International Journal of Engineering &Technology, 7 (1.7) (2018) 54-58

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Detection of novel attacks by anomaly intrusion detection system using classifiers

P. M. Abhinaya *, V. Nivethitha

Assistant Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu – 600062
**Corresponding author E-mail: abhinaya@veltech.edu.in*

Abstract

Nowadays analyzing unsuspicious network traffic has become a necessity to protect organizations from intruders. Really it is a big challenge to accurately identify threats due to the high volume of network traffic. In the existing system, to detect whether network traffic is normal or abnormal we need lots of information about the network. When lot of information is involved in the identification process the relationship between different attributes and the important attributes consider for classification plays an important role in the accuracy. Information gain selection process is used to provide a rank for features. Based on the rank, the most contributed features in the network is found and used to improve the detection rate based on the features selection. In this project, the performance of Lazy and Bayesian classifiers is analysed. In lazy classifier comes there are some algorithms namely, IBK and Kstar. Bayesian classifier comes there are some algorithms namely, Bayes Net, and Naïve Bayes. The performances of Bayesian and lazy classifiers are analysed by applying various performance metrics to identify the best classifier. It is observed that, the efficiency of lazy classifier is better as compared to that of Bayesian classifier.

Keywords: *Information Gain Selection; Kstar; Bayesian; Bayes Net; Classification; IBK; Naïve Bayes; Lazy.*

1. Introduction

Attacker activities are still now rapidly increase using an intrusion detection system detects the intruder on the network. It can be classified into Signature Based Intrusion Detection System, Anomaly Based Intrusion Detection, Wireless Intrusion Detection System (WIDS), Host Based Intrusion Detection System (HIDS), Network Intrusion Detection System (NIDS), and Network Behavior analysis (NBA). NIDS monitoring all devices incoming and outgoing traffic in the network. It is used to monitor all the passing traffic on the subnet and passing traffic is compare to the library of known attacks. Based on known attacks identified behavior of network and inform to the administrator for protecting network. HIDS work on the networks but separately run on host or device. It analyzing both incoming and outgoing packets from the device or in suspicious traffic is detected send alert message to the administrator. It is collecting a snapshot of an existing system file and check for the pervious system file snapshots. If system files are altered or deleted. send notification message to the administrator. NBA determines network traffic to identify attacks that generate unusual traffic flows, such as DDos attack and some form of malware, virus etc. WIPS is used to detect the abnormal traffic in wireless network with support of wireless network protocol. In Signature based Intrusion Detection Systems it maintains database of previous attack signatures and recognized system vulnerabilities. An intrusion or attack is indicated by the storing of Intrusion Detection Systems. Each intrusion leaves signatures behind failed attempt to run an application, nature of data packets, login failures, accessing for files and folders etc. These signatures, called footprints can be used to recognize and avoid similar attacks in future. Based on these detection system identify intrusion attempts. The drawbacks of this detection Systems are signature database must be regularly restructured and sustained. This Detection Systems may be failed to identify a unique attack. An Anomaly Based Intrusion Detection Systems [7] references a baseline or learned pattern of normal system activity to identify active intrusion attempts. Nowadays, many computing infrastructures find it very difficult to prevent unauthorized access and attacks. We have to transfer packets from source to destination as data sharing is one of the key functionality of networking .while transmitting the data in lot of attacks are available to modified the original data and transfer altered data to destination. In some cases attacker send irrelevant request to system which may overwhelm the system and damage the system behavior. To find the behavior of system is complex and gathered lot of information about the network traffic. Information based on network traffic can be classified into three types 1) some features contribute to predict the behavior of network 2) some features do not give any impact 3) some feature gives confusion for predicting the behavior of the network. So collect the contributed feature of network traffic to easily find out the behavior of the network.

2. WEKA

WEKA is a machine learning tool, thereby forming a basis of data mining. It was developed in 1997 by professionals of University of Waikato, New Zealand. Such accumulation of data mining algorithms was issued under the General Public License. WEKA [1] executes calculations for information preprocessing, data preprocessing, association rules, visualization, regression, clustering. For such calculations, it comprises of 49 data preprocessing, 3 association rules and nearby 76 order classifications. The entire

Copyright © 2018 P. M. Abhinaya, V. Nivethitha. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

calculation is performed with the utilization of a Graphical User Interface (GUI) known by explorer that helps in investigating situations from information contained in the dataset. It consists of modules like Knowledge Flow, which is a java based interface and is useful for running machine learning tests, experimenter which is used for making analysis and organizing diverse systems for testing. In addition, various dataset formats like .csv, .arff, .data etc are supported by WEKA in order to extract the relevant data from the crude data.

3. Methodology

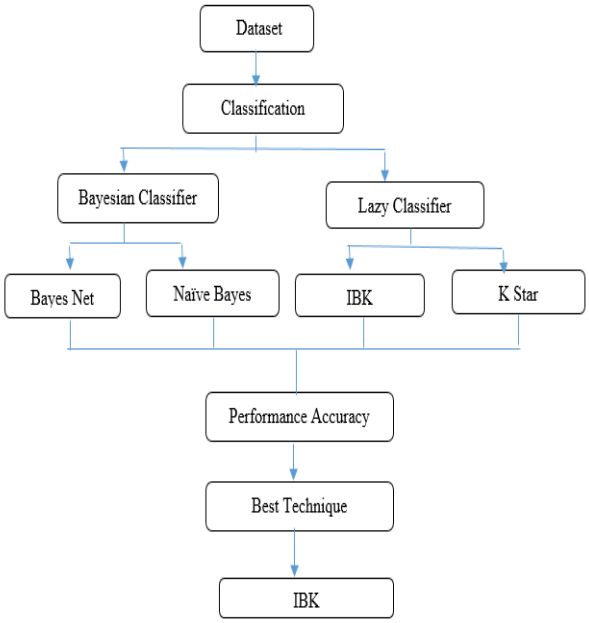


Fig. 1: Types of Classifiers.

Fig.1 describes the overview of working process and gives input to dataset then select classifier either Bayesian or lazy. Calculate accuracy for four classifier and compare to give best classification technique.

3.1. Dataset

There are important issues in the dataset that showed in statistical analysis which highly affects the performance of the systems, and outcomes in a very poor estimation of anomaly detection approaches. NSL-KDD [6] is describe, which consists of selected records of the complete KDD data set. It contains 33,800 record and 41 attributes. The advantages of NSL-KDD dataset are

i) It has better reduction because no duplicate record in the test set.
ii) Classifier will not produce any biased result because no redundant records in the train set.
iii) Classifier will not produce any biased result because no redundant records in the train set.

3.2. Features selection

Feature selection is also called data dimension reduction in predictive analytics; it refers to the process of identifying the few most important variables or parameters which help in predicting the results. Feature selection should be one of the main concerns for a Data Scientist. Accuracy and generalization power can be leveraged by a correct feature selection based on entropy and information gain. Increasing interpretability of the model. The main advantage for feature selection is applied for dataset to reduce the training time and evaluating time.
The important for features selection:

i) It enables the machine learning algorithm to train faster.
ii) It reduces the complexity of a model and makes it easier to interpret.
iii) It improves the accuracy of a model if the right subset is chosen and reduces over fitting.

WEKA supporting many feature selection techniques that help to outcomes of features selection

3.2. Information gain attributes evaluation

The feature selection is classified into Search Method and Attribute Evaluator. The search method is the trial or navigation of different combinations of attributes in the dataset to get the chosen features. The attribute evaluator is the method by which each attribute in the dataset, (column or feature) is examined in the context of the output variable. Some Attribute Evaluator techniques required the use of specific Search Methods. For example as shown Fig.2, the Correlation Attribute Evaluation technique used in the next section can be used with a Ranker Search Method that estimates each attribute and displays the results in a rank order.

$$H(\text{Category}) = -\text{sum}(Pi * \log_2(Pi)) \quad (1)$$

$$\text{InfoGain}(\text{Category}, \text{Attribute}) = H(\text{Category}) - H(\text{Category} \mid \text{Attribute}) \quad (2)$$

Where,
Pi is the probability of the Category i in the dataset.
Entropy basically measures the degree of "impurity". The closest to 0 it is, the less impurity there is in your dataset.

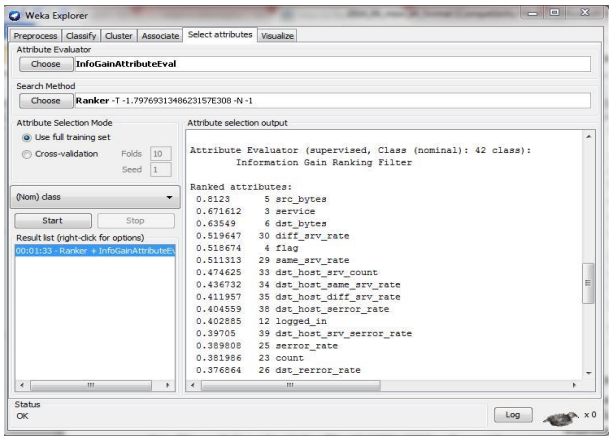


Fig. 2: Attributes in the Dataset.

4. Classification

In this paper, the performance of Bayesian (Bayes Net, and Naïve Bayes) and Lazy classifiers (IBK and Kstar) are analyzed.

4.1. Bayesian classifier and bayes net

Bayesian networks are powerful probabilistic in nature and their use for grouping has become important. Bayesian algorithms [8] predict the class based on the probability of that class. A Bayesian network is a graphical model probability relationship amid a set of variable features. Bayes Net otherwise called as Bayesian networks. The Bayes Net is constructed with the help of minimal attributes. Bayes Nets or Bayesian networks [8] are graphical demonstration for probabilistic relationships amid a set of random variables. A finite set Var= {Var1…Varn} of discrete random variables where the values of each variable Vari may be from a finite set represented by Val (Vari).

$$P(Var_1 \ldots\ldots Var_n) = \prod_i (P(Var_i \mid Pa(var_i))) \quad (3)$$

4.2. Naïve bayes

The probabilistic Naïve Bayes classifier is implementing for classification. Naïve Bayes Simple uses the normal distribution method to model numeric attributes. The Naive Bayes algorithm is fully centred on conditional probabilities. Naïve Bayes is executed using Bayes' Theorem [14] and that is the procedure that calculates the probability by totalling the occurrence of values and combinations of values in the historical data. Bayes Theorem calculates the probability of an event occurring given the probability of another event that has already occurred.

$$P(hyp|d)=(P(d|hypo) * P(hyp))/P(d) \quad (4)$$

Where,
P (hyp|d) is the posterior probability, which gives the probability of hypothesis h given the data d.
P (d|hyp) gives the probability of data d given that the hypothesis hyp was true.
P (hyp) is the prior probability of h, which gives the probability of hypothesis hyp being true.
P (d) gives the probability of the data.

4.3. Lazy classifier and IBK

Lazy Classifier stored all the training instances and does no exact work until classification time. The main advantage for this method [8] is gained in employing that the target function will be evaluating locally such as in the k-nearest neighbor algorithm. Multiple problems are solved simultaneously by lazy learning systems due to local approximation of objective function. IBK is also called k-nearest-neighbor classifier. The number of nearest neighbors can be specified openly in the object editor or by automatic evaluation using leave-one-out cross-validation focus to an upper limit. A kind of different search algorithms [7] can be used to increase speed up the task of searching the nearest neighbors. The algorithm is discussed below

K- Nearest neighbour algorithm
Training
Build the set of training examples *D*.
Classification
Given a query instance x_q to be classified,
Let $x_1 .. x_k$ denote the *k* instances from *D* that are nearest to x_q
Return
$F(x_q) = \arg\max \sum \delta(v, f(x_i))$

K- Nearest neighbour algorithm
Training
Build the set of training examples *D*.
Classification
Given a query instance x_q to be classified,
Let $x_1 .. x_k$ denote the *k* instances from *D* that are nearest to x_q
Return
$F(x_q) = \arg\max \sum \delta(v, f(x_i))$

4.4. KSTAR

The K* algorithm is utilized for cluster analysis targeting at the partition of n observation into k clusters in which each observation belonging to the cluster with the nearest mean. We can describe K* algorithm is an instance based learner which is used for entropy as a distance measure

$$K^*(x_i, y) = -\ln P^*(x_i, y) \quad (5)$$

Where P* means probability of all transformational paths from instance y to x. It can be worthwhile to understand the probability that y will arrive at x via a random walk in feature space.

5. Experimental results

The yield dataset is analyzed with various classification algorithms such as BayesNet, NaïveBayes, IBK, Kstar.

5.1. Correctly and incorrectly instance classification

It can be concluded that accuracy of classifications algorithms like BayesNet, NaiveBayes, IBK and Kstar for the correctly classified instances is more as compared to the incorrectly classified instances. Fig. 3 shows the bar graph of the correctly and incorrectly classified instances of the taken algorithms.

Table 1: Correctly and Incorrectly Instance Classification among Various Algorithm

Algorithms	Correctly Instance Classification	Incorrectly Instance Classification
BayesNet	97.40	2.8
NaiveBayes	2.8	4.43
IBK	99.34	0.65
Kstar	99	1

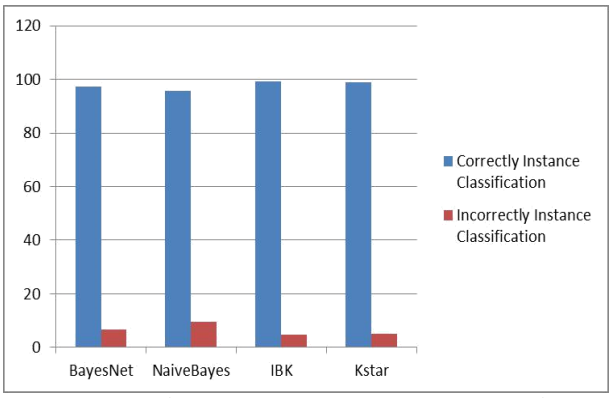


Fig. 3: Bar Graph of Correctly and Incorrectly Instance Classification of Algorithms.

5.2. Errors

a) Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

Mean Absolute Error Computes accuracy of attributes and average magnitude of errors. Actually it is calculate the average of absolute values between predicted observation and absolute observation. Root mean squared value also computes the average of magnitude of errors. From the experiment implemented with the used dataset as shown in Table.2 and Fig. 4.

Table 2: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

Algorithms	Root Mean Squared Error	Mean Absolute Error
BayesNet	0.1543	0.0282
NaiveBayes	0.1937	0.0494
IBK	0.0807	0.0066
Kstar	0.1343	0.0348

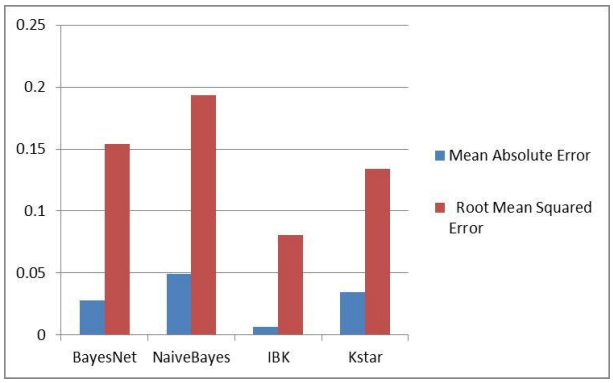


Fig. 4: Bar Graph of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

b) Root Relative Squared Error (RRSE) and Relative Absolute Error (RAE)

These are the errors show performance of every experiment. Absolute error shows physical error and relative error shows information about how much efficient a particular attribute relatively measured. Table.3 and Fig. 5 shown the dataset taken in this experiment to compared four algorithm.

Table 3: Root Relative Squared Error (RRSE) and Relative Absolute Error (RAE)

Algorithms	Root Relative Squared Error	Relative Absolute Error
BayesNet	30.9376	5.6561
NaiveBayes	39.4328	9.9298
IBK	16.1701	1.322
Kstar	26.8804	6.9664

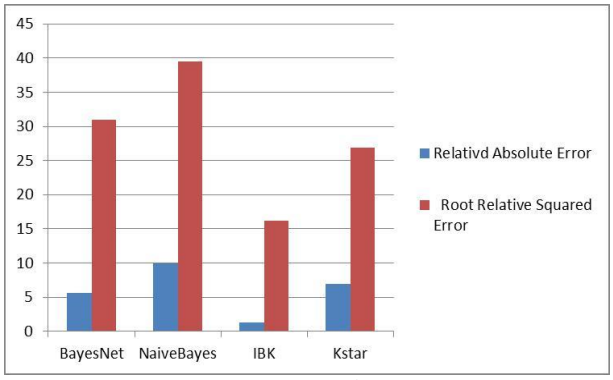


Fig. 5: Comparison Graph of RAE and RRSE.

5.3. Accuracy measurement

The accuracy of algorithms is shown in Table.4 and Fig. 6. It is measured with the help of various parameters

Table 4: Values of TP-RATE, FP-RATE, Precision, Recall, F-Measure for Algorithms

Algorithms	Precision	Recall	TP Rate	FP Rate	F-Measure
BayesNet	0.974	0.974	0.974	0.028	0.997
NaiveBayes	0.956	0.956	0.956	0.046	0.95
IBK	0.993	0.993	0.993	0.007	0.995
Kstar	0.11	0.98	0.99	0.02	0.99

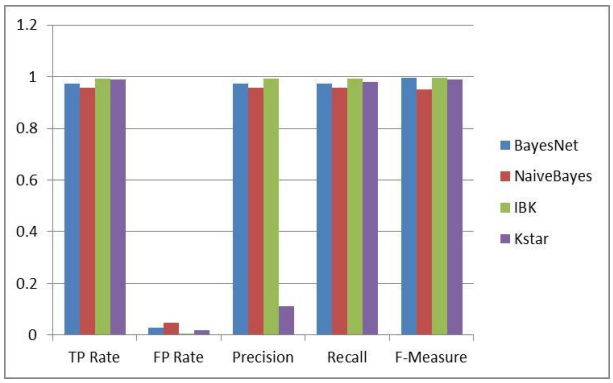


Fig. 6: Graphical Representation of Precision, Recall TP-Rate, FP-Rate, F-Measure Values.

Table 5: Accuracy of Measurement of Algorithms

Algorithms	Accuracy
BayesNet	97.2
NaiveBayes	95.4
IBK	99.3
Kstar	98.9

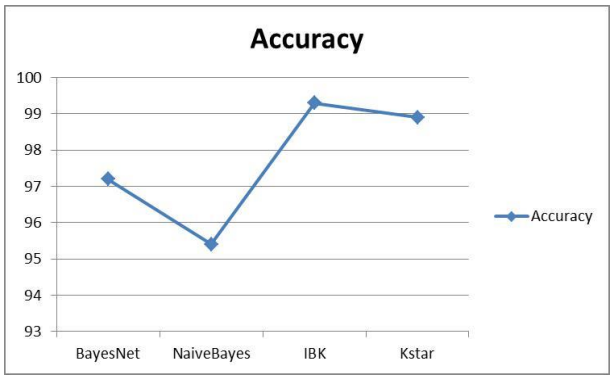


Fig. 7: Comparison of Algorithm Accuracy.

6. Conclusion

In this work, the performance of Bayesian (BayesNet, and Naïve Bayes) and Lazy classifiers (IBK and Kstar) are analysed by applying various performance factors. From the performance analysis, lazy classifier is more efficient than Bayesian classifier. The host based intrusion detection system and network based intrusion detection system can be integrated for better detection. Developing a new IDS schemes for detecting novel attacks rather than individual instantiations.

References

[1] SY Ji, BK Jeong, S Choi and DH Jeong, "A multi-level intrusion detection method for abnormal network behaviors" ELSEVIER: Journal of Network and Computer Applications, vol.62, pp.9-17, 2016. https://doi.org/10.1016/j.jnca.2015.12.004.

[2] Huang L, Milne D, Frank E, Witten IH, "Learning a concept-based document similarity measure", Journal of the Association for Information Science and Technology, pp.1593-608, 2012. https://doi.org/10.1002/asi.22689.

[3] Vaithiyanathan V, Rajeswari K, Kapil Tajane and Rahul Pitale, "Comparison of different classification techniques using different datasets", International Journal of Advances in Engineering & Technology, May 2013.

[4] Sharma TC, Jain M, "WEKA approach for comparative study of classification algorithm", International Journal of Advanced Research in Computer and Communication Engineering. April 2013.

[5] Amor NB, Benferhat S, Elouedi Z, "Naive bayes vs decision trees in intrusion detection systems", ACM symposium on applied computing, vol.14, pp. 420-424, March 2013.

[6] S. Durai , N. Rajkumar, N. K. Manikandan and D. Manivannan "Data Entry Works in computer using Voice Keyboard" , Indian Journal of Science and Technology, Vol 9 (2), January 2016

http://nsl.cs.unb.ca/NSL-KDD/.
https://doi.org/10.17485/ijst/2016/v9i2/85814.

[7] Aljawarneh S, Aldwairi M, Yassein MB,"Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model", Journal of Computational Science, March 2017. https://doi.org/10.1016/j.jocs.2017.03.006.

[8] Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, Vázquez E,"Anomaly-based network intrusion detection: Techniques", systems and challenges. Computers & security, pp.18-28, March 2009. https://doi.org/10.1016/j.cose.2008.08.003.

[9] Alaei P, Noorbehbahani F, "Incremental anomaly-based intrusion detection system using limited labeled data", IEEE: International Conference, pp. 178-184, April 2017. https://doi.org/10.1109/ICWR.2017.7959324.

[10] Van NT, Thinh TN, Sach LT, "An anomaly-based network intrusion detection system using Deep learning. InSystem Science and Engineering", IEEE International Conference pp. 210-214, July 2017.