Efficient Intrusion Detection System using Stream Data Mining Classification Technique

Ketan Sanjay Desale
Assistant Professor
DYPSOEA,
Pune, Maharahtra
desaleketan.10@gmail.com

Chandrakant Namdev Kumathekar
Dept. of Computer Engineering,
DYPSOEA,
Pune, Maharahtra
ckumthekar777@gmail.com

Arjun Pramod Chavan
Dept. of Computer Engineering,
DYPSOEA,
Pune, Maharahtra
Arjun.chavan1010@gmail.com

***Abstract*— Recent emerging growth of data created so many challenges in data mining. Data mining is the process of extracting valid, previously known & comprehensive datasets for the future decision making. As the improved technology by World Wide Web the streaming data come into picture with its challenges. The data which change with time & update its value is known as streaming data. As the most of the data is streaming in nature, there are so many challenges need to face in the sense of security perspective. Intrusion Detection System (IDS) works in the supposition of detecting the intruders to protect the respective system. The research in data stream mining & Intrusion detection system gained high attraction due to the importance of system's safety measure. Algorithms, systems & frameworks that address security challenges have been developed over the past years. In this paper, we present the mechanism to improve the efficiency of the IDS using streaming data mining technique. We apply four selected stream data classification algorithms on NSL-KDD datasets and compare their results. Based on the comparative analysis of their results best method is found out for efficiency improvement of IDS.**

Keywords- Hoeffding, intrusion detecting system, naive bayes, stream data classification, streaming data.

I. INTRODUCTION

Data mining is the withdrawal of hidden predictive information or knowledge from large database. It is strong and new technology has great prospective to companies focus on the most important data in their data repository. Data mining tools predict future drift and behaviors by allowing businesses to make knowledge-dive decisions [1]. Data mining mechanism can answer business or profession questions which were traditionally too time consuming to resolve. In traditional data set, data does not change with time and they are static in nature, whereas streaming data generated continuously. Continuous data, i.e. streaming data is impossible to store, hence it required to be analyzed in single pass [2] [3] [4]. Streaming data can be network data which consists of inbound and outbound traffic of the network.

Security had become major concern in all fields of network & system infrastructure. The main issue is to identify the authorized user & the one who is legitimate to access the system without abusing their privileges. Insider threats as well as outsider threats are rigorous to the network/system, called as intruders. An Intrusion Detection System (IDS) is critical technology to detect such intruders who are harmful to the system. Main goal of the IDS is to protect the system & network from the intruders. IDS keep track of behavior of the activities; if they are malicious to the system then it'll be automatically detected by the IDS [9]. There are two types of IDS i.e. NIDS & HIDS. Network Intrusion Detection System (NIDS) resides on network & observes the malicious traffic passing through the network whereas Host Intrusion Detection System (HIDS) resides on the system & observes inbound & outbound traffic going or coming from/to the system; the example of the HIDS will be firewall [10].

Drawback of current IDS:

- Current IDS does not detect the novel intruders: As some of the IDS work on the signature based technology, there are some predefined signatures in IDS, but as the signatures are predefined they fail to detect the novel intruders.
- False Positive: It occurs when normal is wrongly classified as intruder.
- False Negative: It occurs when an intruder is wrongly classified as normal.

Data mining can significantly improve the intrusion detection. Classifiers classify between normal & the suspicious activity to simplify the IDS working. We can group the difficulties in to three general categories: accuracy, kappa & time. We present the key design elements & categories them into which general issues they focus.

II. CLASSIFICATION

Classification is a process of analysis of data. Classification is a benefit to the streaming data. It is very necessary to classify the streaming data. Classification has more applications that are artifice detection, retailing, analytical modeling, manufacturing and medical analysis [11]. In this chapter we will introduce some classification algorithm of streaming data mining such as Naive Bayes algorithm, Hoeffding Tree algorithm, Accuracy Updated Ensemble algorithm and Accuracy Weighted Ensemble algorithm. The Naive Bayes algorithm is probability based, Hoeffding Tree algorithm is Decision Tree based algorithm

978-1-4799-6892-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCUBEA.2015.98

IEEE computer society

and remaining two algorithms that accuracy updated ensembles and accuracy weighted ensembles are ensemble based algorithms [12].

A. Hoeffding

Hoeffding algorithm is a decision tree learning algorithm and an effective way of classification of data points. In Hoeffding algorithm, classification of different problems must be defined. Classification problems area collection of training examples of the form (p, q), where p is a vector of s attributed and q is a discrete class label and aim is to produce a model of the form q=f(a). Such that the function f(a) predicts the classes j for future examples I with higher accuracy [13].

It consists of the test node, root node and the leaf nodes, where each leaf node denotes prediction of class. In our case major requirement is the classification of streaming data in a single pass. Data streams are to be read in less amount of time for classification. Hoeffding algorithm combines the data into a tree while the model is being built incrementally, even at that time we can use to classify data. These are some advantages discussed, but there are also some disadvantages of this algorithm. If ties occur in the dataset, then holding fails to classify data into the tree. Hoeffding bound ε is given as follows

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

Hoeffding Tree Algorithm:

1. Hoeffding is a Tree with a root node]
2. for all training data do
3. Sort example into leaf l using Hoeffding Tree
4. Update abundant statistics in l
5. Increment ml
6. if ml mod M min = 0 and e.g. at l not of same Class then
7. Calculate l (kl) for each attribute factor
8. Let kp be attribute with highest l
9. Let Cq be attribute with second-highest l
10. Compute hoeffding bound ε
11. If $Cp \neq C\emptyset$; and (l (C*a*) - l (C*p*) >ε or ε < τ) 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

B. Naive Bayes

Naive Bayes classifier is a probabilistic classifier. It is also called as simple Bayes and Independence Bayes classifier. It is based on Bayes Theorem. It has ability to solve diagnostic and predictive problems. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. Bayes Classifier used for classification of Streaming data for finding the Accuracy, statistic kappa, Time these decision 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:

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

C. Classifier Ensemble

Accuracy updated ensemble (AUE) is the logical extension to the accuracy weighted ensemble (AWE) algorithm. It overcomes the drawback of weighting function of the accuracy weighted ensemble by using method of updating classifier according to the current distribution. To achieve this, we renovate only particular algorithmic classifiers. We first considered only current ensemble among all- the top weighted classifiers. Then we use MSE as an entrance for allowing online updating only accurate enough classifiers. Therefore classifiers can enter the ensemble, but will not be updated Algorithm.

Accuracy updated algorithm

Input: D: data stream
n: number of ensemble
Output: O: ensemble of n online classifiers with updated weights

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 C_e € O do
12. if $W_e > \frac{1}{MSE_r}$ & Ce≠Ci then update classifier C_e with x_i

III. PROPOSED SYSTEM

The overall system architecture is designed to support a data mining-based IDS with the properties described throughout this paper. As shown in Figure 1, the architecture consists of network dataset, classifiers, decision parameters, and result analysis. This architecture is capable of supporting

not only analyzing & deciding best classifier, but also improving the efficiency of intrusion detection system. In traditional intrusion detection system, intruders are classified with help of confusion matrix by the characterization in true positive and true negative, but as confusion matrix has drawbacks like false positive and false negative it may possibly for intruders to penetrate into the system.

Drawbacks in the current IDS system cause emerging need of pre-processed data, but as we are dealing with network dataset it is not possible to pre-process data in a single pass.

Proposed System is only intended to improve Intrusion detection efficiency, not to prevent intruders. In our architecture, network dataset is provided to four different classifiers i.e. Naive Bayes, hoeffding tree, accuracy weighted ensemble and accuracy ensemble in the supposition of druthers. Further, the data are classified in terms of decision parameter accuracy, kappa & time. To obtain performance of classifier, mean values are evaluated according to their decision parameters. According to the Comparative analysis of results, best fitted classifier is obtained among all.

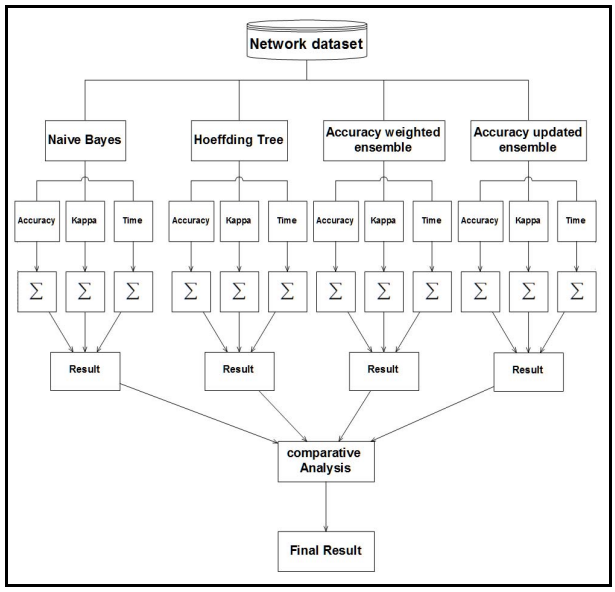

Figure 1. System Architecture

IV. Experimental Setup

Massive Online Analysis (MOA) is an open source framework software environment for implementing different algorithms. MOA tool is used for running different experiments for online learning from evolving massive data streams. MOA software is somehow related to WEKA and it is also written in Java language [13] [14]. The goal of Massive Online Analysis (MOA) tool provides framework for running experiments in data stream mining are as follows:

- Mainly used for storable setting, for massive data streams, for repeatable experiments.
- A set of existing algorithm and easily extensible framework for new data streams from different data sources & several evaluation methods.

Beginning with MOA tool for evaluation of results, certain tasks are carried out, MOA tool provides graphical user interface. There are various streaming classification method used in MOA like Hoeffding (decision tree) Tree, Naive Bayes algorithm, ensemble algorithms etc.

A. *Dataset Used*

Datasets consist of all the information collected during a survey that needs to be analysed. Here we had used network dataset i.e. NSL-KDD, which consists of elected records of the complete KDD Cup 99 dataset [18].

Following table 1 show the detail description of dataset which used for experiment.

TABLE I. Datasets And Their Instances

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

V. Result And Discussion

We evaluated the effectiveness of the proposed technique by performing experiments on MOA tool with selected classifiers. We performed our experiments on a system having configuration of 3GHz Pentium processor, 2GB RAM and 360GB Hard Drive. Following section shows the results of evaluation using MOA in graphical form.

Figure 2. shows that KDD Test+ is selected for evaluation of results. Comparison of parameters shows that Accuracy weighted ensemble classifier gives 95.30 % accuracy but requires 3.76 sec of time. Naive bayes is giving 87.40 % accuracy & taking less time.

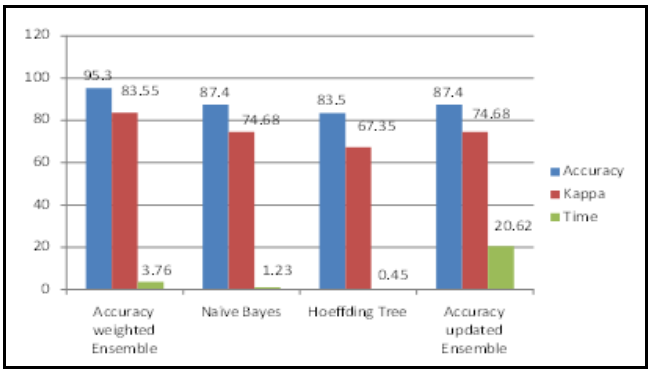

Figure 2. Dataset: KDD Test+

Figure 3. 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.

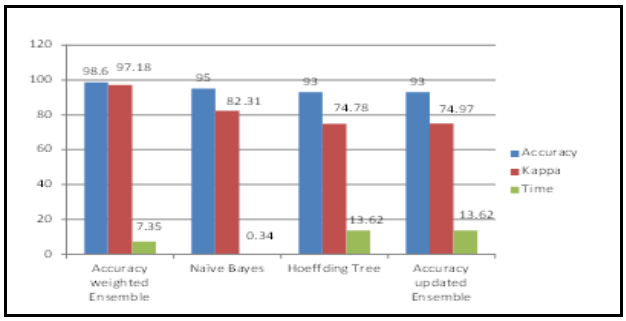

Figure 3. KDD Test-21

Figure 4. shows that KDD Train+20Percent is selected for evaluation of results. Comparison of parameters shows that Accuracy weighted ensemble classifier and Naive bayes both gives same accuracy i.e. 98.60 % and they also takes same time i.e. 7.35 sec.

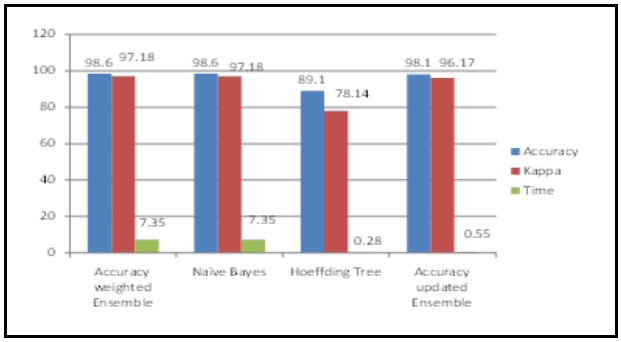

Figure 4. KDD Train+20Percent

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

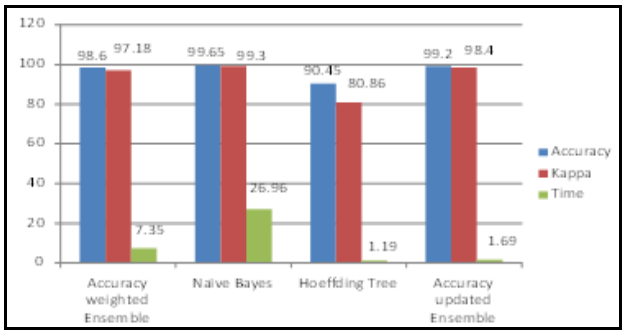

Figure 5. Dataset: KDD Train+

A. *Future work*

From the obtained experimental results we can conclude that accuracy weighted ensemble gives higher accuracy than other classifiers but takes little bit more time whereas hoeffding tree gives less accuracy in less time.

In applications where accuracy has major concern, we can use accuracy weighted ensemble & in the applications where results are required in less time we can use hoeffding tree. As higher accuracy is required in less time we can use AWE+Hoeffding tree as an ensemble.

VI. CONCLUSION

In this paper, we discuss about classification techniques and improving the performance of Intrusion detection system using four classifiers i.e. Naive Bayes classifier, Hoeffding tree classifier, Accuracy Updated Ensemble classifier and Accuracy Weighted Ensemble Classifier. At the time of literature survey, we found that the Naive Bayes is probability based classifier Hoeffding Tree is decision tree based classifier and the remaining two, Accuracy Updated Ensemble and Accuracy Weighted Ensemble are ensemble based algorithms.

Results generated from analysis of these classifiers, we obtained that Naive Bayes and Hoeffding Tree classifier hand out best results than Accuracy Updated Ensemble and Accuracy Weighted Ensemble Classifier. By observations of both, the best classifier Naive Bayes has more accuracy, but it takes more time whereas Hoeffding tree classifier gives accuracy nearest to the Naive Bayes classifier and it takes less time than naive Bayes classifier.

Network dataset applied to best fitted classifier to improve the performance of Intrusion Detection System (IDS).

ACKNOWLEDGMENT

We are thankful to faculty of Department of Computer Science, Savitribai Phule Pune University for their support. The product of this research paper would not be possible without all of them.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining: Concepts and Techniques”, 3rd edition, Morgan Kaufmann, 2011. (1st ed., 2000-2001) (2nd ed., 2006)

[2] B. Babcock, M. Datar, and R. Motwani ,”Load Shedding Techniques for Data Stream Systems” (shortpaper), Proc. of the 2003 Workshop on 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] B. Babcock, S. Babu, M. Datar, R. Motwani, and J.Widom,” Models and 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: A Mobile and Distributed DataStream Mining System for Real-Time

Vehicle 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 the Workshop on Management and Processing of Data Streams, San Diego, 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 problems and preliminary results", Proceedings of the 2003 ACM SIGMOD Workshop on Management and Processing of Data Streams. In cooperation with the 2003 ACM-SIGMOD International Conference on Management of Data, San Diego, CA,June 8, 2003.

[8] Gaber, M, M., Krishnaswamy, S., and Zaslavsky,A., "On-board Mining of Data Streams in Sensor Networks", Accepted as a chapter in the forthcoming book Advanced Methods of Knowledge Discovery from Complex Data, (Eds.) Sanghamitra Badhyopadhyay, Ujjwal Maulik, Lawrence Holder and Diane Cook, Springer Verlag.

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

[10] [10] R. Heady, G. Luger, A. Maccabe, and M. Servilla, "The architecture of a network level intrusion detection system", Technical report, Computer Science Department, University of New Mexico, August 1990.

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

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

[13] Shabiashabir khan,M.A.Peer, S.M.K.Quadri, "Comparative Study of Streaming Data Mining techniques", International conference on computing for sustainable Global Development (INDIA.com).2014.

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

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

[16] Gaber, M, M., Zaslavsky, A., and Krishnaswamy,S., A Cost-Efficient Model for Ubiquitous Data Stream Mining", the Tenth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia Italy, July 4-9.

[17] Gaber, M, M., Zaslavsky, A., and Krishnaswamy,S.," Towards an Adaptive Approach for Mining Data Streams in Resource Constrained Environments", the Proceedings of Sixth International Conference on Data Warehousing and Knowledge Discovery { Industry Track (DaWak 2004), Zaragoza, Spain, 30 August { 3 September, Lecture Notes in Computer Science (LNCS),Springer Verlag.

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