Increasing efficiency of Intrusion Detection System using Stream Data Mining Classification

Miss. Akanksha M. Shrishrimal[1], Mr.Khwaja Aamer[2], Mr. Syed A.H[3]

[1]*Student,Computer department,AEC Beed,Maharashtra,India*
[2]*Lecturer,Computer department,AEC Beed,Maharashtra,India*
[3]*Lecturer, Computer department,AEC Beed,Maharashtra,India*

ABSTRACT

Data mining is the process of extracting knowledge or information from the previously known and comprehensive datasets for the future decision making. Data mining is the not only withdraw of hidden predictive information but also a strong technology that has great perspective to companies to focus on the most important data in their data repository. Data is growing day by day and this growth has created so many challenges in data mining. The improved technology by World Wide Web is streaming data. The streaming data come into the picture with its challenges and it is known as the data which change with time and update its value. In the sense of security perspective, as the most of the data is streaming in nature, there are so many challenges need to face. The Intrusion Detection System (IDS) work in the supposition of detecting the intruders to protect the respective system. Due to the importance of system's safety measure the research in data stream mining and Intrusion detection system gained high attraction. In this paper, we represent the mechanism to improve the efficiency of the IDS using different streaming data mining classification technique. We apply four selected stream data classification algorithms on NSL-KDD datasets and compared their results. Based on the comparative analysis of their results best method is found out for efficiency improvement of IDS.

Keyword:- *Hoeffding, intrusion detection system,naïve base, stream data classification, streaming data.*

1. INTRODUCTION

Data mining is the process of extracting information or knowledge from previously known and comprehensive datasets. Data mining tools predict future drift and behaviors by allowing businesses to make knowledge-dive decisions. As well as Data mining mechanism can solve business questions which were too time consuming to resolve. By considering the nature we can differentiate the data into two types, that is static data and continuous data. The data does not change with time and they remain static that is called static data, whereas continuous data changes with time and update its value, also called as streaming data. This type of data is impossible to store, hence it required to be analyze in single pass.

In all fields of system and network infrastructure security has become major concern. The main issue is to identify the authorized user and the one who legitimate to the system without abusing their privileges. Intruders that are insider threats as well as outsider threats are rigorous to the system/network. An Intrusion Detection System (IDS) technology used to detect this type of intruders which are very harmful to the system. So the main goal of the IDS is

to protect the system or network from the intruders. IDS system may observe on the behaviors of the system's activities, if they are malicious to the system then it will be detected by the IDS.

There are two types of Intrusion Detection System (IDS) i.e. Network Intrusion Detection System (NIDS) , which resides only on network and observe the malicious traffic passing through the network from source to the destination. And Host Intrusion Detection System (HIDS), which resides only on system infrastructure and observes the inbound and outbound traffic going or coming from or to the system.

Drawback of current IDS:

- Current IDS does not detect the novel intruders, as some of IDS system having predefined signatures so that they does not detect the novel intruders.
- False positive:
- False negative:

Classifiers classify between normal and the suspicious activity to simplify the IDS working. We can group the difficulties into three categories i.e. accuracy, kappa and time. We present the key design elements and categories them into which general issues they focus.

1.1 Stream data mining

Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities. Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches, and sensor data. Data stream mining can be considered a subfield of data mining, machine learning, and knowledge discovery.

Detection Method

Signature-based

Signature-based IDS refers to the detection of attacks by looking for specific patterns, such as byte sequences in network traffic, or known malicious instruction sequences used by malware. This terminology originates from anti-virus software, which refers to these detected patterns as signatures. Although signature-based IDS can easily detect known attacks, it is impossible to detect new attacks, for which no pattern is available.

Anomaly-based

Anomaly-based intrusion detection systems were primarily introduced to detect unknown attacks, in part due to the rapid development of malware. The basic approach is to use machine learning to create a model of trustworthy activity, and then compare new behavior against this model. Although this approach enables the detection of previously unknown attacks, it suffers from false positives:previously unknown legitimate activity may also be classified as malicious.

2. CLASSIFICATION

Classification is the process of classifying data as well as analysis of data. As the streaming data is generated continuously so that it is impossible to store, so it required to analyze them. It is very necessary to classify the streaming data. In this paper we introduce some classification algorithms of streaming data mining such as Naïve Bayes algorithm, Hoeffding Tree algorithm, Accuracy Updated Ensemble algorithm and Accuracy Weighted Ensemble algorithm. The Naïve Bayes algorithm is probability based algorithm whereas Hoeffding Tree algorithm

is the decision tree based algorithm and remaining both algorithms are Accuracy Updated Ensemble and Accuracy Weighted Ensemble are ensemble based algorithms.

2.1 Hoeffding Tree

Hoeffding tree is the decision tree learning algorithm and which is the effective way of classification of data points. Here classification of different problems must be defined.It consist of the test node, root node and leaf nodes where each leaf node denotes prediction of class. In this case major requirement is the classification of the streaming data in a single pass. Hoeffding tree algorithm combines the data into a tree while the model is being built incrementally. These are some advantages of Hoeffding, there are some of disadvantages also that is if ties occur in the dataset, then holding fails to classify the data into tree. Hoeffding bound is given as follows:

$$\epsilon = \sqrt{\frac{R^2 \log(1/\delta)}{2n}}$$

HoeffdingTree Algorithm:

1. Hoeffding is a Tree with a root node
2. for all training data do
3. Sort example into leaf 1 using Hoeffding Tree
4. Update abundant statistics in 1
5. Increment m1
6. if m1 mod M min=0 and e.g. at 1 not of same class then
7. Calculate 1(k1)for each attribute factor
8. Let kp be attribute with highest 1
9. Let Cq be attribute with second highest 1
10. Compute Hoeffding Tree Bound
11. If *Cp* _ *CØ*; and (l (C*a*) - l (C*p*) >_ or _ < _) then
12. Replace *l* with an internal node that splits on *Cp*
13. for all branches do
14. Add a new leaf with initialized sufficient statistics
15. End for
16. End if
17. End if
18. End for

2.2 Naïve Bayes

Naïve Bayes algorithm is the probability based algorithm. It I also called simple Bayes and independent Bayes classifier. It is based on Bayes theorem. Baysian classification provides a useful perspective for understanding and evaluating many learning algorithms. Bayes classifier used for classification of streaming data for finding the Accuracy, statistic kappa, Time these desion parameters.

Assume,

P(H/X)= Posterior Probability

P(H)= Prior Probability

P(X/H)= Posterior Probability of X conditioned on H

P(X)= Prior Probability of X*

Formula for Bayes Theorem is:

$$P(H/X) = \frac{P(X/H)P(H)}{P(X)}$$

2.3 Classifier Ensemble

Accuracy Updated Ensemble(AUE) and Accuracy weighted Ensemble (AWE) algorithms are ensemble based algorithms whereas AUE algorithm is the logical extension as well as it overcomes the drawbacks of weighting function of the AWE algorithm by using method of updating classifier according to the current distribution. We first consider only the current ensemble among all weighted classifiers. Then we use MSE as an entrance for allowing online updatingonly accurate enough classifiers.

Accuracy Updated Ensemble algorithm:

1. C = NULL
2. for all data chunks x_i D do
3. Train classifier $C_i$on x_i;

4. Compute error MSE of $C_i$via cross validation on x_i;
5. Derive weight W for $C_i$using (3);
6. for all classifiers Cr C do
7. Apply C_r on x_i to derive MSE i;
8. Compute weight W_i based on (3);
9. O =n of the top weighted classifiers in C _{C_i};
10. C = C _ {C_i};
11. for all classifiers Ce € O do
12. if $W_e > \frac{1}{MSE_r}$&Ce_C_i then update classifier Ce with x_i

3. SYSTEM DESIGN AND PROPOSED SYSTEM

Here we are designed the overall system architecture for supporting data mining based IDS with its properties which is described throughout this paper. As shown in fig.1 this system architecture consists of network datasets, classifiers, decision parameters and result analysis. In our architecture network dataset is provided to four different classifiers that is Naïve Bayes ,Hoeffding Tree, Accuracy updated ensemble and accuracy weighted ensemble in the supposition of the druthers. Again this data are classified in term of three decision parameter that is accuracy, kappa and time. To obtain performance of classifier, mean values are evaluated according to their decision parameters. Then according to the comparative analysis of results, best fitted classifier is obtained among these classifiers.

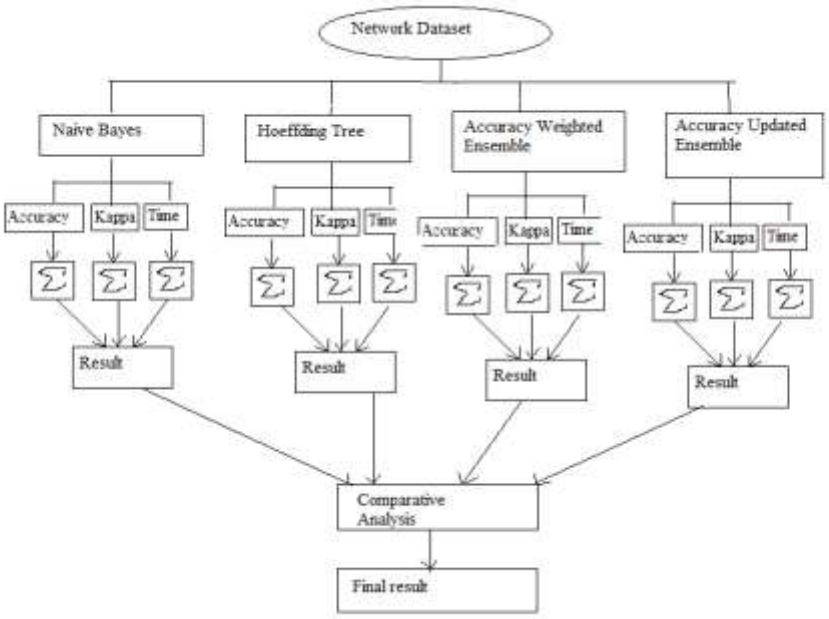

Fig. 1 System Architecture

This architecture is capable of supporting not only analyzing and deciding best classifier but also improving the efficiency of the intrusion detection system. Proposed system is used not only to prevent intruders but also intended improve Intrusion detection efficiency. In traditional IDS intruders are classified with the help of confusion matrix by characterization in true negative and true positive, but it having some drawback like false-*+ positive and false negative. Again current IDS does not detect novel intruders, there are some predefined signatures in IDS , so as the signatures are predefined they fail to detect the novel intruders.

4. EXPERIMENTAL SETUP

Massive Online Analysis (MOA) tool is an open source software platform which is only used for implementing different algorithms. This tool is somehow related to WEKA as well as it is also written in Java language. The MOA tool is used for running different experiments for online learning from evolving massive data streams. It provides framework for implementing experiments in data stream mining.

In these paper, there are various classification method used in MOA like Hoeffding Tree, Naïve Bayes, Enemble algorithms etc.

Beginning with MOA tool for obtaining the best results, certain tasks are carried out, MOA tool provides graphical user interface.

4.1 Dataset used

Dataset having all the information which were collected during a survey and that needs to be analyzed. Here we used NSL-KDD .

Following table shows the detail description of the dataset which used for the experiment.

Table 1- Dataset used

Dataset	Instances

Test+	22544
Train+	25192
Train+_21	11850
Train+_20%	125972

5. RESULT AND DISCUSSION

We performed our experiments on MOA tool with selected classifiers. Following section shows the results of evaluation using MOA in graphical form.

Chart 1 shows that KDD Test+ is selected for evaluation ofresults. Comparison of parameters shows that Accuracyweighted ensemble classifier gives 95.30 % accuracy butrequires 3.76 sec of time. Naive bayes is giving 87.40 %accuracy & taking less time.

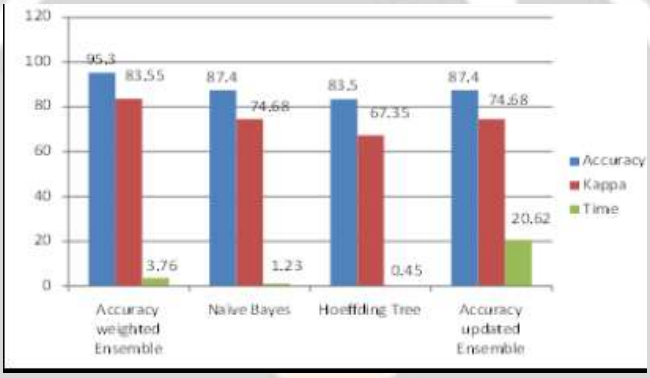

Chart 1: Dataset: KDD Test+

Chart 2 shows that KDD Test-21 is selected for evaluation of results. Comparison of parameters shows that Accuracy weighted ensemble classifier gives 98.60 % accuracy but requires 7.35 sec of time. Naive bayes is giving 95.00 % accuracy& taking less time.

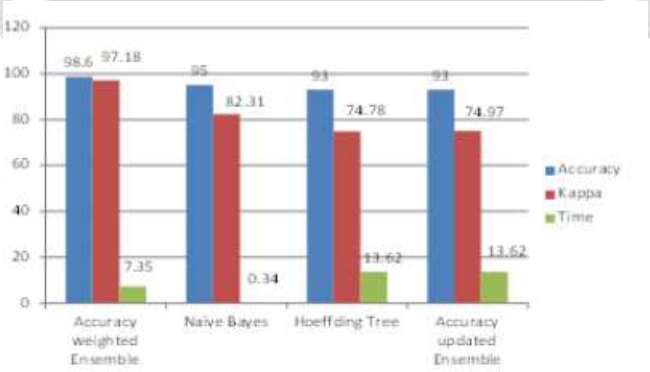

Chart 2: KDD Test-21

Figure 3.shows that KDD Train+20Percent is selected forevaluation of results. Comparison of parameters shows that Accuracy weighted ensemble classifier and Naive bayesboth gives same accuracy i.e. 98.60 % and they also takes same time i.e. 7.35 sec.

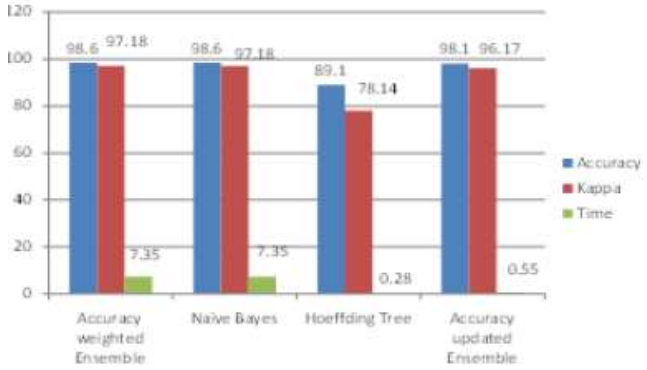

Chart 3: KDD Train+20Percent

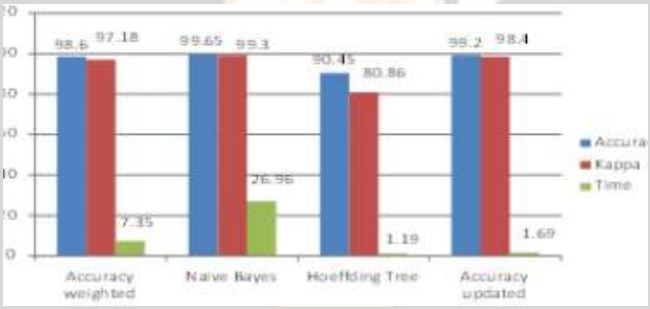

Chart 4: Dataset: KDD Train+

Figure 5.shows that KDD Train+ is selected for evaluation of results. Comparison of parameters shows that Naïve bayes classifier gives 99.65 % accuracy aboubut requires 26.96 sec of time. Accuracy weighted ensemble is giving 98.60 % accuracy& taking less time.

6. CONCLUSION

In this paper, we discuss about how to improve the efficiency of Intrusion Detection System by using classification techniques as well as four classifiers i.e. Naïve Bayes classifier, Hoeffding tree classifier, Accuracy Updated Ensemble classifier and Accuracy Weighted Ensemble classifier.Results generated from analysis of these classifiers, weobtained that Naive Bayes and Hoeffding Tree classifierhand out best results than Accuracy Updated Ensemble andAccuracy Weighted Ensemble Classifier. By observations ofboth, the best classifier Naive Bayes has more accuracy, butit takes more time whereas Hoeffding tree classifier givesaccuracy nearest to the Naive Bayes classifier and it takesless time than naive Bayes classifier.

7. ACKNOWLWDGEMENT

We are thankful to faculty of Department of Computer Science, DrBabasahebAmbedkarMarathwada University, Aurangabad for their support.The product of this research paper would not be possiblewithout all of them.

8. REFERENCES

[1] Jiawei Han, MichelineKamber, Jian Pei, "Data MiningConcepts andTechniques", 3rd edition, Morgan Kaufmann, sss2011. (1st ed., 2000-2001) (2nd ed., 2006)

[2] Babcock, M. Datar, and R. Motwani ,"Load Shedding Techniques for Data Stream Systems" (shortpaper), Proc. of the 2003 Workshopon Management and Processing of Data Streams, June 2003I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[3] Bifet, Albert. "Mining Big Data in Real Time",Informatica37,pp:15-20, 2013.

[4] Babcock, S. Babu, M. Datar, R. Motwani, and J.Widom,"Modelsand issues in data stream systems", Proceedings of PODS, 2002.

[5] H. Kargupta, R. Bhargava, K. Liu, M. Powers, P.Blair, S. Bushra, J.Dull, K. Sarkar, M. Klein, M. Vasa,and D. Handy, "VEDAS: AMobile and Distributed DataStream Mining System for Real-TimeVehicle Monitoring", Proceeding of SIAM International Conference on Data Mining, 2004.

[6] N. Tatbul, U. Cetintemel, S. Zdonik, M. Cherniack,M.Stonebraker"Load Shedding on Data Streams", Proceedings of theWorkshop on Management and Processing of Data Streams, SanDiego, CA, USA, June8, 2003.

[7] G. Dong, J. Han, L.V.S. Lakshmanan, J. Pei, H.Wang and P.S.Yu,"Online mining of changes from data streams: Research problemsand preliminary results", Proceedings of the 2003 ACM SIGMODWorkshop on Management and Processing of Data Streams.Incooperation with the 2003 ACM-SIGMOD International Conferenceon Management of Data, San Diego, CA,June 8, 2003.

[8] Gaber, M, M., Krishnaswamy, S., and Zaslavsky,A., "On-boardMining of Data Streams in Sensor Networks", Accepted as a chapterin the forthcoming book Advanced Methods of Knowledge Discoveryfrom Complex Data, (Eds.) SanghamitraBadhyopadhyay, UjjwalMaulik, Lawrence Holder and Diane Cook, Springer Verlag.

[9] Manish Kumar, Dr. M. Hanumanthapaa, "Intrusion Detection Systemusing Stream Data Mining and Drift Detection Method",4thlCCCNT -2013 July 4-6,2013, Tiruchengode,India.

[10] R. Heady, G. Luger, A. Maccabe, and M. Servilla, "Thearchitecture of a network level intrusion detection system", Technicalreport, Computer Science Department, University of New Mexico,ugust 1990.

[11] Aggarwal, J. Han, J. Wang, and P. S. Yu," On DemandClassification of Data Streams", Proc. 2004 Int.

[12] S. Muthukrishnan, "Data streams: algorithms and applications",
[13] Proceedings of the fourteenth annual ACM-SIAM symposium on
[14] discrete algorithms. (2003).

[15] Shabiashabirkhan,M.A.Peer, S.M.K.Quadri, "Comparative Study ofStreaming Data Mining techniques", International conference on
[16] computing for sustainable Global Development (INDIA.com).2014.

[17] Albert Bifet , Geoff Holmes, Richard Kirkby , Bernhard Pfahringer,"MOA: Massive Online Analysis" Journal of Machine LearningResearch 11 (2010) 1601-1604.

[18] G. Cormode, S. Muthukrishnan,"What's hot and what's not: trackingmost frequent items dynamically",PODS 2003: 296-306

[19] Gaber, M, M., Zaslavsky, A., and Krishnaswamy,S., A Cost-EfficientModel for Ubiquitous Data Stream Mining", the Tenth InternationalConference on Information Processing and Management ofUncertainty in Knowledge-Based Systems, Perugia Italy, July 4-9.

[20] Gaber, M, M., Zaslavsky, A., and Krishnaswamy,S.," Towards anAdaptive Approach for Mining Data Streams in ResourceConstrained Environments", the Proceedings of Sixth InternationalConference on Data Warehousing and Knowledge Discovery {Industry Track (DaWak 2004), Zaragoza, Spain, 30 August { 3September, Lecture Notes in Computer Science (LNCS),SpringerVerlag.

[21] http://www.iscx.ca/NSL-KDD/