INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

AN UPDATED ID3 TECHNIQUE FOR ACCURATE CLASSIFICATION INTRUSION DATA SET

Manu Bijone[1], Jitendra Dangra[2]

[1]M.Tech-Computer Science and Engineering, Lakshmi Narain College of Technology, Indore India,

Email:mbijone@gmail.com

[2]Asst. Prof. - Dept. of IT/CS, Lakshmi Narain College of Technology, Indore India,

Email: jitendra.dangra@gmail.com

Abstract: - Data classification algorithms are very important in real world applications like- intrusion classification, heart disease prediction, cancer prediction etc. This paper presents a novel decision tree based technique for data classification. Basically it is an enhanced variant of ID 3 algorithm. ID3 is a popular and common decision tree based technique for data classification. In this paper, an upgraded version of ID3 is proposed. This version calculates information gain in a different way by giving more weightage to more important attribute instead of an attribute which is having more different values. The fundamentals of data classification are also discussed in brief. The experimental results have proven that the accuracy of the presented method is better.

Keywords: Network Security, Intrusion Detection, Intrusion Prevention, Data Set, Computer Networking.

INTRODUCTION:

A network security system typically relies on layers of protection and consists of multiple components including networking monitoring and security software in addition to hardware and appliances. All components work together to increase the overall security of the computer network. Security of data can be done by a technique called cryptography [16]. Network security covers a variety of computer networks, both public and private, that are used in everyday jobs conducting transactions and communications among businesses, government agencies and individuals. Networks can be private, such as within a company, and others which might be open to public access [16].

Because of large volumes of security audit data as well as complex and dynamic properties of intrusion behaviors, the optimization of the performance of IDS becomes an important open problem that is receiving more and more attention from the research community [17, 18]. Uncertainty to explore if certain algorithms perform better for certain attack classes constitutes the motivation for the reported herein.

The Data mining-based intrusion detection systems (IDSs) have demonstrated high accuracy, also good generalization to novel types of intrusion and robust behavior in a changing environment in recent years. A major problem faced by them is the intensive computation required in the model generation phase.

Network Based IDS

Because they only scrutinize network traffic [1], the NIDS do not benefit from running on the host. They are often run on dedicated machines that observe the network flows sometimes in conjunction with a firewall. That's why the security threats of the clients do not affect the monitors.

One major shortcoming of NIDS is that they are oblivious to local root attacks [17, 18]. The authorized user of the system that attempts to gain additional privileges will not be deleted if attack is performed locally. The authorized user of the system may be able to set up an encrypted channel when accessing the machine remotely.

Host Based IDS

The HIDS have an ideal vantage point [1, 3]. An HIDS runs on the machine it monitors, HIDS can theoretically observe and log any event occurring on the machine [17, 18].

It means that an HIDS can only be trusted up to the point where the system was compromised. As network attacks have increased in number severity over the past few years, intrusion detection system (IDS) is becoming a critical component to secure the network. Because of large volumes of security audit data as well as complex and dynamic properties of intrusion behaviors, the optimization of the performance of IDS becomes an important open problem that is receiving more and more attention from the research community. Uncertainty to explore if certain algorithms perform better for certain attack classes constitutes the motivation for the reported herein.

The Data mining-based intrusion detection systems (IDSs) have demonstrated high accuracy, also good generalization to novel types of intrusion and robust behavior in a changing environment in recent years. A major problem faced by them is the intensive computation required in the model generation phase.

The intrusion detection systems [2, 3, 4] are based on either signature based techniques or the statistical based techniques. The signature based techniques make use of the training data or the signature to detect & prevent the intrusion. Therefore the signature based techniques are not good enough to detect the novel intrusion attacks. Whereas the statistical based techniques have an advantage over the signature based techniques that they can also detect the novel attacks. One most common method for classification is decision tree based classification.

RELATED WORK:

In [6] **Jake Ryan et al** have worked on the concept of the neural network. The neural network performs learning on the basis of the test data and then it performs predictions. It can classify that the behavior of the node as normal or abnormal. **Denning D.E et al** [7, 8] has presented a sequential rule based model for the prediction of abnormal behavior. Sequential rules are based on the sequential data base. The training data is stored in the sequence in which they occur in the sequence data set & then the sequence rule mining algorithm is applied on the training data set to identify the patterns of the normal behavior and the abnormal behavior. The system developed in [9] has more accuracy in identifying whether the records are normal or attack one. **Dewan M et al** [10] proposed an improved version of the self adaptive Bayesian algorithm (ISABA). It is based on the concept of the Bayesian network but accuracy rate is below expectation. **S.Sathyabama et al** [11] presented a method based on the clustering. In this method the similarity based records are stored in the clusters. Also the dissimilar records are called the outliers. For the outliers the alarm is raised & the record is checked for the abnormal behavior. To protect from attack, the paper [19] introduces a Model for Cryptosystem Using Neural Network, which is of high security and low cost. Based on this model separate Encryption and Decryption Algorithm is presented. Also Training Algorithm for Multi-layered Neural Network is provided to hold secure multimedia data. The proposed work finds its application in medical imaging systems, military image database communication and confidential video conferencing, and similar such application [19]. **Amir Azimi Alasti et al** [21] proposed a self-organizing map (SOM) based method for the intrusion detection. The map has been used successfully to classify the data records. In this method the false positive

alerts have been reduced up to a good extent. Wi-Fi technology is vulnerable to many attacks due to the lack of security, limitation of capability, power limitations, resource handling etc. Security is more and more important, and wireless monitoring and shielding are of prime importance for network security [12]. In order to satisfy secure communication between all nodes, this paper proposes Advanced Technique for Monitoring and Shielding in Wi-Fi Technology. Idea proposed in [12] explores various security issues of IEEE 802.11 based wireless network and analyzes numerous problems in implementing the wireless monitoring and shielding system. To protect from attack, the system analyzes wireless network protocols efficiently and flexibly, reveals rich information of the IEEE 802.11 protocol such as traffic distribution and different IP connections, and graphically displays later. **Alan Bivens et al** [13] have proposed a self-organizing map (SOM) based classification technique. The false negative has been reduced in this model. The authors of [5, 14] proposed the ensemble approach. This approach is a fusion of many existing algorithm. The experimental results have shown that it has outperformed many existing techniques. It has also outperformed support vector machine. This paper [15] presents a method which is based on the concept of the dimension reduction. The dimension reduction is achieved by the feature extraction technique of the data mining. **Aly Ei-Senary et al** [20] has proposed a fusion of the apriori and the kuok algorithm. This proposed model also uses the concept of the fuzzy set i.e. partial membership function.

Objective

- The objective is to classify the information of a flow available as normal or attack by an updated decision tree based classifier.
- The accuracy of the proposed decision tree based classifier will be better as compared to that of the existing classification techniques.

PROPOSED METHODOLOGY:

INPUT:

1. TRAINING DATA
2. TESTING DATA

OUTPUT:

1. CLASSIFICATION TREE

PROCEDURE:

1. IF THE INPUT DATA SET IS EMPTY THEN RETURN A SINGLE NODE WITH VALUE "UNSUCCESS".
2. IF EVERY RECORD OF THE INPUT DATA SET CONTAINS SIMILAR VALUE FOR THE TARGET ATTRIBUTE THEN CREATE A NODE CONTAINING THAT VALUE AND RETURN.
3. OTHERWISE IF ALL THE RECORDS OF THE TRAINING DATA SET CONTAINS NO THEN CREATE A NO NODE AND STOP
4. CALCULATE THE MODIFIED INFORMATION GAIN OF ALL THE ATTRIBUTES USING THE FOLLOWING MODIFIED GAIN CALCALATION FORMULAE:

The Gain (D, A) is information gain of example set D on AN attribute A is defined as FOLLOWS:
Modified Gain (D, A) = (Entropy(D) - S ((|Da| / |D|) *Entropy(Da))) * V

Where:
V IS THE WEIGHT ASSOCIATED WITH EACH ATTRIBUTE, ACCORDING TO THE IMPORTANCE IN

RESULT Da= subset of D for which attribute A has value a.
| Da| = TOTAL number of elements in Da

SELECT AN ATTRIBUTE (DISCUSSED IN ATTRIBUTE SELECTION) WITH THE HIGHEST INFORMATION GAIN AND CONSTRUCT A DECISION NODE

5. REPEAT STEP 4 FOR EACH ATTRIBUTE

RESULT

ANALYSIS

The data set used in the experimental study is kdd 99 data set. We have used following attribute for the classification.

We implemented the existing ID3 and the modified Id3 on JAVA platform. The experimental study has proven that the accuracy of the proposed ID3 is better as compared to the existing one.

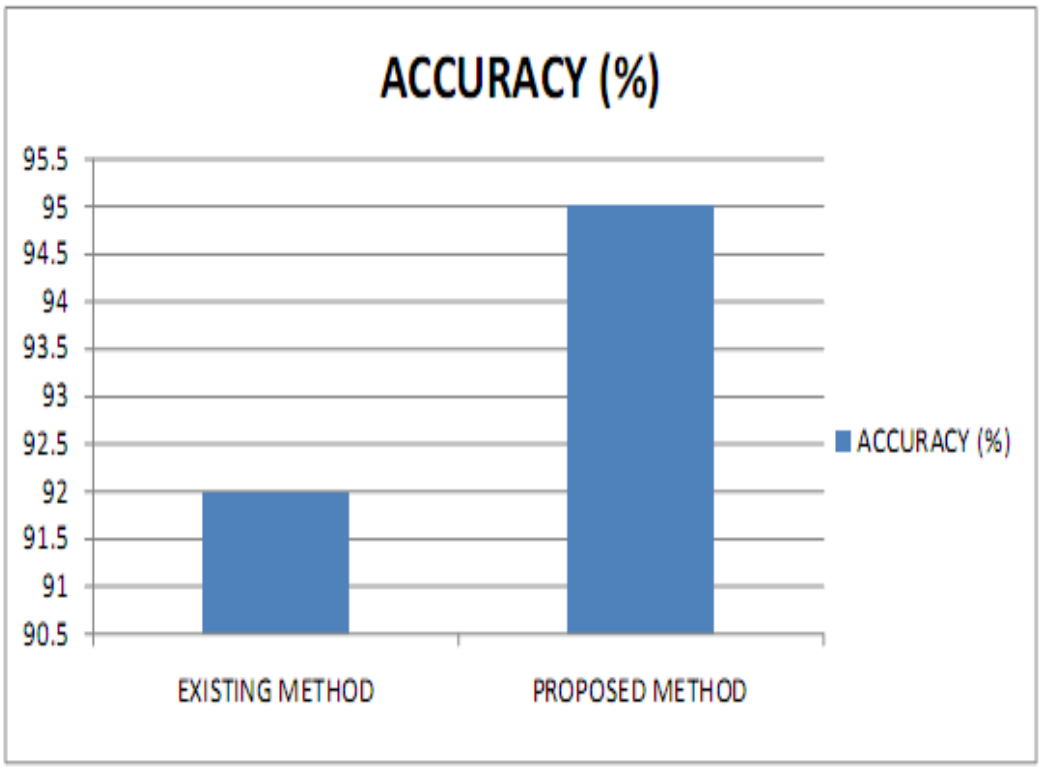

Fig: Proposed technique with respect to existing technique

CONCLUSION:

In this paper, existing approaches are analyzed and problem with network security is identified. This paper focus on why intrusion detection is needed during communication over network? In this paper, an enhanced data classification technique is also proposed to overcome the network security issue. Proposed methodology calculates information gain in a different way by giving more weight-age to more important attribute instead of an attribute

which is having more different values. The accuracy of proposed classifier is better than existing approach.

REFERENCES:

[1] Singh Vijendra. Efficient Clustering For High Dimensional Data: Subspace Based Clustering and Density Based Clustering, Information Technology Journal; 2011, 10(6), pp. 1092-1105.
[2] D Breiman, L., Friedman, J. H., Olshen, R. A., and Stone C. J., "Classification and Regression Trees", Wadsworth International Group. Belmont, CA: The Wadsworth Statistics/Probability Series1984.
[3] Quinlan, J. R., "Induction of Decision Trees". Machine Learning; 1986, pp. 81-106.
[4] Quinlan, J. R. Simplifying "Decision Trees. International Journal of Man-Machine Studies", 1987, 27:pp. 221-234.
[5] Gama, J. and Brazdil, P. "Linear Tree. Intelligent Data Analysis", 1999, 3(1): pp. 1-22.
[6] Langley, P. "Induction of Recursive Bayesian Classifiers". In Brazdil P.B. (ed.), Machine Learning: ECML,1993, pp. 153-164, Springer, Berlin/Heidelberg~lew York/Tokyo.
[7] Witten, I. & Frank, E,"Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.ch. 3,4, pp 45-100.
[8] Yang, Y., Webb, G. "On Why Discretization Works for Naive-Bayes Classifiers", Lecture Notes in Computer Science, 2003, pp. 440 – 452.
[9] H. Zantema and H. L. Bodlaender, "Finding Small Equivalent Decision Trees is Hard", International Journal of Foundations of Computer Science; 2000, 11(2):343-354.
[10] Huang Ming, Niu Wenying and Liang Xu , "An improved Decision Tree classification algorithm based on ID3 and the application in score analysis", Software Technol. Inst., Dalian Jiao Tong Univ., Dalian, China, June 2009.
[11] Chai Rui-min and Wang Miao, "A more efficient classification scheme for ID3", Sch. of Electron. & Inf. Eng., Liaoning Tech. Univ., Huludao, China; 2010, Version1, pp. 329-345.
[12] Shyam Nandan Kumar, "Advanced Technique for Monitoring and Shielding in Wi-Fi Technology, International Journal of Research in Engineering Technology and Management, vol. 2, issue 3, pp. 1-6, 2014.
[13] Chen Jin, Luo De-lin and Mu Fen-xiang," An im pr oved ID3 decision tree algorithm",Sch. of Inf. Sci. & Technol., Xiamen Univ., Xiamen, China, 2009, pp. 127-134.
[14] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann, 2006, ch-3, pp. 102-130.
[15] Ravi Jeet Singh, "A Survey of Modern Classification Techniques", International Journal for Scientific Research & Development, Volume 1 Issue 2 , 2013.pp. 209 – 211.
[16] Shyam Nandan Kumar, "Review on Network Security and Cryptography", International Transaction of Electrical and Computer Engineers System, vol. 3, no. 1, pp.1-11, 2015, doi: 10.12691/iteces-3-1-1.
[17] Manu Bijone, "A Survey on Secure Network: Intrusion Detection & Prevention Approaches", American Journal of Information Systems, vol. 4, no. 3, 2016, doi: 10.12691/ajis-4-3-2.
[18] Manu Bijone, and Jitendra Dangra, "A Survey of Signature Based & Statistical Based Intrusion Detection Techniques", IJSRD - International Journal for Scientific Research & Development, Vol. 4, Issue 08, pp. 583-585, 2016.
[19] Shyam Nandan Kumar, "Technique for Security of Multimedia using Neural Network", International Journal of Research in Engineering Technology and Management, vol. 2, issue 5, pp.1-7, 2014.
[20] Breast Cancer Statistics from Centers for Disease Control and Prevention, http://www.cdc.gov/cancer/breast/statistics/
[21] Iu Yuxun and Xie Niuniu, "Improved ID3 algorithm",Coll. of Inf. Sci. & Eng., Henan Univ. of Technol., Zhengzhou, China, 2010, pp.465-573.