A Case Study: Stream Data Mining Classification

Ketan Sanjay Desale
ME scholar
Department of Computer Engineering
DYPSOET, SPPU

Roshani Ade
Assistant Professor
Department of Computer Engineering
DYPSOET, SPPU

ABSTRACT
Continuous and unending growth of data created so many challenges in data mining task. Data mining is extraction meaningful information i.e. knowledge from large datasets for the future decision making. The data which is continuously generating with changing values is known as streaming data. We face many problems with streaming data as we are unable to store it and process it. Network data is one of the best examples of streaming data. Intrusion Detection System (IDS) used to detect the malicious user to protect the network. System's safety in a network is a prime important factor. In this paper, we present comprehensive approach to improve performance of IDS by applying some classification techniques with streaming dataset. For the experiment purpose we created our own network dataset which shows significant accuracy in results after applying classifiers.

Keywords
Data mining, Naive Bayes, Hoeffding tree, Intrusion Detection System (IDS)

1. INTRODUCTION
Data mining is technique of analyzing data from different sources and summarizing it into fruitful information. Data mining software is one of a number of analytical tools for inspecting data. It allows users to rehash data from many different aspects, classify it, and outline the relationships determine [1, 2]. Technically, data mining is the step of finding interconnection or patterns among a huge number of fields in relational databases. Data mining techniques can take the benefits of mechanization on current software and hardware platforms, and can be carried out on new systems as existing platforms are raised and new products developed. The data which generated continuously and rapidly is called streaming data [3]. The process of taking knowledge from this continuous and rapid data is called streaming data mining. Streaming data require interpret in one pass because it's impervious to store. Streaming data can be networked data it subsists of inbound and outbound traffic [4]. Nowadays use of internet services are increases such as internet banking online shopping and many other. This internet services required security and privacy.

On the other hand, our computers are under attacks and vulnerable to many threats. There are many tools available for attacking and intruding networks. An intrusion can be defined as any set of actions that threaten the security requirements such as integrity, confidentiality, availability of a computer/network resource [5]. Example of intruder is denial of service (DOS), worms and virus and many more. Intrusion detection system (IDS) is technology use to detect intruders which are harmful to the system. Main goal of intrusion detection system is protect the system and network from malicious activity and intruders. There are two types of IDS i.e. NIDS and HIDS [6]. 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 [7].

2. LITERATURE SURVEY
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][12]. In this chapter we will introduce some classification algorithm of streaming data mining such as Naïve Bayes algorithm and Hoeffding Tree algorithm. The Naive Bayes algorithm is conditional probability based, Hoeffding Tree algorithm is Decision Tree based algorithm. In our previous, work we take the four classifiers which are naive byes, hoeffding tree, accuracy updated ensemble and accuracy weighted ensemble. From the results we obtain the best two classifiers which are naive byes and hoeffding tree classifier.

2.1 Hoeffding Tree
Hoeffding algorithm is a decision tree learning algorithm and an effective way of classification of data points. In Hoeffding tree 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[8]. 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.

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 _

11. If Cp _ CØ; and (l (Ca) - l (Cp) >_ 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 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 [9]. 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.

Let us 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) is the logicalextension 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 [10]. 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 xi D do

3. train classifier Ci on xi;

4. compute error MSE of Ci via cross validation on xi;

5. derive weight W for Ci using (3);

6. for all classifiers Cr C do

7. apply Cr on xi to derive MSE i;

8. compute weight Wi based on (3);

9. O =n of the top weighted classifiers in C U {Ci};

10. C = C U {Ci};

11. for all classifiers Ce € O do

12. if Ce ≠ Ci then update classifier Ce with xi

3. PROPOSED SYSTEM

The total system architecture is designed to support a data mining-based approach 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 analyzing & deciding best classifier for streaming data. In thisarchitecture, 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.

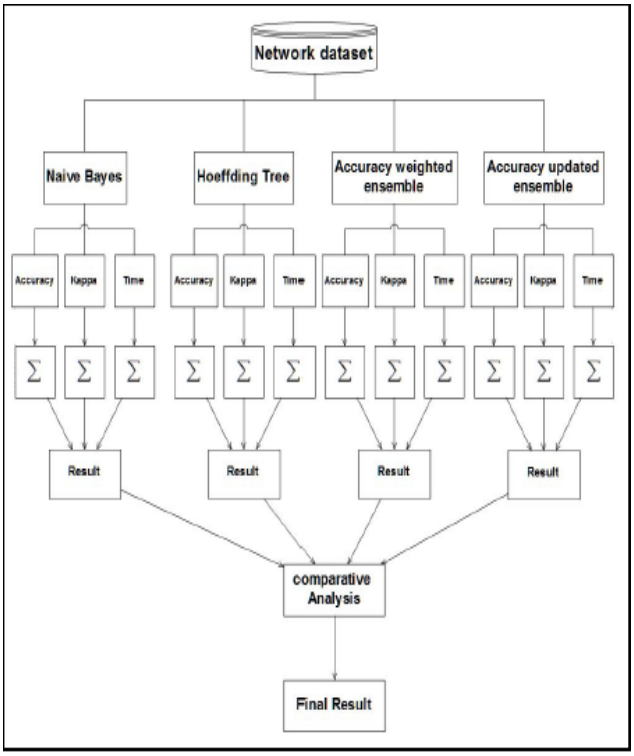

Fig 1 : System Architecture

4. 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]. 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, Naïve Bayes algorithm, ensemble algorithms etc.

4.1 Dataset Used

Dataset is a collection of set of information which is comprised of separate elements, but can be manipulated as a unit by a computer. Here we collected a network dataset on college network. A Network dataset is designed to support network analysis. It includes lines representing the path of flow in the network, increased with other features, topology, and attributes that model network-relevant properties such as impedance and capacity of flow. Network datasets can represent transportation networks.

For creating network dataset Wireshark Network Analyzer tool is used for capturing the packets. From that captured packets we restricted the packets of specific protocols like TCP, UDP, DHCP, ICMP and ARP. We focused on attributes like time, source address, destination address, protocol type, length, source port, destination port. Then we apply these datasets to two classifiers which are Naïve Byes Classifier and Hoeffding Tree classifier. We collect this dataset in specific time periods. We were collecting our data set over three weeks on timely basis. There are 3478 instances in our dataset. This dataset used for classifying normal and anomaly data and improving the performance of IDS.

5. EXPERIMENTAL RESULTS

We developed an application in Java language with user interface. It is use for running different experiments for learning from evolving abundant data streams. The goal of application is to provide framework for running experiments in data streams mining. Before beginning application for evaluation of result certain steps are carried out.

1. Click configure button for configuration.

2. Choose evaluate prequential

3. Select classifier (Hoeffding Tree / Naive Bayes)

4. Browse .arff file from system

5. Set instances as available

6. Click on RUN button.

So accordingly, evaluation of results will been generated and accuracy and time parameters has been considered for measuring the performance.

5.1 Accuracy

It means that how much our system is accurate enough to classify between normal and anomalous behavior. It is calculated as,

$$A = \frac{TN + TP}{TP + TN + FP + FN} \quad (1)$$

Where,

- TP – is the True Positives which mean positive cases are correctly identified.
- TN – is the True Negatives which mean positive cases are incorrectly identified.
- FP – is the False Positives which mean negative cases are incorrectly identified as positive.
- FN – is the False Negatives which mean positive cases are incorrectly identified as negative.

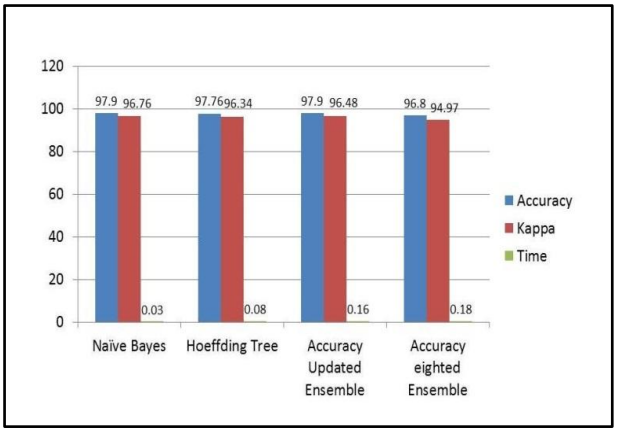

Fig. 1: Experimental results

Figure 1 shows the result of classifiers on collage network dataset. From the graph it clear that Naïve Bayes classifier gives 97.9 % accuracy. Also Hoeffding Tree is giving 96.76 % accuracy but taking more time than Naïve Bayes. In ensemble approach Accuracy Updated Ensemble gives 97.9 % accuracy in much more time. Similarly, Accuracy eighted Ensemble gives 96.8 % accuracy by taking highest time amongst four classifiers.

6. TIME TO BUILD

It is the time required to build the model. It is calculated inseconds.Experimental results show that Naive Bayes is taking much less time as compared to Hoeffding tree. Here Naive Bayes is taking 0.03 sec time which is very less as compared to Hoeffding Tree 0.08 sec. ensemble approach takes more time as compared to Naive Bayes and Hoeffding Tree algorithm.

7. CONCLUSION AND FUTURE SCOPE

In this paper, we have used classification techniques to improve the performance of intrusion detection system using two classifiers i.e. Naive Bayes classifier and Hoeffding tree classifier. At the time of literature survey, we studied four classifiers and found that the Naïve Bayes is conditional probability based classifier and Hoeffding Tree is decision tree based classifier and the remaining two i.e. Accuracy Updated Ensemble and Accuracy Weighted Ensemble are ensemble based algorithms. Result analysis on readymade dataset using these classifiers, we obtained that Naïve Bayes and Hoeffding Tree classifier hand out best results than Accuracy Updated Ensemble and Accuracy Weighted Ensemble classifier.

In this paper we carried out experiment on our own created network dataset. The result analysis shows that classifier Naïve Bayes has more accuracy, and it takes less time whereas Hoeffding tree classifier gives accuracy nearer to the Naive Bayes classifier but it takes more time than it.

8. REFERENCES

[1] Shabiashabir khan, M.A.Peer, S.M.K.Quadri, "Comparative Study of Streaming Data Mining

techniques", International conference on computing for sustainable Global Development, 2014.

[2] C. Aggarwal, J. Han, J. Wang, P. S. Yu, "for Clustering Evolving DataStreams", Proc. 2003 Int. Conf. on Very Large DataBases, Berlin, Germany, Sept. 2003.

[3] Gaber, M, M., Zaslavsky, A., and Krishnaswamy,S.," Towards an Adaptive Approach for Mining DataStreams in Resource Constrained Environments", theProceedings of Sixth International Conference on DataWarehousing and Knowledge Discovery IndustryTrack (DaWak 2004), Zaragoza, Spain, 30 August 3September, Lecture Notes in Computer Science (LNCS),Springer Verlag.

[4] P. Domingos and G. Hulten, "General Method for Scaling Up Machine Learning Algorithms and its Application to Clustering", Proceedings of theEighteenth International Conference on MachineLearning, 2001, Williamstown, MA, Morgan Kaufmann.

[5] K. S. Desale, Rohani Ade, "Genetic Algorithm based Feature Selection Approach for Effective Intrusion Detection System", 2015 International Conference on Computer Communication and Informatics (ICCCI - 2015), Jan. 08 10, 2015, Coimbatore, INDIA

[6] G. Dong, J. Han, L.V.S. Lakshmanan, J. Pei, H.Wang and P.S. Yu, "Online mining of changes from datastreams: Research problems and preliminary results", Proceedings of the 2003 ACM SIGMOD Workshop on Management and Processing of Data Streams. In cooperation with the 2003 ACMSIGMOD International Conference on Management of Data, San Diego, CA, June 8, 2003.

[7] Theodoros Lappas and Konstantinos Pelechrinis, "Data Mining Techniques for (Network) Intrusion Detection Systems".

[8] S. Muthukrishnan, "streams: algorithms and applications", Proceedings of the fourteenth annual ACMSIAM symposium on discrete algorithms. (2003).

[9] Muamer N. Mohammada, Norrozila Sulaimana, Osama Abdulkarim Muhsinb, "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment", Procedia Computer Science 3, 2011, pp. 1237 – 1242.

[10] Muamer N. Mohammada, Norrozila Sulaimana, Osama Abdulkarim Muhsinb, "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment", Procedia Computer Science 3, 2011, pp. 1237 – 1242.

[11] C. Aggarwal, J. Han, J. Wang, and P. S. Yu," A Framework for Projected Clustering of High Dimensional Data Streams", Proc. 2004 Int. Conf. onVery Large Data Bases, Toronto, Canada, 2004.

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

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