3 by Quinlan. Splitting occur according to gain. And due to high splitting size the time and space complexity rise and this is controlled by pruning in which leaf nodes replaces the branches not satisfying the test [2, 3, 6].

### *C. C5.0:*

Just like C4.5, C5.0 is also a decision tree algorithm. It is an improvement over C4.5 with higher detection rate, greater speed, efficiency and small in size. Boosting helped in improvement but can't help [5] in greater noise scenario so it removes the extra unused attribute using the method called Winnowing [5].

### *D. Random Tree:*

There is an equal probability of sampling all the generated trees. At each level it randomly selects the no. of attributes which are used to make the tree and allows the class probability estimation and operate no pruning step [11].

### *E. Random Forest:*

A group of methods increase the accuracy in classification which is more accurate than the basic one. The main concept is the classifier generation using replacement method. Because each classifier denotes the decision tree, it is hence called random forest. Random attribute selection result in splitting the tree. Major advantage is the scalability and fast learning [11]. This algorithm helps in more accurate detection [24].

### *F. SVM:*

This algorithm works to maximize the class level separation [11] by increasing training set margins
SVM is used for numeric prediction as well as classification.

It has two variations:

a) C- SVM:
b) One-Class SVM

When the learning is supervised, C- SVM is used and in case of unsupervised One- Class SVM is preferred.

In SVM the training set is transformed into the higher dimensional data using technique. This technique is Non-Linear mapping.

Out of newly created data best hyper plane is selected on the basis of high margin distance among the set. Vectors close to the hyper plane are selected to be the whole training set reducing the data set size [15].
SVM is inefficient in dealing with the multiclass datasets due to which detection rate is not much high [5].

Table 1. Classification algorithms on different data sets Comparison

| ***Algorithm Authors, Year*** | ***Dataset*** | ***Results*** |
|---|---|---|
| ID3 Improved<br>(G. Zhai, C. Liu)<br>2010 | Online data | Less time complexity compared to ID3<br>Low False detection rate |
| C4.5 & C4.5 with Pruning<br>(N. G. Relan, D. R. Patil)<br>2015 | KDDCup 99 & NSL_KDD | C4.5 with pruning is more accurate as compared to C4.5 without pruning |
| Random Forest & Random Tree<br>(K. S. Elekar)<br>2015 | KDDCup 99 | Random Tree algorithm detects R2L and U2R attacks.<br>Random Forest algorithm detects DoS and Probe attacks. |
| SVM<br>(M. V. Kotpalliwar, R. Wajgi)<br>2015 | KDDCup 99 | High Validation and Classification Accuracy.<br>High Time complexity |

## IV. HYBRID METHODS

Sometimes the signature learning methods and Anomaly learning methods are inefficient in high detection accuracy [18] and therefore other detection methods like Hybrid methods come in use. Hybrid methods give proper results with higher True detection rate.

### *A. K-Means combined with C4.5*

A Supervised clustering algorithm which works in two phase:

a) Selection
b) Classification

In the first phase closer clusters are found using the Euclidean distance of the data set and in second phase test tuple is classified as normal and attack. It uses the mean to designate Cluster center.

K-Means method when combined with C4.5 method eliminates the problems of Class Dominance with Forced Assignment Problem because in K-Means method each sample must belong to only one cluster but if not then C4.5 does the work by classification on the basis of decision rule [10].

### *B. K-Means combined with Naïve Bayes*

K- Means combined with unsupervised learning method like Naïve Bayes Classification method improves the detection rate. In case of Naïve Bayes classification the different class attribute values are independent of each other.

This method has two stages. In stage I Nature of attack is analyzed and then similar attributes are grouped as pre-classification component. In stage II the output cluster is reclassified into different classes of attack and also not classified data if any in the stage I is classified in this stage.

When compared to simple Naïve Bayes method, this Hybrid method gives more accurate results.

### *C. J48 combined with Random Tree & Random Forest*

Different methods work efficiently for different data and attacks but none of them works efficiently for all kind of attacks in together. Like J48 method efficiently deals with normal attacks, Random Tree method with U2R and R2L attacks whereas Random Forest with Dos and Probe attacks.

Just to overcome the problem of increased time & space complexity due to individual method for different attack set, these methods can be used in combination [13] to deal with the attack set. Combination of two methods out of three detects a category of attacks. As compared to other combination Random Forest and Tree method has higher detection rate for the probe attacks.

Different Hybrid methods with results on datasets are compared in the Table 2.

Table 2. Comparison of different classification algorithm

| Hybrid Methods | Datasets | Results |
|---|---|---|
| K-Means combined with Naïve Bayes<br>( Z. Muda, W. Yassin, M. N. Sulaiman, N. I. Udzir)<br>2011 | KDDCup99 | The detection rate of Probe, Normal, U2R, DoS increases higher than Naïve Bayes method separately. |
| K-Means combined with C4.5<br>( A. P. Muniyandi, R. Rajeswari, R. Rajaram )<br>2012 | KDDCup99 | The True Positive Rate, Precision and F-Measure increases as compared to K-Means and C4.5 separately. |
| J48 combined with Random Tree<br>(K. S. Elekar)<br>2015 | KDDCup99 | The detection rate and false attack detection rate is highly increased for DoS, U2R, R2L, Normal attacks |

## V. CONCLUSION

Various data mining techniques have been discussed for the increase in the rate of detection and the reduction of false detection rate which were implemented in past. Through the time these techniques have evolved better with higher efficiency and accuracy. These improvements and their evolution has been discussed and reviewed in this paper. This includes dataset selection criteria, feature selection factors, clustering and classification methods and hybrid methods.

Classification algorithms are only capable of detecting the intrusion which is known. Among different classification algorithms for intrusion detection, Decision Tree Method has shown the better results with higher detection. C4.5 method combined with pruning has the highest detection rate for NSL_KDD dataset in comparison to detection rate of basic detection tree. The hybrid method K-means with C4.5 methods have been observed to have a higher detection efficiency than the K-means with Naïve Bayes method.

Among the different Decision tree algorithms reviewed for different attacks, C4.5 algorithms detects Normal attacks, Random Tree detects R2L and U2R attacks and Random Forest method detects DoS and Probe attacks with greater efficiency. Similarly The detection rate and false attack detection rate is highly increased for DoS, U2R, R2L and Normal attacks by implementing J48 algorithm with Random Tree.

For the unknown attacks only clustering algorithms fall incapable due to their high false positive results. For it the Hybrid algorithms like CVC (Classification via Clustering) is implemented. It increases the accuracy by reducing the false positive results.

## REFERENCES

[1] W. Pu, W. Jun-qing, *"Intrusion Detection System with the Data Mining Technologies"*, In the Proceedings of the 2011 IEEE International Conference on Communication Software and Networks (ICCSN), China, 2011, ISBN: 978-1-61284-486-2.

[2] S. K. Sahu, S. Sarangi, S. K. Jena, *"A Detail Analysis on Intrusion Detection Datasets",* In the Proceedings of the 2014 IEEE International Advance Computing Conference (IACC), India, 2014, ISBN: 978-1-4799-2572-8.

[3] N. G. Relan D. R. Patil, "*Implementation of Network Intrusion Detection System Using Variant of Decision Tree Algorithm"*, In the Proceedings of the 2015 IEEE International Conference on Nascent Technologies in the Engineering Field (ICNTE),India, 2015, ISBN: 978-1-4799-7263-0.

[4] G. Kayacik, A. N. Zincir-Heywood, M. I. Heywood. *"Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets",* In the Proceedings of the 2005 IEEE Annual conference on privacy, security and trust, Canada, 2005.

[5] M. Kumar, M. Hanumanthappa, T. V. Suresh Kumar. *"Intrusion Detection System Using Decision Tree Algorithm"*, In the Proceedings of the 2012 IEEE International Conference on Communication Technology (ICCT), China, 2012, ISBN: 978-1-4673-2101-3.

[6] Wang, B. Chen, "*Intrusion Detection System Based On Multi-Strategy Pruning Algorithm of the Decision Tree"*, In the Proceedings of the 2013 IEEE International Conference on Grey Systems and Intelligent Services, China, 2013, ISBN: 978-1-4673-5248-2.

[7] N. M. Prajapati, A. Mishra, P. Bhanodia, *"Literature Survey- IDS for Ddos Attacks",* In the Proceedings of the Conference on IT in Business, Industry and Government (CSIBIG), India, 2014, ISBN: 978-1-4799-3064-7.

[8] G. Zhai, C. Liu, *"Research and Improvement on ID3 Algorithm in Intrusion Detection System"*, In the Proceedings of the 2010 IEEE International Conference on Natural Computation (ICNC), China, 2010, ISBN: 978-1-4244-5961-2.

[9] Thakur, N. Markandaiah, D. S. Raj. *"Re- Optimization of ID3 and C4. 5 Decision Tree"*, In the Proceedings of the 2010 IEEE International Conference on Computer and Communication Technology (ICCCT), India, 2010, ISBN: 978-1-4244-9034-9.

[10] A.P. Muniyandi, R. Rajeswari, R. Rajaram, *"Network Anomaly Detection by Cascading K-Means Clustering and C4. 5 Decision Tree Algorithm",* Procedia Engineering, Vol. 30, pp. 174-182, 2012.

[11] P. Aggarwal, S. K. Sharma, *"An Empirical Comparison of Classifiers to Analyze Intrusion Detection"*, In the Proceedings of the 2015 IEEE International Conference on Advanced Computing Communication Technologies (ACCT), India, 2015, ISBN: 978-1-4799-8488-6.

[12] Y. J. Zhao, M. J. Wei, J. Wang *"Realization of Intrusion Detection System Based on the Improved Data Mining Technology"*, In the Proceedings of the 2013 IEEE International Conference on Computer Science Education (ICCSE), Sri Lanka, 2013, ISBN: 978-1-4673-4463-0.

[13] K. S. Elekar, *"Combination of data mining techniques for intrusion detection system",* In the Proceedings of the 2015 IEEE International Conference on Computer, Communication and Control. IEEE, India, 2015, ISBN: 978-1-4799-8164-9.

[14] S. Sahu, B. M. Mehtre, *"Network Intrusion Detection System Using J48 Decision Tree"*, In the Proceedings of the 2015 IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), India, 2015, ISBN: 978-1-4799-8792-4.

[15] M. V. Kotpalliwar, R. Wajgi, *"Classification of Attacks Using Support Vector Machine (SVM) on KDDCUP' 99 IDS Database",* In the Proceedings of the 2015 IEEE International Conference on Communication Systems and Network Technologies (CSNT), India, 2015, ISBN: 978-1-4799-1797-6.

[16] S. Agrawal, J. Agrawal, *"Survey on Anomaly Detection Using Data Mining Techniques"*, Procedia Computer Science, Vol. 60, p.p: 708-713, 2015.

[17] Kruegel, F. Valeur, G. Vigna, *"Intrusion Detection and Correlation: Challenges and Solutions"*, Springer Science and Business Media, Inc. Boston, 2005, ISBN: 978-0-387-23399-4.

[18] Z. Muda, W. Yassin, M. N. Sulaiman, N. I. Udzir, "Intrusion Detection Based on K Means Clustering and Naïve Bayes Classification", In the Proceedings of the 2011 IEEE International Conference on Information Technology in Asia (CITA 11), Malaysia,pp.1-6, 2011, ISBN: 978-1-61284-130-4.

[19] U. Bashir, M. Chachoo, *"Intrusion Detection and Prevention System: Challenges and Opportunities"*, In the Proceedings of the 2014 IEEE International Conference on Computing for Sustainable Global Development (INDIACom), India, 2014, ISBN: 978-93-80544-12-0.

[20] M. Padmadas, N. Krishnan, J. Kanchana, M. Karthikeyan, *"Layered Approach for Intrusion Detection Systems Based Genetic Algorithm*", In the Proceedings of the 2013 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). India, 2013, ISBN: 978-1-4799-1597-2.

[21] M. Aslahi-Shahri, R. Rahmani, M. Chizari, A. Maralani, M. Eslami, M. J. Golkar, A. Ebrahimi, *"A Hybrid Method Consisting of GA and SVM for Intrusion Detection System"*, Neural Computing and Applications, Vol. 27, 2016, p.p. 1669–1676.

[22] S. B. Kotsiantis, *"Decision Trees: A Recent Overview",* Artificial Intelligence Review, Vol. 39, 2013, p.p. 261–283.

[23] Han, M. Kamber, J. Pei, *"Data mining: Concepts and Techniques"*, Elsevier, 2011, ISBN: 978-0-12-381479-1

[24] Zhang, M. Zulkernine, A. Haque, *"Random-Forests-Based Network Intrusion Detection Systems*", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 38, 2008.

[25] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani , "*A Detailed Analysis of the KDD CUP 99 Data Set*", IEEE Symposium on Computational Intelligence for Security and Defense Applications, Canada, 2009, ISSN: 2329-6267.

[26] R. Kumar, *"A Review of Network Intrusion Detection System using Machine Learning Algorithms"*, International Journal of Computer Sciences and Engineering (IJCSE), Vol. 5, Issue-12, p.p. 94-100, 2017

**Authors Profile**

*Mr. Ramakant Soni* completed his Bachelor of Technology in Computer Science from Rajasthan Technical University, Kota, India in Year 2010 and Master of Technology in Multimedia & Data Management from Mewar University, Chittorgarh, India in year 2015. He is currently working as Assistant Professor in Department of Computer Science, B. K. Birla Institute of Engineering & Technology, Pilani. He is a member of Institution of Engineers (India) since 2013. His main research work focuses on Image Processing, Data Analytics, Data Mining and Machine Learning. He has 8 years of Teaching experience and 5 years of Research Experience.

***Mr. Pradeep Singh Shekhawat*** completed his Bachelor of Engineering in Computer Science from Rajasthan University, Jaipur, India in Year 2009 and Master of Technology in Multimedia & Data Management from Mewar University, Chittorgarh, India in year 2015. He is currently working as Assistant Professor in Department of Computer Science, B. K. Birla Institute of Engineering & Technology, Pilani. He is a member of Institution of Engineers (India) since 2013. His main research work focuses on Networking, Cloud computing, Internet of Things and Data analytics. He has 9 years of Teaching experience and 5 years of Research Experience.