ISSN No:-2456-2165

Generalized Flow Performance Analysis of Intrusion Detection using Azure Machine Learning Classification

Dr. Narasimha Chary CH[1]
(Associate Professor)
Dept of CSE , Sri Indu College of Engineering and Technology (Autonomous) Sheriguda, Hyderabad, TS, India- 501510

Dr. Srihari Chintha[2]
(Associate Professor)
Department of CSE Vishwa Vishwani Institute of Systems & Management Survey No. 128, Boston House, Thumkunta Post, Shamirpet Road, Hakimpet (via), Thumkunta, Telangana 500078

E.Rajendra[3]
(Associate Professor)
Dept of CSE , Sri Indu College of Engineering and Technology (Autonomous) Sheriguda, Hyderabad, TS, India- 501510

Dr. Sunke Srinivas[4]
(Associate Professor)
Holy Mary Institute of Technology and Science

Abstract:- The development of real-world databases presents computing difficulties for a single computer. Cloud-based systems, on the other hand, can handle massive quantities of data management activities for large-scale real-world data set calculations. The study focuses on a new Generalized Flow inside the cloud computing platform, Microsoft Azure Machine Learning Studio (MAMLS), which analyses multi-class and binary classification data sets to maximise overall classification accuracy. To begin, each data set is split into training and testing sets. Following that, the training data is utilised to create classification model parameters. Reduce the dimensionality of your data to enhance classification accuracy. Data-centered information increases overall classification accuracy by reducing multi-class classification to a series of hierarchical binary classification problems. Finally, the performance of the improved classification model is tested and appraised. The proposed study assessed algorithm performance utilising 82,332 test samples from a recent data set, UNSW NB-15. It took 6 seconds to train 1,75,341 network instances using the suggested two-class forest decision model. At 99 percent, 94.49 percent, 91.79 percent, and 90.9 percent, the multi-level forest decision- making model recognised attack types such as generics, feats, shellcodes, and worms, respectively.

***Keywords*:-** *Azure Machine Learning; Decision Forest; Intrusion Detection; Locally Deep SVM; Mutual Information; UNSW NB-15.*

I. INTRODUCTION

Business optimization, health care, financial trading, data visualisation, medical/healthcare data, and the Internet of Things are just a few of the real-world applications that have benefited from big data analysis in recent years. [1] In general, efficient management of volume, speed, variety, and complexity [of data] is required for big data processing[2]. The advancement, storage, and computing paradigms were all used to develop a plethora of new methods. Multi-threaded computational single-node memory analytics over cached in- memory data may be broadly divided into the following categories: • Consider the following examples: MATLAB; R; Weka; RapidMiner; and so on

Data is stored across many clusters in a distributed parallel/distributed storage system such as the Hadoop Distributed File System, and parallel computation is performed with the help of a Message Passing Interface (MPI). Adapting advanced machine learning (ML) algorithms to new data paradigms is required for the majority of real-world applications. In the Microsoft Azure Machine Learning Studio (MAMLS), users may train and deploy machine learning models on a computer cloud-based platform that can be accessible through a web-based user interface. Microsoft-developed categorization modules make a variety of machine learning methods available. With this work, we modify existing machine learning modules in MAMLS and combine them with 'R' modules to generalise the flow, which scans a small portion of "big data" that can be acquired from a variety of applications, and maximise multi-class binary classification accuracy with the least

ISSN No:-2456-2165

amount of manual intervention. In order to report credit score, pathology, and anomaly issues, a broad range of classificators has been investigated thus far. Improved model parameters, followed by a feature score, have been created for credit scoring models, as shown in [3], for instance. Logistics regression, decision trees, nearest neighbours, linear discrimination analysis, neural networks and SVMs were some of the well-known classifiers [5]. The typical breast cancer detection and diagnosis have been applied by different neural network classification models for the creation of expert diagnostic systems[8] in the context of applications of pathology classification. As an additional measure to decrease data dimensionality, feature ranking techniques based on Spearman rank coefficients as well as Fisher-scoring algorithms were extensively used to achieve this [3]. the number [5] the number [8] [5] Two significant contributions are made by this essay. first, The first step is to apply various ranking methods and discover that the classification accuracy is optimised for both public and local data sets for an optimal collection of top ranking features for both public and local data sets. Developing a generalised flow that reduces dimensionality while optimising classification parameters via multi-class and binary classification tasks is the second step in the process of increasing overall classification accuracy. An improved classification accuracy has been demonstrated on a local medical picture data set using a customised modular (generalised flow) representation of the cloud-based MAMLS platform, which has been benchmarked in terms of classification performance on public data set 2 and has been shown to optimise classification accuracy on a local medical picture data set.

II. RELATED WORK

➢ *Azure Machine Learning (AML)*

Recently, Azure Machine Learning (AML) has used all fields of medical services, industry, administration and engineering to improve prediction accuracy[16]. AML is basically a cloud computer that focuses on paradigm shifting. It is because of the dual processing location and changes in data storage from the company computers to servers on the internet[17, 18] and also because of the high availability and elastic processing power of the computing resources that are used so that we can communicate about a new computational model[19]. Azure machine learning acts as a cloud-based computing platform, accessible through a web interface. As shown in Fig. 1, the categorization models of Microsoft include a range of ML techniques. Moreover, Azure's ability to leverage pre-built Machine Learning Program Interface (API) for product recommendation.

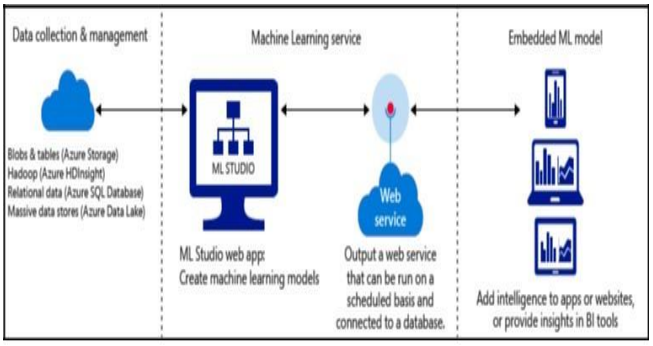

Fig 1 The Basic Workflow for Azure Machine Learning [21]

➢ *Classification Model Development for Predicting*

Two types of support vectors may be used for model generation: a multi-class neural network and a multi-class decision forest. These courses are focused on a supervised approach to learning. Therefore, a tagged dataset with a label column is needed. For example, this neural network model is useful in predicting binary output, including therapy or remedy.

After the data is prepared, the algorithm is selected and a prediction model is created. This is where methods for machine learning are useful. A subset of the prepared data is used to train the model and the model may then be tested using another set or other ready data for performance assessment. Different algorithms may be tested while assessing the performance and the one carrying out the test can be selected.

➢ *Data Pre-processing and Feature Selection*

The particular variables required to build the model in most classification jobs frequently need to be identified. The Machine Learning Studio has two function selection modules to help choose the most appropriate variable for modelling applications such as the linear discriminatory analysis and filter-based selection of features[23]. There are defined objectives that affect the selection of an appropriate strategy. In this instance it is predicted by classifying models in order to get the highest accuracy and descriptive mining. In this research, three methods spanning from MA multiclass algorithms offered in the Azure ML Studio are analysed in comparison.

- *Two Class Vector Support Machine*

✓ *Multi-Class Neural Network*
✓ *Multi-Class Forest Decision*

The following defines the problem of classification: If the sampled input X= (X1, X2, …,…, Xd), the input X1 is indicative of an item in sample size, the aim is to examine the mapping of X Y, representation of a class with y My.

ISSN No:-2456-2165

➢ *Two-Class Neural Network*

The multilayer perceptron (MLP) is one of the most important models in the Artificial Neural Network[25]. A neural network consists of a collection of linked layers in which the inputs are guided to the output by a sequence of weighted nodes and edges. The input layer receives signals from external nodes. The artificial neural network would help doctors to model, analyse and comprehend complex clinical information across a broad variety of medical applications. The calculations are performed for each node situated in the hidden layers and in the output layer in order to calculate network output for a given input. The value is calculated by applying a function triggered on the sum weighed for each node. Many processing elements form a neural network (NN), whereas the PEs consist of weighted interconnections.

➢ *Class Support Vector Machine*

In 1992, Bernhard presented SVMs at the Conference on Learning Theory (COLOT) to support vector machines. The SVM is based on statistical theory of learning and is widely considered as a type of a kernel-based learning algorithm. SVMs are essentially models of learning which are supervised and help in data analysis and pattern identification. They may be used both for regression and classification tasks. If a sample set of samples is allocated from one of the two classes, additional samples are assigned to one class or the other using the SVM algorithm. One of the first machine learning methods included vector support systems. Although the newest advancements have created more accurate algorithms, this method can work on basic data sets when it emphasises speed rather than accuracy.

➢ *Two-Class Decision Forest*

One of the ensemble's categorization learning approaches is the decision forest algorithm, which works via the production of various decision-making bodies and hence the vote on the favourite class of output. Voting is basically the aggregate variant, such that a standard frequency histogram of labels is a classification decision forest output. The ultimate judgement of the ensemble favours the trees with the highest certainty. A set of fundamental tests are performed on each tree class, increasing the tree structure level till decision is taken on the leaf node.

III. METHODOLOGY

This section examines the various components of in-depth experimentation. This article focuses on eight two-class and three multi-class classification techniques. Classification models have been created in four distinct stages: pre-processing, feature selection using mutual information, adjust hyperparameters and workflow prediction building. UNSW NB- 15 includes basically 47 features and two class names. The dataset includes continuous, discrete and symbolic features, which are preprocessed in different ranges. During the experiment, all nominal attributes were converted into integer. Numerical features with a wide range are difficult to handle. The logarithmic scale was used to minimise its range of values. For Boolean features, no scaling was required. Min-max standardisation was used to determine the lowest and highest value of each feature in the [0, 1] range.

$$V' = \frac{V - min_i}{max_i - min_i} \quad (1)$$

In (1), min and max the lowest and maximum values of every function are defined as "i." Every V-function is scaled to V." The score feature was utilised to prioritise characteristics followed by the construction of processes for categorization jobs. When tested, mutual information provided relatively better outcomes than other function selection approaches based on filters. Mutual information is a measure of information between a random feature "x" and an objective variable "y" or the label[27], as the name indicates. The mutual information between two variables is (2) as described in (2) and (2) (3).

$$I(X;Y) = \iint p(x,y) log \frac{p(x,y)}{p(x)\,p(y)} dxdy \quad (2)$$

$$S = argmax\, I(X_S;Y) |S| = n \quad (3)$$

In (2) p(x,y) indicates the function of joint probability density, p(x), p(y) are marginal density functions. "n" refers to the number of selected features and is known as shared mutual information in the context of feature selection. The subset of chosen characteristics is called XS as indicated in (3). Table 1 shows the distribution of training and test data sets. In the training and testing distributions, there was no redundancy unlike benchmark datasets[17, 18]. As stated before, mutual information was employed as an Azure Machine learning studio feature score approach. The outstanding characteristics, as mentioned in Table 2, are provided as input to the various classifiers to achieve the best predictions. A 10 times cross validation was used to accurately evaluate the performance of all the machine learning algorithms examined in the study and a different test set was used for the assessment. In machine learning research, cross validation becomes crucial to control overfitting and confirm the capacity of algorithms to generalise independent data (testing set)[28].

ISSN No:-2456-2165

Table 1 Dataset distribution

Class	Training Samples	Testing Samples
Normal	56000	37000
Analysis	2000	677
Backdoor	1746	583
Reconnaissan		
ce	10491	3496
Shellcode	1133	378
Worms	130	44
DOS	12264	4089
Fuzzers	18184	6062
Generic	40000	18871
Exploits	33393	11132
Total	1,75,341	82,332

Table 2 List of Salient Features

Sl No	Name of Feature	The Feature Score
1	Ct_State_Ttl	0.686
2	Dttl	0.56
3	Sttl	0.27
4	Dinpkt	0.23
5	Smean	0.20
6	Rate	0.199
7	Ct_Dst_Sport_Ltm	0.196
8	Sload	0.190
9	State	0.1875
10	Dload	0.1872
11	Sbytes	0.185
12	Dpkts	0.175
13	Dbytes	0.171
14	Dur	0.158
15	Ackdat	0.156
16	Dmean	0.147
17	Synack	0.138
18	Tcprtt	0.131

➢ *Averaged Perceptron*

Averaged perceptron is a simple neural network structure used to categorise data using a linear function. In order to assess the expertise of two classes of Averaged Perceptron, MAMLS provides the possibility to select a single value or multiple values as learning rates. The learning rate for the best configuration of the stochastic descent optimizer was tuned to several parameters like 0.1, 0.5 and 1.0. The advantage of utilising a parameter range is that the model produces the optimum model over multiple possibilities.

➢ *Bayes Point Machine*

Bayes point machine is built upon the Bayesian principle to categorise network instances effectively by picking a Bayes point (average). Iterations are typically set between 5 and 100. This value shows the number of times the algorithm iterates through the training data. Numerous trials were done in different iterations within this range, but the results were not sufficiently compelling and the training time during these trials was longer. However, the results were good in setting the number of training iterations at 30 for Bayes Point Machine, which was the foundation for maintaining 30 as the number of training iterations for this experiment.

➢ *Boosted Decision Tree*

The Boosted Decision Tree is an ensemble model primarily intended to correct mistakes in previously constructed trees. The competency of the two-class Boosted Decision Tree was examined using four essential hyper-parameter values as given in Table 3. Here the maximum number of leaves indicates the maximum leaves in any tree. The size of the tree may be raised by changing this value, however excessive and extended workout duration has been increased by the number of leaves. The minimum number of leaf node samples refers to the number of situations in which a leaf node is created. Value 10 means that the training data has 10 examples that fulfil the same requirements as the established rules. The initial learning rate was 0.2 which basically indicates the pace of convergence. In addition, 100 decision-making bodies have been established in the ensemble. More than 100 trees are also provided but again, the training period is significantly longer, which is thus not recommended.

Table 3 Critical Parameters used for Configuring boosted Decision Trees

Max. Leaves per Tree	Minimum Number of Samples per Leaf Node	Learning Rate	Number Constructed of Trees
20	10	0.2	100

➢ *Decision Forest*

Two-class decision forests is one of the most favoured models of binary classification as proposed by Team Azure[15]. There are two resampling techniques available to construct a two-class forest model, namely replicating and bagging. Replicate technique trains each tree on the same training data while the aggregation or bagging of bootstraps allows each tree to develop on a new sample. It should be noted that the values provided in Table 4 represent optimum outcomes when assigned.

Table 4 Critical Parameters used for Configuring Decision Forest

Number of Trees	Decisionmaximum Decision Tree	Depth of number Splits Per Node	Randomminimum Number of Samples Per Leaf Node
8	32	128	1

ISSN No:-2456-2165

Maximum depth increased to maximum accuracy of 1 although overfitting was seen, resulting in a longer training period (not desirable). The number of random divisions indicates the number of divisions produced per node from which the optimal division can be selected. The minimum sample number per leaf node refers to the number of instances required for creating the leaf. It has been tried to detect if improved outcomes might be achieved by altering, but not effective, the values of essential parameters.

The kernel Sigmoid does not meet the theorem of Mercer. Large values cannot thus be applied to sigmoid sharpness. Smaller numbers like 1 can influence the threshold when utilised. Table 6 demonstrates the essential default settings for local deep SVM models of the two classes.

Table 5 Critical Parameters used for Configuring Locally Deep SVM

Sigmoid Depth of the Tree Lambda		Lambda Theta	Lambdatheta Prime	Sharpness
3	0.1	0.01	0.01	1

- *Support Vector Machine*

Two SVM classes utilise regularisation of L1 (Lasso) to manage overfitting. The Lambda W=0.001 default value was specified as weight as a non-zero value is desirable to control the overfitting degree.

- *Logistic Regression*

Tolerance for optimization is a threshold that is typically given while building an L-BFGS (Broyden-Fletcher-Goldfarb-Shanno) optimization model for two class regression [15]. This model requires correct adjustment of the L1 and L2 values set to 1 and 1 respectively. The memory capacity in megabytes utilised by L-BFGS optimizers was set to 20, which shows the previous gradients kept in memory for consecutive stages. If the memory size is greater, then it slows down the training process and the model is defective in every way. Two class logistic regression model as described in Table 7 have been constructed and tested with three significant parameters. Regularization is commonly performed in order to reduce overfitting in classification issues.Table 7. Critical parameters used for configuring logistic regression.

Table 6 Logistic Regression

L1 Regularization	L2 Regularization	Memory size used by L-BFGS
1	1	20

IV. CONCLUSION

In this work, eight two- and three multi-class classification models utilising UNSW NB-15 data sets were constructed. On the basis of an empirical study, it can be said that forest decision has performed well. Since the tests on local computers take a long time to do, the Microsoft Azure Machine Learning Studio (MAMLS) has been chosen. Besides the conventional performance measurements such as accuracy, accuracy, retrieval, f1-score and AUC, the suggested study additionally included training time and time for evaluating the algorithms' efficacy. The study suggested highlights that MAMLS may be used in big datasets as an expedient Integrated Development Environment (IDE). In the future, it will be interesting to use different datasets for intrusion detection and then measure the performance of different classifiers. Experts have long called on the research community to experiment with various datasets and to offer new network intrusion detection algorithms

Another way to investigate in the future may be to use predictive models as scalable web services to enhance MAMLS' capabilities. It will be technically difficult to develop a MAMLS- based wrapper method. Such wrapper-based techniques can help to show the efficacy of MAMLS, which will finally lead to a sensible appraisal of its calculation performance.

REFERENCES

[1] Hakimi, Zahra, Karim Feaz, Morteza Barati, "A Flow-based Distributed Intrusion Detection System Using Mobile Agents," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 3, no. 6, pp. 732-740, 2013.

[2] Jang-Jaccard J, Nepal S., "A survey of emerging threats in cybersecurity," *Journal of Computer and System Sciences*, vol. 80, no. 5, pp. 973-93, 2014.

[3] Yan F, Jian-Wen Y, Lin C., "Computer network security and technology research," *In 2015 Seventh International Conference on Measuring Technology and Mechatronics Automation*, IEEE, pp. 293-296, 2015.

[4] CH Nnarasimha Chary "AN EFFICIENT SURVEY ON VARIOUS DATA MINING CLASSIFICATION ALGORITHMS IN BIO INFORMATICS" INTERNATIONAL JOURNAL OF ENGINEERING & TECHNIQUES, 2018 , vol., no. 4 , issue 2 2008

[5] Aburomman, Abdulla Amin, and Mamun Bin Ibne Reaz, "Review of IDS Development Methods in Machine Learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 5, pp. 2432-2436, 2016.

[6] Buczak, Anna L., and Erhan Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153-1176, 2015.

[7] Othman, Suad Mohammed, Fadl Mutaher Ba-Alwi, Nabeel T. Alsohybe, and Amal Y. Al-Hashida, "Intrusion detection model using machine learning algorithm on Big Data environment," *Journal of Big Data*, vol. 5, no. 1,pp. 34, 2018.

[8] Tchakoucht, Taha AIT, and Mostafa Ezziyyani, "Building a fast intrusion detection system for high-speed-networks: probe and DOS attacks detection," *Procedia Computer Science*, vol. 127, pp. 521-530, 2018.

[9] Dr.CH.NARASIMHA CHARY,"COMPREHENSIVE STUDY ON MULTI-OPERATOR BASE STATIONS CELL BINARY AND MULTI-CLASS MODELS USING AZURE MACHINE LEARNING", " A JOURNAL OF COMPOSITION THEORY,Vol 14,Issue 6,2021

[10] Suthaharan, Shan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 4, pp. 70-73, 2014.

[11] DrSunke Srinivas,Dr.Narasimha Chary CH "PRIVACY PRESERVING MEDIA SHARING WITH SCALABLE ACCESS CONTROL AND SECURE DEDUPLICATION IN MOBILE CLOUD COMPUTING" , The International journal of analytical and experimental modal analysis Volume XV, Issue I, January/2023.

[12] Casado, Ruben, and Muhammad Younas, "Emerging trends and technologies in big data processing," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 8, pp. 2078-2091, 2015.

[13] Narasimha chary Cholleti1Research Scholar, Dept. of Computer Science & Engineering, Sri Satya Sai University of Technology &Medical Sciences,, Dr. Tryambak Hirwarkar "ANALYZING SECURITY OF BIOMEDICAL DATA IN CANCER DISEASE" JOURNAL OF CRITICAL REVIEWS ,VOL 7, ISSUE 9, 2020.

[14] Team, AzureML, "AzureML: Anatomy of a machine learning service," *In Conference on Predictive APIs and Apps*, pp. 1-13. 2016.

[15] Elshawi, Radwa, Sherif Sakr, Domenico Talia, and Paolo Trunfio, "Big data systems meet machine learning challenges: Towards big data science as a service," *Big data research*, vol. 14, pp. 1-11, 2018.

[16] Moustafa, Nour, and Jill Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," *In 2015 military communications and information systems conference (MilCIS)*, IEEE, pp. 1-6, 2015.

[17] NARASIMHA CHARY CHOLLETI, TRYAMBAK HIRWARKAR "BIOMEDICAL DATA ANALYSIS IN PREDICTING AND IDENTIFICATI CANCER DISEASE USING DUO-MINING", Advances in Mathematics: Scientific Journal 9 (2020), no.6, 3487–3495,Spec. Issue on RDESTM-2020

[18] Moustafa, Nour, and Jill Slay, "The significant features of the UNSW-NB15 and the KDD99 data sets for network intrusion detection systems," *In 2015 4th international workshop on building analysis datasets and gathering experience returns for security (BADGERS)*, IEEE, pp. 25-31, 2015.

[19] Ribeiro, Mauro, Katarina Grolinger, and Miriam AM Capretz, "Mlaas: Machine learning as a service," *In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 896-902, 2015.

[20] Tafti, Ahmad P., Eric LaRose, Jonathan C. Badger, Ross Kleiman, and Peggy Peissig, "Machine learning-as-a-service and its application to medical informatics," *In International Conference on Machine Learning and Data Mining in Pattern Recognition*, Springer, Cham, pp. 206-219, 2017.