Optimization and Reduction of Time Complexity of Ids Algorithms Using Data Mining Technique

N.K.Barpanda
Reader, Dept.of Electronics, Sambalpur University, Odisha, India.

Abstract- The continuous development and rapid expansion of World Wide Web and local network systems have changed the computing world in the last decade. With the rapid expansion of computer networks during the past decade, security has become a crucial issue for computer systems. It is the IDS which protect to our computer network. Different classification and clustering algorithms have been proposed in recent years for the implementation of intrusion detection systems. In this paper, multiple algorithms are analyzed to find the optimal algorithm. Then we reduce the time complexity of optimal algorithm by eliminating some features without altering the efficiency.

Keywords- *Data Mining, Intrusion Detection, Attribute, Classification And Clustering Algorithm, Time Complexity, Weka.*

I. INTRODUCTION

Intrusion detection System (IDS) is a type of security management system for computers and networks [1]. An intrusion detection system (IDS) inspects all outbound and inbound network action and find out the doubtful patterns that may point to a network or system intrusion or attack from someone trying to crack into or conciliation a system. IDS gathers and observed information from different areas inside a network of systems to find out probable safety breaches, which contain together called intrusions (attacks exterior from the association) and misuse (attacks from inside the association). IDS use susceptibility assessment, it is an expertise which is design and developed to appraise the security of a network [2]. Data mining techniques can be used to detect intrusions. Applications of data mining have presented a collection of research efforts on the use of data mining in computer security. In the context of security of the data we are looking for the information whether an information security breach has been experienced [3]. This data could be collected in the perspective of discovering attacks or intrusions that aim to break the privacy and security of services, information in a system or alternatively, in the context of discovering evidence left in a computer system as part of criminal activity. There are four major categories of networking attacks: Denial of Service, Probing, User to Root and Remote to Local.

Intrusion detection system is the area where data mining concentrate heavily. There are two fold reasons for this first an IDS is very common and very popular and extremely critical activity. Second, large volume of the data on the network is dealing so this is an ideal condition for the data mining to use it. The data mining technology has the enormous benefits in the data extracting attributes and the rule, so it is significant to use data mining methods in the intrusion detection [4]. A significant problem of IDS is how to efficiently divide the normal behavior and the abnormal behavior from a huge number of raw information's attributes, and how to effectively generate automatic intrusion rules following composed raw data of the network. To accomplish this, different data mining methods must be studied, like classification, correlation analysis of data mining methods and so on [4]. The ever rising new intrusion or attacks type poses severe difficulties for their detection. The human labeling of the accessible network audit information instances is generally tedious, expensive as well as time consuming. This paper focuses on study of existing intrusion detection task by using data mining techniques and discussing on various issues in existing IDS based on data mining techniques.

A. WHAT IS IDS?

An intrusion detection system (IDS) inspects all outbound and inbound network action and find out the doubtful patterns that may point to a network or system intrusion or attack from someone trying to crack into or conciliation a system. IDS gathers and observed information from different areas inside a network of systems to find out probable safety breaches, which contain together called intrusions (attacks exterior from the association) and misuse (attacks from inside the association). IDS use susceptibility assessment, it is an expertise which is design and developed to appraise the security of a network [2]. Data mining techniques can be used to detect intrusions. Applications of data mining have presented a collection of research efforts on the use of data mining in computer security. In the context of security of the data we are looking for the information whether an information security breach has been experienced [3]. This data could be collected in the perspective of discovering attacks or intrusions that aim to break the privacy and security of services, information in a system or alternatively, in the context of discovering evidence left in a computer system as part of criminal activity. There are four major categories of networking attacks: Denial of Service, Probing, User to Root and Remote to Local.

Intrusion detection system is the area where data mining concentrate heavily. There are two fold reasons for this first

an IDS is very common and very popular and extremely critical activity. Second, large volume of the data on the network is dealing so this is an ideal condition for the data mining to use it. The data mining technology has the enormous benefits in the data extracting attributes and the rule, so it is significant to use data mining methods in the intrusion detection [4]. A significant problem of IDS is how to efficiently divide the normal behavior and the abnormal behavior from a huge number of raw information's attributes, and how to effectively generate automatic intrusion rules following composed raw data of the network. To accomplish this, different data mining methods must be studied, like classification, correlation analysis of data mining methods and so on [4]. The ever rising new intrusion or attacks type poses severe difficulties for their detection. The human labeling of the accessible network audit information instances is generally tedious, expensive as well as time consuming. This paper focuses on study of existing intrusion detection task by using data mining techniques and discussing on various issues in existing IDS based on data mining techniques.

B. WHAT IS ATTACK?

Attack is an unwanted information or unauthorized access to our network which cause damage to our records.

C. TYPES OF ATTACK

Following are the four major categories of networking attacks:

Denial of Service (DoS): In DoS attack, legitimate networking requests are not served because attacker makes the resources either too busy or full to serve the request. Hence the legitimate user cannot access the services of a machine or network resources. Example: apache, mail bomb, back etc.

Probing (Probe): In probing, attacker scans a machine or a network device for gathering the information about weaknesses or vulnerabilities that can be exploited later to compromise the target system. Example: saint, mscan, nmap *etc.*

User to Root (U2R): In U2R attacks, an authorized user attempt to abuse the vulnerabilities of the system in order to gain privilege of root user for which they are not authorized. Example: perl, xterm, Fd-format etc.

Remote to Local (R2L): In this type of attacks, a remote user tries to gain access as a local user to a local machine by sending packets to a machine over the internet. An external intruder exploits vulnerabilities of the system to access the privileges of a local user. Example: xlock, phf, guest *etc.*

II. LITERATURE SURVEY

Memon V I, Chandel G S [5] presented work is a grouping of three data mining methods to decrease false alarm rate in IDS that is called a hybrid IDS which has k-Means, K-nearest neighbor and Decision Table Majority method for anomaly detection. Presented hybrid IDS evaluated over the KDD-99 Data set; such type of data set is used worldwide for calculating the performance of various IDS. Initially clustering executed via k-Means over KDD99 data sets then executed two-classification method; KNN followed by DTM. The presented system can detect the intrusions and categorize them into four types: Remote to Local (R2L), Denial of Service (DoS), User to Root (U2R) and Probe.

Wankhade K, Patka S, Thool R [6] presents a hybrid data mining approach encompassing feature selection, filtering, clustering, divide and merge and clustering ensemble. An approach for evaluating the number of the cluster centroid and selecting the suitable early cluster centroid is presented.

Dhakar M, Tiwari A [7], in perspective to enhance performance, the work presents a model for IDS. This improved model, named as REP (Reduced Error Pruning) based IDS Model gives output with greater accuracy along with the augmented number of properly classified instances. It uses the two algorithms of classification approaches namely, K2 (BayesNet) and REP (Decision Tree). Here REP provides an effective classification along with the pruning of tree with quick decision learning capability.

Subramanian P.R and Robinson J.W [8] have discussed on network security through Intrusion Detection Systems (IDSs) with data mining approaches. This model uses binary classifier (C4.5) and multi boosting technique. Here binary classifier is used to classify bit by bit transmission of the packet and used for each type of attack to improve the accuracy and to reduce the variance and bias multi boosting technique is used.

Chandolikar N.S and Nandavadekar V.D [9] presented an approach for intrusion detection using J48 decision tree classifier and also compared with some other tree based algorithms in which J48 tree shows the best performance. To evaluate the performance of the algorithm correctly classified instances, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Root Relative Squared Error and Kappa statistics measures are used. Barot V and Toshniwal D [10] presented a hybrid model that ensembles Naive Bayes (statistical) and Decision Table Majority (rule based) approaches. Naive Bayes predicts quickly because of less complex functioning of it and processes training data set only once to store statistics. Decision Table Majority (DTM) is a classifier that matches each of the attribute values all together. This model uses sequential reclassification approach for combining rule base classifier. Here correlation based feature selection (CFS) algorithm is used for attribute selection using BestFirst search. Author used KDDCUP'99 data set for their experiment.

Om H and Kundu A [11] presented a hybrid model that combines K-Means and two classifier methods: K-nearest neighbor and Naive Bayes. This model uses entropy based feature selection method for attribute selection. It applies K-Means clustering algorithm for clustering purpose (used number of clusters five) which is followed by K-nearest neighbor (KNN) and Naïve Bayes classification algorithms for detecting intrusions. The model shows better approach than only K-Means and K- Means, KNN. Author also used the KDD99 cup data set for performing their experiment.

Thakur M R & Sanyal S [12] suggested a multi-dimensional method towards intrusions or attacks detection. Network system usage various parameters like destination and source IP addresses; destination and source ports; outgoing and incoming network traffic information rate and amount of CPU cycles per request are split into numerous dimensions. Observing raw bytes of information corresponding to the values of the network factors, an established function is inferred throughout the training phase for every measurement. This grown-up function takes a measurement value as an input and returns a value that represents the level of anomaly in the system usage relating to that dimension. This mature function is referred as *entity Anomaly pointer*. *Entity Anomaly pointer* recorded for every of the measurement are then used to produced a Universal *Anomaly Pointer*, a function with n variables (n is the number of dimensions) that provides the U*niversal Anomaly Factor*, a pointer of anomaly over the system usage based on all the measurements measured together. The U*niversal Anomaly pointer* inferred throughout the training phase is then used to find out anomaly over the network traffic throughout the detection phase. Network traffic data encountered through the detection phase is fed back to the system to develop the maturity of the *Entity Anomaly Pointers* and hence the U*niversal Anomaly Pointer*.

Pathak V and Ananthanarayana V. S [13] have suggested a multi-threaded K-Means clustering approach. In this approach they have used six threads which run in parallel. Out of which five threads are used to cluster the data and the last sixth thread is
used to take decision classify the data. Out of five threads, each is used to identify particular type of attack and normal or abnormal data. Author used KDD99 training data set for their experiment. Proposed approach i.e. multi-threaded K-Means gives better result in comparison to K-Means.

Wang P and Wang J Q [14] discussed about data mining which is popularly known as an important way to mine useful information from large volumes of data which is noisy, fuzzy, and random. In this, present the whole techniques of the IDS along with data mining method in details. Author mainly discussed about three data mining based approaches: Classification, Association and Sequence rules. Also discussed the system architecture of the IDS.

III. METHODOLOGY

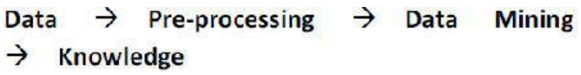
A. DATABASE DESCRIPTION

The proposed system is evaluated using publicly available NSL-KDD intrusion detection dataset which is an enhanced version of the KDD99 intrusion detection dataset. KDD99 dataset is the only well-known and publicly available data set in the area of intrusion detection [14]. It is still widely used in evaluating the performance of proposed intrusion detection algorithms. On the KDD99 intrusion detection dataset 78% of training instances and 75% of test instances are duplicated. Hence the NSL-KDD dataset is generated by removing redundant instances in both the training and test data of the KDD99 intrusion detection dataset [12]. This dataset has 41 features and one class attribute. The training data contains 24 types of attacks and the testing data contains extra 14 types of attacks. The attacks in this dataset are categorized in one of the four attack categories (DoS, Probing, User to Root and Remote to Local attacks).

Though NSL-KDD dataset is enhanced version of the KDD99 dataset we observed two basic problems in this dataset. First as shown in Fig. 2 there are ambiguities in some records of the testing dataset. That is some records have same value for all the 41 features, however they are labeled to different classes (one as normal and the other as attack). The second observation we made is there is a feature called *num_outbounds_cmds* which has a value of zero for all the records in both the training and testing data. This feature will not have any contribution in identifying attacks from normal profiles. Hence we made two improvements in using NSL-KDD dataset: we removed all ambiguous records and the *num_outbounds_cmds* feature from the dataset.

B. WEKA SOFTWARE

Here we have used weka data mining tools for analysis of classification and clustering algorithms. WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. WEKA supports many data mining tasks such as data re-processing, classification, clustering, regression and feature selection to name a few. The workflow of WEKA would be as follows:

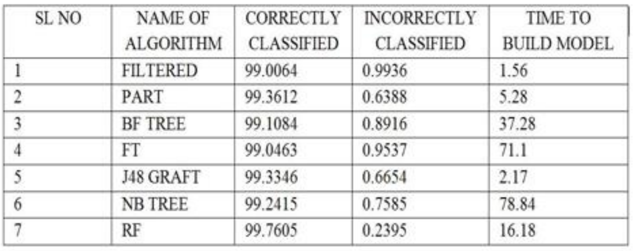


Data → Pre-processing → Data Mining → Knowledge

IV. RESULT ANALYSIS

A. *COMPARISION OF CLASSIFICATION ALGORITHMS:*

We have recorded many classification algorithm's output by using WEKA data mining tools, which is given in form of following table.

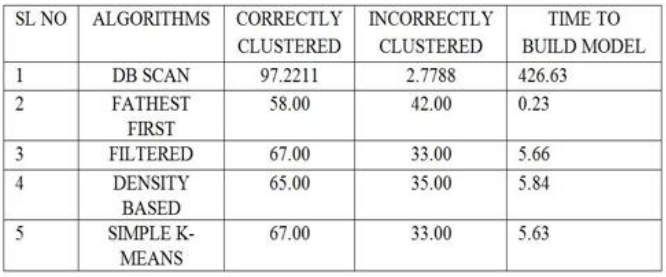


SL NO	NAME OF ALGORITHM	CORRECTLY CLASSIFIED	INCORRECTLY CLASSIFIED	TIME TO BUILD MODEL
1	FILTERED	99.0064	0.9936	1.56
2	PART	99.3612	0.6388	5.28
3	BF TREE	99.1084	0.8916	37.28
4	FT	99.0463	0.9537	71.1
5	J48 GRAFT	99.3346	0.6654	2.17
6	NB TREE	99.2415	0.7585	78.84
7	RF	99.7605	0.2395	16.18

From above we found that RF (Random Forest) algorithm is the best trained algorithm to determine attack having the efficiency 99.76%.When we applied it for test set, it's efficiency came to 98.7% which is the optimal one.

B. *COMPARISION OF CLUSTERING ALGORITHMS.*

Here, again we recorded outputs many clustering algorithms and analyzed through the following tables.



SL NO	ALGORITHMS	CORRECTLY CLUSTERED	INCORRECTLY CLUSTERED	TIME TO BUILD MODEL
1	DB SCAN	97.2211	2.7788	426.63
2	FATHEST FIRST	58.00	42.00	0.23
3	FILTERED	67.00	33.00	5.66
4	DENSITY BASED	65.00	35.00	5.84
5	SIMPLE K-MEANS	67.00	33.00	5.63

In our study we A complete description of all 41 features is available [10], [16]. Instead of describing all the features, here we divide them into three groups and provide descriptions and examples for each group.

Group 1 includes features describing the *commands* used in the connection (instead of the commands themselves). These features describe the aspects of the commands that have a key role in defining the

Group 2 includes features describing the *connection specifications*. This group includes a set of features that present the technical aspects of the connection. Examples of this group include: protocol type, flags, duration, service types, number of data bytes from source to destination, etc..

Group 3 includes features describing the *connections to the same host in last 2 seconds*. Examples of this group are: number of connections having the same destination host and using the same service, % of connections to the current host that have a rejection error, % of different services on the current host, etc..

During inspection of the data it turned out that the values of six features (land, urgent, num_failed_logins, num_shells, is_host_login num_outbound_cmds) were constantly zero over all data records (see [10] for descriptions). Clearly these features could not have any effect on classification and only made it more complicated and time consuming. They were excluded from the data vector. Hence the data vector was a *35 dimensional* vector.

As a result we find same efficiency of optimal algorithm and time required to build the model is reduced.

V. CONCLUSION

It has been discussed about different classification and clustering algorithms to design IDS. At last the optimal algorithms Random Forest and DB Scan are found. Then the time complexity of both the optimal algorithms reduced by eliminating some of features which has no impact to detect the attack. The efficiency of IDS can be improved by applying some hybrid algorithms which is the future work.

It is clearly found that DB scan having the highest efficiency compared to other algorithms which is 97.7%. So if clustering concept is coming then we must choose DB scan for designing IDS.

VI. REFERENCES

[1] D. E. Denning, "An intrusion detection model," *IEEE Transactions on Software Engineering*, vol. 13, no. 2, pp. 222– 232, 1987.

[2] James Cannady, "Artificial neural networks for misuse detection," Proceedings of the 1998 National Information Systems Security Conference (NISSC'98), Arlington, VA, 1998.

[3] J. Ryan, M. Lin, and R. Miikkulainen, "Intrusion Detection with Neural Networks," *AI Approaches to Fraud Detection and Risk Management: Papers from the 1997 AAAI Workshop,* Providence, RI, pp. 72-79, 1997.

[4] K. Fox, R. Henning, J. Reed, and R. Simonian, "A neural network approach towards intrusion detection," Proceedings of 13th National Computer Security Conference, Baltimore, MD, pp. 125-134, 1990.

[5] P. Lichodzijewski, A.N. Zincir Heywood, and M. I. Heywood, "Host-based intrusion detection using self-organizing maps," *Proceedings of the 2002 IEEE World Congress on Computational Intelligence*, Honolulu, HI, pp. 1714-1719, 2002.

[6] H. Debar, M. Becker, and D. Siboni, "A neural network component for an intrusion detection system," Proceedings of 1992 IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, California, pp. 240 – 250, 1992.

[7] Daivid Poole, Alan Makworth, and Randi Goebel, Computational Intelligence, New York: Oxford University Press, 1998.

[8] Sergios Theodorios and Konstantinos Koutroumbas, *Pattern Recognition*, Cambridge: Academic Press, 1999.

[9] Kristopher Kendall, “A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems,” *Masters Thesis, MIT,* 1999.

[10] Srinivas Mukkamala, “Intrusion detection using neural networks and support vector machine,” *Proceedings of the 2002 IEEE International* Honolulu, HI, 2002.

[11] R. Cunningham and R. Lippmann, “Improving intrusion detection performance using keyword selection and neural networks,” *Proceedings of the International Symposium on Recent Advances in Intrusion Detection,* Purdue, IN, 1999.

[12] R.K.Das, N.K.Barpanda and P K.Tripathy,”Network reliability optimization problem of interconnection Network under node edge failure model”, Applied Softcomputing-Elseviewer, vol.12, no.8,2322-2328.

[13] C. Sinclair, L. Pierce, and S. Matzner, “An application of machine learning to network intrusion detection,” Proceedings of 15th Annual Computer Security Applications Conference (ACSAC '99), Phoenix, AZ, pp. 371-377, 1999.

[14] R. A. Kemmerer and G. Vigna, “Intrusion detection: a brief history and overview,” Computer, vol. 35, no. 4, pp. 27–30, 2002.

[15] Piero P. Bonissone, “Soft computing: the convergence of emerging reasoning technologies,” *Soft Computing Journal*, vol.1, no. 1, pp. 6-18, Springer-Verlag 1997.

[16] MIT Lincoln Laboratory, http://www.ll.mit.edu.

[17] Premansu Sekhara Rath, Dr N K Barpanda and S. Panda,” A Soft Computing Approach for Detection and Classification of Attacks”, IARJET INTERNATIONAL JOURNAL,ISSN 2394_1588,VOL 2,Jan_2016

[18] Sanjay Sharma and R. K. Gupta,” Intrusion Detection System: A Review”, International Journal of Security and Its Applications Vol. 9, No. 5 (2015), pp. 69-76. Opinder Singh & Dr. Jatinder Singh,” Comparative study of various Distributed Intrusion Detection Systems for WLAN”, Global Journal of researches in engineering Electrical and electronics engineering Volume 12 Issue 6 Version 1.0 May 2012.

[19] Abebe Tesfahun, D. Lalitha Bhaskari,” Effective Hybrid Intrusion Detection System: A Layered Approach”, J. Computer Network and Information Security, 2015, 3, 35-41 Published Online February 2015 in MECS.

[20] Alireza Shameli-Sendi, Naser Ezzati-jivan, Masoume Jabbarifar, and Michel Dagenais,” Intrusion Response Systems: Survey and Taxonomy”, IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.1, January 2012.

[21] Mohan V. Pawar, Anuradha J,” Network Security and Types of Attacks in Network”, International Conference on Intelligent Computing, Communication & Convergence(ICCC-2014)Conference Organized by Interscience Institute of Management and Technology,Bhubaneswar, Odisha, India

[22] K. Jungwon, J. B. Peter, A. Uwe, G. Julie, T. Gianni and T. Jamie, “Immune System Approaches to Intrusion Detection – A Review”, Natural Computing: an international journal, vol. 6, Issue 4, (2007) December.

[23] E. J. Derrick, R. W. Tibbs and L. L. Reynolds, “Investigating New Approaches to Data Collection, Management and Analysis for Network Intrusion Detection”, ACMSE, Winston-Salem, N. Carolina, USA, (2007) March 23-24, pp. 283-287.

[24] K. K. Bharti, S. Shukla and S. Jain, “Intrusion detection using clustering”, Special Issue of IJCCT, International Conference [ACCTA-2010], vol. 1, Issue 2, (2010) August 3-5, pp. 3-4.

[25] M. Panda and M. R. Patra, “Ensembling Rule Based Classifiers for Detecting Network Intrusions”, IEEE International Conference on Advances in Recent Technologies in Communication and Computing, (2009), pp. 19-22.

[26] V. I. Memon and G. S. Chandel, “A Design and Implementation of New Hybrid System for Anomaly Intrusion Detection System to Improve Efficiency”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, vol. 4, Issue 5, (Version 1), (2014) May, pp. 01-07.

[27] K. Wankhade, S. Patka and R. Thool, “An efficient approach for Intrusion Detection using data mining methods”, International Conference on Advances in Computing, Communications and Informatics (ICACCI), Print ISBN:978-1-4799-2432-5 INSPEC Accession no. 13861274, (2013) August 22-25, pp. 1615-1618.

[28] M. Dhakar and A. Tiwari, “A New Model for Intrusion Detection based on Reduced Error Pruning Technique” International Journal of Computer Network and Information Security, (2013), pp. 51-57.

[29] P. R. Subramanian and J. W. Robinson, “Alert over the attacks of data packet and detect the intruders”, Computing, Electronics and Electrical Technologies (ICCEET), IEEE International Conference on ISBN: 978-1-4673-0211-1, (2012) March 21-22, pp. 1028-1031.

[30] N. S. Chandolikar and V. D. Nandavadekar, “Efficient algorithm for intrusion attack classification by

analyzing KDD Cup 99", Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on ISSN :2151-7681, (2012) September 20-22, pp. 1 - 5.

[31] V. Barot and D. Toshniwal, "A New Data Mining Based Hybrid Network Intrusion Detection Model" IEEE International Conference on Print ISBN: 978-1-4673-2148-8, (2012) July 18-20.

[32] H. Om and A. Kundu, "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system", Recent Advances in Information Technology (RAIT), IEEE International Conference on Print ISBN:978-1-4577-0694-3, (2012) March 15-17, pp. 131-136.

[33] M. R. Thakur and S. Sanyal, "A Multi-Dimensional approach towards Intrusion Detection System" International Journal of Computer Applications, vol. 48, no. 5, (2012) June, pp. 34-41.

[34] V. Pathak and V. S. Ananthanarayana, "A novel Multi-Threaded K-Means clustering approach for intrusion detection" Software Engineering and Service Science (ICSESS), IEEE 3rd International Conference on Print ISBN: 978-1-4673-2007-8, (2012) June 22-24, pp. 757-760.

[35] P. Wang and J. Q. Wang, "Intrusion Detection System with the Data Mining Technologies" IEEE 3rd International Conference on Print ISBN: 978-1-61284-485-5, (2011) May.

[36] Z. Muda, W. Yassin, M. N. Sulaiman and N. I. Udzir, "Intrusion Detection based on K-Means Clustering and Naive Bayes Classification", 7th IEEE International Conference on IT in Asia (CITA), (2011).

[37] D. Md. Farid, N. Harbi and M. Z. Rahman, "Combining naive bayes and decision tree for adaptive intrusion detection", International Journal of Network Security & Its Applications (IJNSA), vol. 2, no. 2, (2010) April.