RESEARCH ARTICLE OPEN ACCESS

Comparison of Classification Algorithm based on Machine Learning Techniques and Secure Network Information in Lesser Time

Sarvottam Dixit [1] Aaina [2]
[1] Pro-Vice Chancellor in Mewar University Chittorgarh
[2] Research Scholar in Mewar University, Chittorgarh - Rajsthan

ABSTRACT
One of the difficult challenges of the current Network Security sector is its network difficulty. For policy infringements or dubious traffic, a network must be continually monitored. Therefore, it is necessary to create an intrusion detection system that can monitor the system for any damaging actions and provide the ultimate control with findings. The creation of a system that can identify network intrusion may play a huge role in data mining. Data mining is a process through which large data archives can extract valuable knowledge. The network's traffic may be widely classified in two classes - normal and abnormal - to identify intrusion. In our study we investigated the categorization of the traffic in the network by various classification approaches and machine learning algorithms. We have discovered nine appropriate classifications out of the classification approaches such as Naïve Bayes, IBK, J48, Random Forest and Decision Stump. We focused on boosting, bagging, and mixing (storing) and analyzed their accuracy and reliability out of the many machine learning methods. Comparisons were conducted using the WEKA tool is given below accordance to particular efficiency measurements. A 10-fold cross validation was done to simulate these categorization frameworks.

Keywords: **-** lazy Classification, meta classifier, Random Forest, classification Rule, Trees**.**

I. INTRODUCTION

The constantly increasing load of network activity has made packet prioritization a fascinating problem in the modern society. A tremendous quantity of data flow, including harmful data, is required for a network. An organization must facilitate the flow of the network and discover any incursion that violated the company's regulations. An intrusion detection system that would be sufficiently effective to intrusion detection systems therefore needs to be created. A network must also be secured from potential assaults. The methods used for intrusion detection may be generally grouped into two main categories of NIDS and Web application firewalls that identify both the host and the networking. NIDS are located tactically in the networks at nodes so that the incoming traffic on a whole network may be analyzed and combined with its own libraries of prepared assaults. A message is issued to the administrator upon noticing an anomalous network activity or when disclosing an intrusion. HIDS only operates on single hosts or connected devices, instead. It monitors incoming and outgoing packets on the device and gives the administration an alert to detect suspected packets. Two forms of NIDS are generally: - abnormality and signatures based [1]. For a given vulnerability a signature-based system is established, hence it has a lower number of false positives, which offers less adjustability. While an abnormality system is much more flexible and will look for potential threats that are not defined, this leads to more fake positive. It can only recognize assaults without accurately identifying the sort of violence. The network traffic is categorized as normal and abnormal as a detection system of intruders. Network traffic is anomaly if the behavior of communication activity differs from the usual networking activity patterns. The effectiveness of the Malware depends on the classification technique. The algorithm is significant for its time complexity consuming in the selection procedure of an algorithm.

For the purposes of classification of network traffic in the above two categories [2], data mining method is utilized. It includes the extraction and processing of enormous amounts of data. To construct a model of conventional direct and utilize the model of decision making and forecasts, machine training techniques are also used. Before they can be used for extremely sensitive applications like machine learning algorithms, the efficiency and effectiveness of these approaches need be evaluated [3].

II. ALGORITHMS FOR WEKA TOOLS

WEKA is a technology used both for data mining as well as for learning algorithms. It was originally implemented in 1997 by the University of Waikato in New Zealand [4]. It consists of many discrete optimization techniques. One of the drawbacks of this application is that it is only possible to support data sets in the ARFF and CSV formats (comma values). It was designed in C initially, but then revisited in

JAVA. It contains computer program for interacting with computer systems. It includes 49 tools for data preprocessing, 15 attribute evaluators, 76 classification techniques and 10 search techniques. It consists of three distinct sorts of GUIs: "The Explorer," "The Experimenter" and "The Knowledge Flow." WEKA offers the possibility to create any new algorithms for machine learning. There are viewing tools and several modules for carrying out the required activities.

To categorize network traffic to ordinary and anomalous categories in principle, classification techniques or classifiers are needed. The aim is to obtain high exactness and precision and to categorize the items behind classifying technologies. Eight types may be broadly categorized in WEKA, which includes several machine learning techniques for each classification. Clearly introduced here are the categorization systems.

Bayes Classification: It derives from earlier based classification investigations and is linked to the probabilistic group. A probabilistic summary must be kept for each class. This summary stores the likelihood function of each characteristic and the likelihood of the class. The graphical models show knowledge of uncertain areas. As graphs [5], dependent variables are shown in nodes and probabilistic weights are applied to edges that link together the relevant random number nodes. When a new instance is found, an update of the recorded probability with the class [6] is only made by the algorithm. In this procedure, the order of training events and the occurrence of classifications mistakes have no impact. It must therefore simply forecast the class based on the value of the class components. There are 13 classifiers in these categories, but only three of which are acceptable with our data set.

Classifying function: It uses the extrapolation and computer program idea Data input to output is translated. It uses the technique to estimate the iterative parameter. In all, 18 classifiers in this category are available, of which Two of our collections are comparable.

Lazy categorization: it requires the whole support vectors to be preserved and relevant data just after the categorisation period. The main advantage of this classification system is that the moving charged is locally approximated [5]. To resolve numerous issues simultaneously, the goal function is approached locally for every query of the system. But the drawback is that a lot of storage capacity is needed for all training instances to be stored simultaneously. It also takes time. In this category 5 classification devices are available, however only two are consistent with our data set.

Metaphysical Classification: These classifier sets are important to determine the best collection of characteristics that can be utilized in the basic classifier training [7]. These classifiers may be utilized to build adaptive control machine learning algorithms and to make predictions of these new models. The category has 26 classifiers, 21 of which are acceptable with our dataset. My Classifier: My representatives are Classificatory of Multi-Instances [8]. It comprises of several examples in an example but only for all occurrences is one class seen. It is therefore an unstructured approach of learning. The 12 classificatory in this category are inconsistent with our dataset.

Misc. categorizer: This subcategory is composed of many classifier kinds. Only two of them are consistent with our data set out of three.

Classification Rules: Association Rule is utilized between all attributes to improve classification model. The accurate quantity of the forecast is specified in a percentage or exact format by the word coverage. The rules of affiliation exclude each other. Most surveyed classifications are available under this categorization, while 8 are in line with our given dataset.

Trees: This is a method for creating a tree flow diagram in which the goal associated with implementation is tested for every node, representing each branching with the outcome of each test. Trees: Predicting and explanatory is the model developed. The expected classes are shown as the tree branches. There are 16 classifications, of which 10 are regarded as satisfactory.

In this paper we will explain in further depth the classification methods we have performed. BayesNet: It is a strategy that works on the fundamental assumption for Bayes and builds a Bayesian network [9] after computing the average of a condition for each node. This graphics concept is a conventional technique and depicts, using a guided acyclic graph, a series of arbitrary parameters together with their dependence.

IBK: It means visual representations of training cases [10] for instance and it does not infer or forecast a set of rules or a tree of decisions. The memory is examined for the current training instances after several training instances have been saved. It therefore takes time and distance.

J48: It's an enhanced version of C4.5 with several extra functions surrounding the ID3 technique for dealing with difficulties ID3 unable to address. [11]. This approach, however, consumes knowledge and power. It first creates a tree using the technique integral images and then uses heuristic criterion. Accurate and accessible principles are used to create the tree.

Random Forest: This technique for classification employs ensemble approaches to get greater prediction effectiveness. The production is based on the decision tree algorithm in the context of individual trees. The classifier is very precise and can accommodate several factors. Decision Stump: a stump is a one-stage algorithmic study model. In other words, it is a statistical method with a single element (root) instantaneously connected to the output layer (its leaves). A stumbling block for the choice forecasts the relevance of only one element. Sometimes they are called 1-rules.

We utilized several machine learning methods in this paper. They construct and use a framework based on inputs to make decisions and predictions. The algorithms we utilize

are AdaBoost, Stacking and Picking. They were addressed below in depth.

AdaBoost: This represents an algorithm of adaptive boost [12]. It is an ensemble-based technique that is based on learning algorithm, initiated by a basic classifier. Then there is a second classifier to focus on cases in the training data that were incorrectly acquired from the basic classification. Add further classifications continue until a certain limit in numerous models or in correctness is reached. For the basic classifier enhancing utilizes the J48 algorithm. Boosting helps to improve the precision of any algorithm.

Bagging: Aggregating [13] is an ensemble approach creating several Learning selected features and classification for individual instances. Finally, by use of average or overwhelming voting the outputs of these several classifiers are linked. Since each sample is distinct from the other, the focus and perspective of each classification model on the issue are distinctive. The basic classifier is used by the J48 as well. Bagging lowers variation and contributes to prevent overfitting. It enhances machine learning algorithms' accuracy and robustness.

Stacking: The grouping or mixing of various algorithms on the learning algorithm is another ensemble procedure. A Meta classifier is produced that learns to anticipate each classifier and to predict accurately data not shown. The two grades that are employed are J48 and IBk, and Regression models is used in the Meta classifier. Blended is essentially the mixing of many technique types. So, we use J48, under the part of the tree, and the IBk, under the lazy section, that is, totally distinct algorithm sets. They can have a unique view of the situation and produce various meaningful forecasts. Logistic regression is a common and accurate approach to discover how the projections The various approaches may also be combined in accordance with the aforementioned. It generates binaries outcomes and is suitable for categorizing binary texts.

III. ANALYSIS OF MEASURING PERFORMANCES

Classificatory effectiveness may be evaluated with various measurements such as accuracy, selectivity, sensitivities, training duration etc. The base from which various parameters can be computed is a confusion matrix. A confusion matrix can be used to calculate the number of cases precisely or inexactly anticipated by a model of categorization. As illustrated in Table I, Usually 4 parameters TP, FN, FP, and TN are expressed in the confusion matrix[7]. The factors are described succinctly following.

True Positive (TP): The instances are shown as frequent in precise forecasts.

False negative (FN): This implies an incorrect prediction, i.e., it recognizes as normal occurrences of assaults.

False positive (FP): offers an indication of the usual frequency of assaults identified.

True negative (TN): Instances accurately discovered during an assault are indicated.

Table I.

		Predicted value	
		Normal measure	***Anomaly measure***
Actual	Normal	Ture Positive	False Negative
	Anomaly	False Positive	Ture Negative

Performance Measures

ROC: This sentence is essential to establish the curves among true and false positives negatives (TPRs) (FPR). The region below the curve is called AUC, the ROC value. The bigger the area of the curve, the higher the ROC value is.
Sensitivity: It is also called a genuine positive rate and delivers the accurately recognized positive results. Sensitivity: Therefore, the system is likely to be able to properly predict positive cases [1].
Ture Positive / (Ture Positive+False Negative) =Sensitivity
Specificity (SPC): it is also known as a real negative interest rate or offers an adequate estimate of the actual negativity. Thus, the algorithm is likely to predict negative cases properly.
TN / (FP+TN) Specificity=TN
Precision: It predicts that a positive forecast is right. Precision:
TP / (TP+FP) Precision=TP
Precise: when stated in percentages, the percentage of relevant forecasts shows the accuracy. The confusion matrix may be computed using the formula:
Precision=(Ture Positive+Ture Negative)/(Ture Positive+Ture Negative+False Positive+False Negative) Accuracy
Kappa: is used to evaluate the information collection computation reliability. There are 2 principles: 0 and 1.0 and complete variances; 1 is perfectly agreed.
Important Mean Error (IME): For an algorithm to be optimally performed, this error should be minimal. It means that a software generates a generic violation.
The F1 score is a fourier transform of sensitivity and precision. The sophisticated modules can be evaluated with this evaluation metrics.
(2 TP+FP+FN) =2*TP / F=2*
False positive rate (FPR): it shows that an algorithm can anticipate occurrences as regular assaults.
RFF = RF / (RF + RF)
-SPC
The False Discovery Rate (FDR) measures the chance of a false positive forecast.
RAF / (RAF + RAF) =1

-PPV
Negative predictive rate (NPV): It reflects an algorithm's probability of properly detecting attacks.
TN / (TN+FN) NPV=TN /
Training time: it's time to create the strategy on the dataset. Training time: The measurement is generally in seconds. The lower the correction factor, the stronger the classifiers are.

IV. DATASET USED

Three datasets were utilized to do the comparative study. Table II provides three dataset descriptions, namely several characteristics and occurrence number. These web-based data sets were gathered (such https://www.unb.ca/cic/datasets/dohbrw-2020.html). The data set 1, in the.csv format, is also in the same format, with the dataset 2 and the dataset 3.

Table II Data sets

Dataset name	No. Of Attributes	No. Of Instances
Dataset 1	5	150
Dataset 2	9	1253
Dataset 3	9	2924

Comparison of Measuring of Algorithm

The six classification techniques are evaluated by the following variables

(a) Size of the dataset (b) Number of classifying (c) Time to classify

The classification algorithms are split into two partitioning groups. The first is to independently evaluate partitioned classification algorithms with non-partitioning-based methods and draw the findings.

A. Comparison of different data sets of classification

The WEKA (3.7.10) includes three datasets and the findings linked to the time needed to construct classification and classification number. Table III describes the duration to build partition-based classification classifications and non-partitioning methods using distinct information sizes. The result is that the data set size decreases the time it takes to categorize itself. For all three datasets, farthest took the less time to construct classes, whereas Random Forest took the largest proportion.

TABLE III: Algorithms time taken

Algorithm	Time DATA SET 1	Time DATA SET2	Time DATA SET 3
Naïve Bayes	0.10	0.13	.67
Random Forest	18.55	45.95	120.41
IBK	0.01	0.03	0.09
J48	1.06	1.72	6.17
Decision Stump	0.8	0.16	1.61

Weka (3.7.10) received three datasets and documented the outcome of the classification process. The comparative analysis of the results is presented in Fig 1.

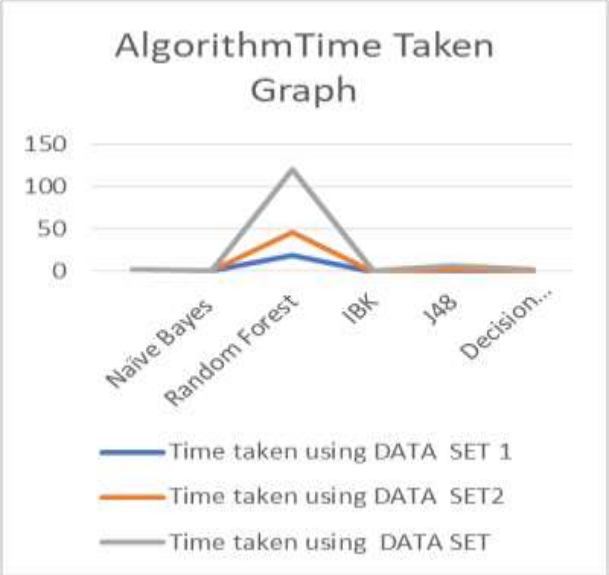

Fig 1 – Time graphical representation to establish classifications

The range of different knowledge capacity technique categorization is described in Table 3. The default settings for the number of classifications have been taken in partitioning-based classification. The value of k is not specified; WEKA (3.7.10) does not describe the possibilities. Inference reached is the number of classifications that are determined by decision stumps that are the same datasets. Largest number of classifications were produced by the random forest.

Table IV: Distinct information size Number of categorization generated

Algorithm	Number of Classification DATA SET 1	Number of Classification using DATA SET 2	Number of Classification using DATA SET 3
Naïve Bayes	2	2	2
Random Forest	7	11	21
IBK	2	2	2
J48	4	9	6
Decision Stump	2	2	2

Weka (3.7.10) has received three data sets, the results of which have been recorded in relation to the number of classifications produced. The statistical analysis of the results is presented in Fig 2.

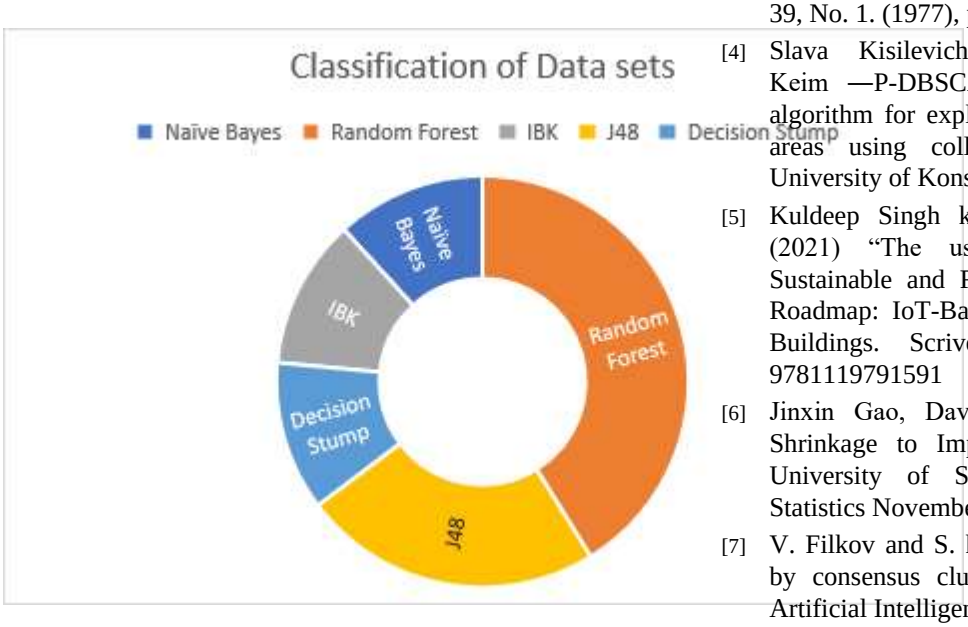

Fig 2: Graphics depiction of classification number generated by different information sizes

CONCLUSION

A comparison research was conducted on the classification of algorithms in three distinct datasets. Different classification methods are evaluated based on data quantity, time taken to create classifications and classification number. We also examined the effectiveness of several WEKA classifiers and found that for this purpose Random Forest & Bayes Net are adequate. Also compared with data mining algorithms, boosting may be concluded to be the optimum concentration. The experimental data of several classification methods are graphically represented. Decision The algorithm of Stump required less time to classify, and the greatest classification number was the Random Forest approach.

REFERENCE

[1] Yuni Xia, Bowei Xi ―Conceptual Clustering Categorical Data with Uncertainty‖ Indiana University – Purdue University Indianapolis Indianapolis, IN 46202, USA

[2] Himadri Chauhan, Vipin Kumar, Sumit Pundir and Emmanuel S.Pilli,"A Comparative Study of Classification Techniques for Intrusion Detection", IEEE International Symposium on Computational and Business Intelligence,2013

[3] A. P. Dempster; N. M. Laird; D. B. Rubin ―Maximum Likelihood from Incomplete Data via the EM Algorithm‖ Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp.1-38.

[4] Slava Kisilevich, Florian Mansmann, Daniel Keim ―P-DBSCAN: A density-based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos, University of Konstanz

[5] Kuldeep Singh kaswan, Jagjit Singh Dhatterwal (2021) "The use of Machine Learning for Sustainable and Resilient Building" Digital Cities Roadmap: IoT-Based Architecture and Sustainable Buildings. Scrivener Publishing Press ISBN: 9781119791591

[6] Jinxin Gao, David B. Hitchcock ―James-Stein Shrinkage to Improve K-meansCluster Analysis‖ University of South Carolina, Department of Statistics November 30, 2009

[7] V. Filkov and S. kiena. Integrating microarray data by consensus clustering. International Journal on Artificial Intelligence Tools, 13(4):863–880, 2004

[8] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. In Proceedings of the thirty-seventh annual ACM Symposium on Theory of Computing, pages 684–693, 2005

[9] E.B Fawlkes and C.L. Mallows. A method for comparing two hierarchical clustering's. Journal of the American Statistical Association, 78:553–584, 1983

[10] M. and Heckerman, D. (February, 1998). An experimental comparison of several clustering and initialization methods. Technical Report MSRTR-98-06, Microsoft Research, Redmond, WA.

[11] "Machine Learning and Deep Learning Algorithms for IoD" in book entitled "Internet of Drones: Opportunities and Challenges" in "Apple Academic Press (AAP), Canada, Publishing date Feb 2022, Hard ISBN: 9781774639856 (https://www.appleacademicpress.com/the-internet-of-drones-ai-applications-for-smart-solutions/9781774639856).

[12] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30.