Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 7, No. 4, December 2019, pp.1667-1674

Dynamic filtering of malicious records using machine learning integrated databases

Ahmed Abbood Ali[1], Ahmed Raee AL-Mhanawi[2], Aqeel Kamil Kadhim[3]
[1]Directorate General of Education Baghdad- Alkarkh-1
[2]Directorate General of Education AL-Qadisiyah
[3]Al-Imam Al-Kadhum University College for Islamic Science

ABSTRACT

Machine Learning, Deep Learning and Predictive Analytics are the key domains of research in assorted domains of implementations including engineering, finance, economics, real time imaging and many others. The researchers are working on different tools and technologies including open source and own developed frameworks so that the higher degree of accuracy can be achieved. The research reports from Market Research News US predicted that the global market size of machine learning based implementations will exceed 20 billion dollars in year 2024. Most of the government and social services are nowadays in process to be deployed with the advanced technologies of machine learning and deep learning so that the minimum error factor can be there. The key players in the industry include; Google, Facebook, IBM Watson, Baidu, Apple, Microsoft, Wipro, Amazon, Intel, Nuance and many others which are working on the advanced algorithms and implementation perspectives of machine learning.

Keywords: Machine Learning, Malware Analysis, Knowledge Discovery

Corresponding Author:

Ahmed Abbood Ali
Departement, [1]Directorate General of Education Baghdad
University #
Address, Bghdad , Iraq
E-mail: ahmed_swe.@yahoo.com

1. Introduction

The domain of knowledge discovery and predictive analytics is more focused and dependent towards machine learning and deep learning-based applications. Enormous algorithms and methodologies are available in machine learning for scientific applications and solutions for real world problems [1]. Broadly, there are three types of approaches in the machine learning which are widely integrated for the problem solving and predictive mining. These approaches include; supervised learning, unsupervised learning and reinforced learning [2, 3]. These approaches are used as per the specific domain of implementation and accuracy required. The industry of deep learning is very closely associated with machine learning to integrate the higher degree of performance and accuracy with the minimum error rate [4, 5]. The classical applications of machine learning include the following perspectives of Computer Vision and Graphics, Engineering Optimization, Biomedical and Bio-Informatics, Software Engineering and Internet Frauds Detection, Customer Relationship Management, Time Series Forecasting, Data mining and Predictive Mining, Chemical Informatics, Web and Mail filtering, Wireless Network Analytics, Adaptive Web Applications and Analysis, Natural language processing (NLP), Automatic taxonomy construction, Automatic summarization, Grammar Evaluations, Language Analytics, Speech recognition, Handwriting recognition, Optical character recognition, Speech Processing synthesis, Sentiment Data Analysis, Machine Process Automation and translation, Query Execution and Processing, Text mining and simplification, Information Retrieval and Predictive Mining,

Pattern recognition, Optical character recognition, Image recognition, Facial recognition system, Handwriting recognition, Speech recognition, Recommendation system, Content-based filtering, Collaborative filtering, E-Commerce, Hybrid recommender systems, Search engine Optimization, Robot Locomotion, Social Engineering and many others [6, 7, 8].

2. Results

Predictive Analysis on Malicious Records

Following are the prominent tools and software, libraries used for the machine learning and data science-based implementations [9,10].

Table 1. Machine Learning Libraries and Toolkits

CNTK	Apache SystemML	Caffe
Deeplearning4j	ELKI	GNU Octave
H2O	Keras	KNIME
Mahout	Mallet	mlpack
MXNet	OpenNN	Orange
PyTorch	RapidMiner	scikit-learn
Shogun	Spark MLlib	TensorFlow
Theano	Weka	Yooreeka

However, several software libraries are widely used, still, Weka is one of the powerful tools that are used by the researchers and data scientists. Weka is having a huge set of machine learning and data science-based algorithms including big data analytics [11, 12, 13]. Weka can be used with Command Line Interface as well as Graphical User Interface (GUI) for implementation of the algorithms. Besides, the in-built and pre-loaded packages in Weka, are assorted extension packages which can be integrated for the advanced applications.

Predictions using Machine Learning Algorithms in Weka

Here is the example of using classifier-based machine learning algorithm using Weka in which the classification problem of the dataset of network traffic is used. In the following example, the dataset "Malicious_Traffic_Records.arff" is used for training the classifier model. In this example, there is the penetration of network traffic in three phases. Based on the penetration score obtained in sequence, the final class attribute is determined [14, 15].

Malicious_Traffic_Records.arff

@relation Malicious_Traffic_Records
@attribute Network_Traffic_Parameter1 numeric
@attribute Network_Traffic_Parameter2 numeric
@attribute Network_Traffic_Parameter3 numeric
@attribute class {1, 2}
@data
90, 89, 89, 1
89, 90, 98, 1
78, 67, 78, 2
67, 71, 78, 2
69, 78, 78, 2

60, 79, 78, 2

The test dataset "testMalicious_Traffic_Records.arff" is used to predict or determine the class or stream of the Malicious_Traffic_Records who obtained Network_Traffic_Parameter in specific sequence. This problem is hereby solved using Weka with the integration of machine learning algorithm of J48 classifier. In the following dataset, we must determine the classes (streams) of the Malicious_Traffic_Records on the basis of their performance (scores) in the examination.

testMalicious_Traffic_Records.arff

@relation Malicious_Traffic_Records
@attribute Network_Traffic_Parameter1 numeric
@attribute Network_Traffic_Parameter2 numeric
@attribute Network_Traffic_Parameter3 numeric
@attribute class {1, 2}
@data
98, 91, 90, ?
89, 67, 78, ?
78, 67, 78, ?
77, 71, 78, ?
90, 78, 78, ?
10, 10, 10, ?
40, 40, 78, ?
30, 30, 80, ?
98, 97, 94, ?

In the option of malware analysis, there are multiple options including Explorer, Experimenter, Knowledge Flow, Workbench and Simple CI. For traditional implementations, the usage of Explorer is done by the data scientists in which there are user friendly interfaces to choose the dataset and applying different algorithms without cramming any instruction or syntax of the algorithmic implementation [16, 17, 18].

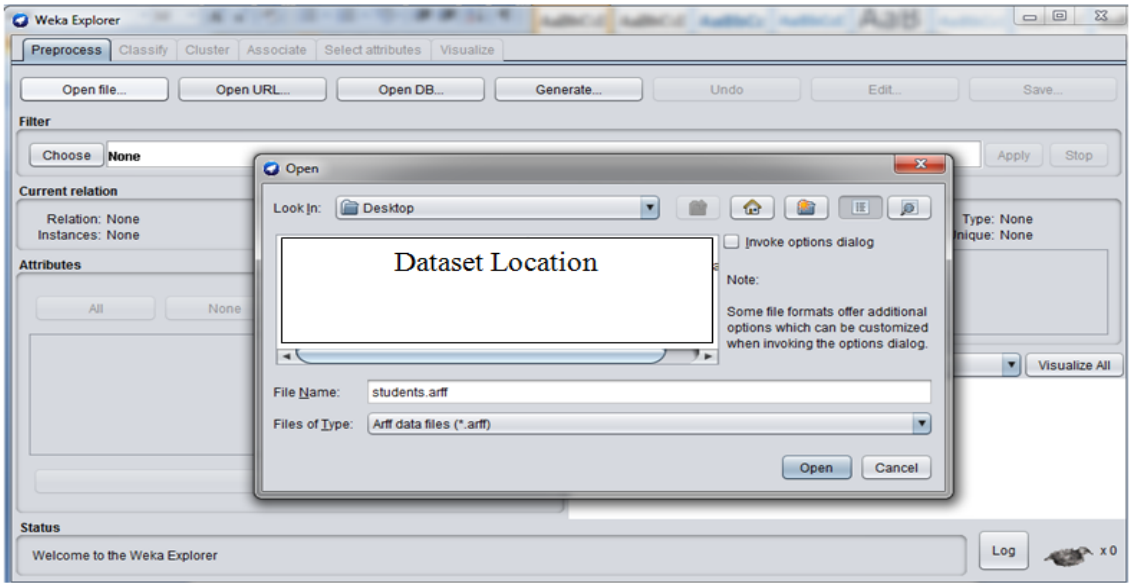

Figure 1. Reading the Training Dataset for Machine Learning

From the Preprocess Tab in Weka as shown in figure, the training data can be selected. In the option to open file, the data scientist can select the dataset that is required to be trained for modeling and processing for the classifier as per the current scenario [19, 20].

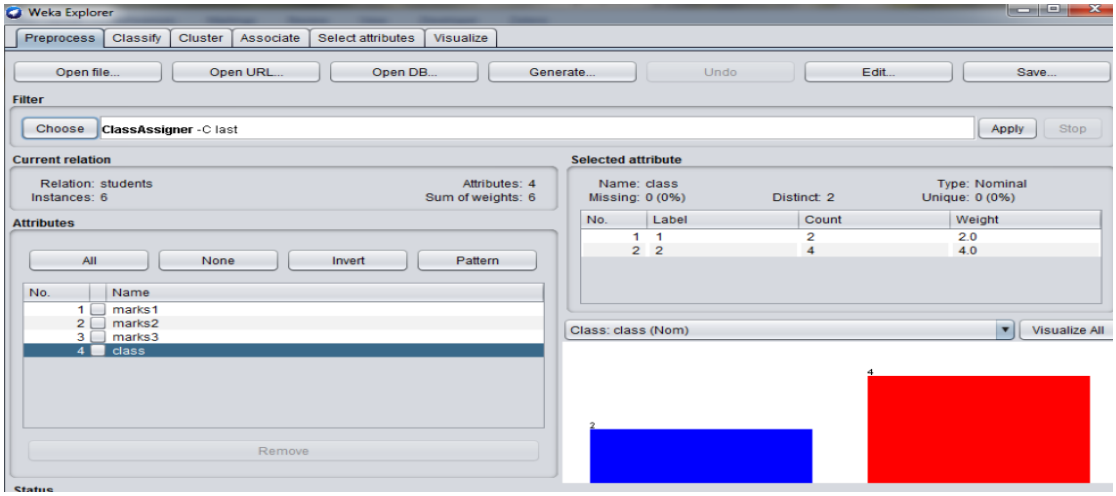

Figure 2. Assigning the Target Class

Once the training dataset is selected and imported to the Weka interface, the target class is required to be mentioned. As per the training data, the attribute "class" is used here as the target. It means that the "class" is the determined value on the combinations and associations of other attributes Network_Traffic_Parameter1, Network_Traffic_Parameter2 and Network_Traffic_Parameter3.

In Weka, there are assorted algorithms for data science and machine learning which can be called and attached with the dataset to be processed. The figure presents the option to select the classification algorithm of J48 so that the classification model can be built based on the training data selected earlier.

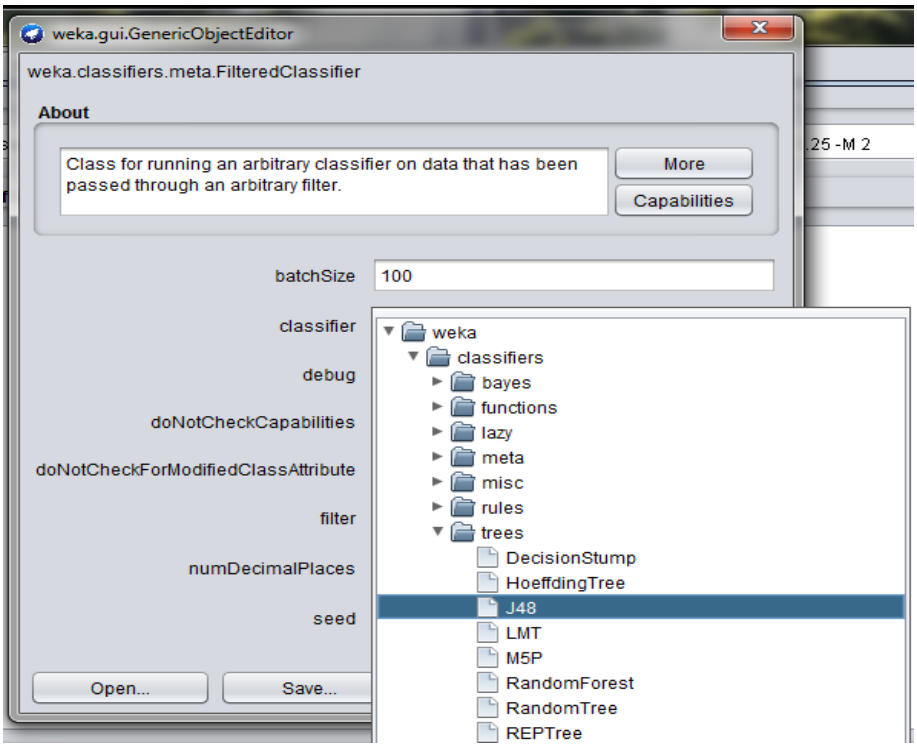

Figure 3. Selection of the Classification Algorithm

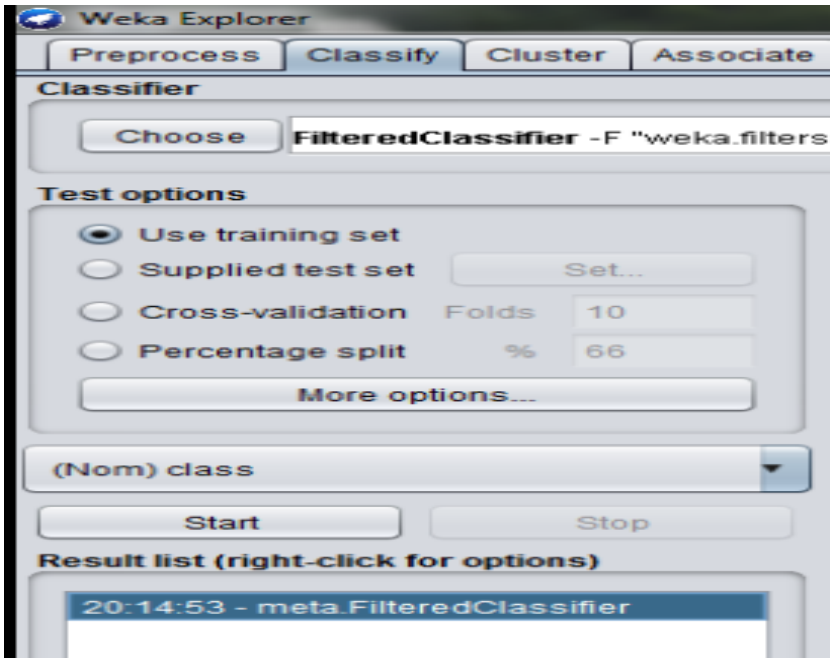

Figure 4. Running the Classifier on Training Dataset

Once the dataset and classification model are invoked, there is a need to run the classifier so that the model can be trained. In this way, the classification model gets fit with the association of determining attributes and the determined attribute. In this example of Malicious_Traffic_Records, the determining attributes are Network_Traffic_Parameter1, Network_Traffic_Parameter2 and Network_Traffic_Parameter3. The determined attribute or target is the class that is having association with the determining attributes or dependent attributes in the training dataset.

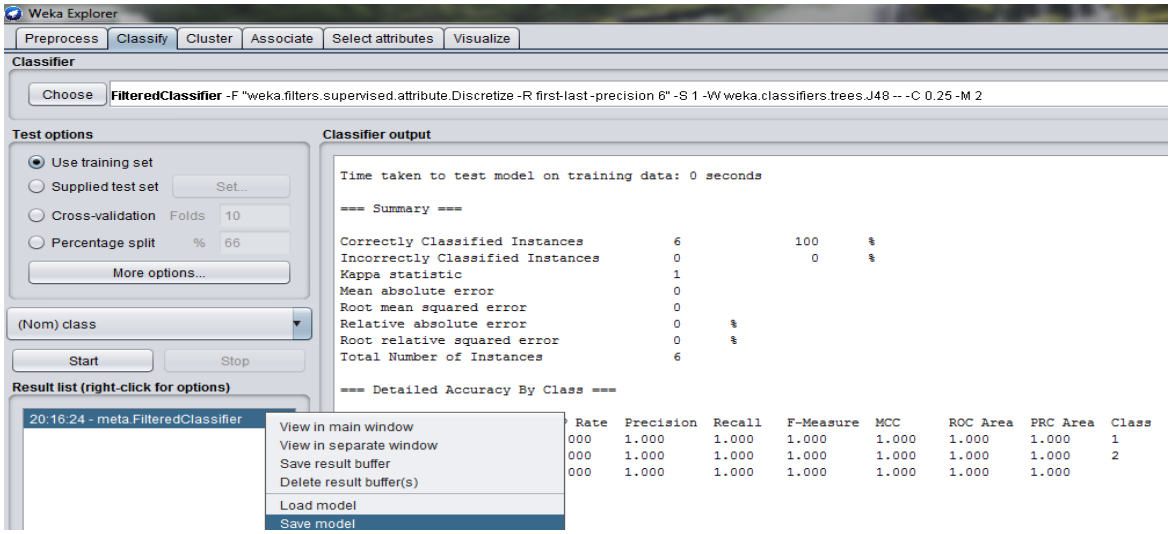

Figure 5. Saving the Model for Prediction of Test Data

The classification algorithm for machine learning is executed and afterward is required to be spared with the goal that the expectation on the testing or approval dataset should be possible. It implies that the prepared model is having the affiliation capacities and the scientific display of the considerable number of characteristics. These scientific capacities of display are additionally required for the

expectation of testing information which are not having the classes. The anticipated class of the testing or approval information is resolved on the capacities made by the prepared arrangement of display, according to the executed algorithm.
For the forecast of the preparation dataset, the test dataset is coordinated with the spared model. The spared grouping model is stacked in the Weka dashboard and after that the choice of "Provided test set" is utilized for testing information. The testing dataset is called utilizing "Set" alternative, so it tends to be anticipated with the spared characterization display.
In the wake of stacking the spared grouping model of AI, the test (approval) dataset is perused with the goal that the obscure classes (spoke to as "?") in the testing information can be anticipated.
To view the predicted classes, the option of “Output Predictions” in the “Classifier Evaluation Options” is set with the “PlainText” so that the unknown classes can be viewed on the Weka interface.

After loading the saved model and invoking the testing dataset, the re-evaluation of the model is done specifically to the supplied test set. It is used so that the supplied testing dataset can be assigned to the predicted classes on the same mathematical functions and model as used in the earlier classification model.

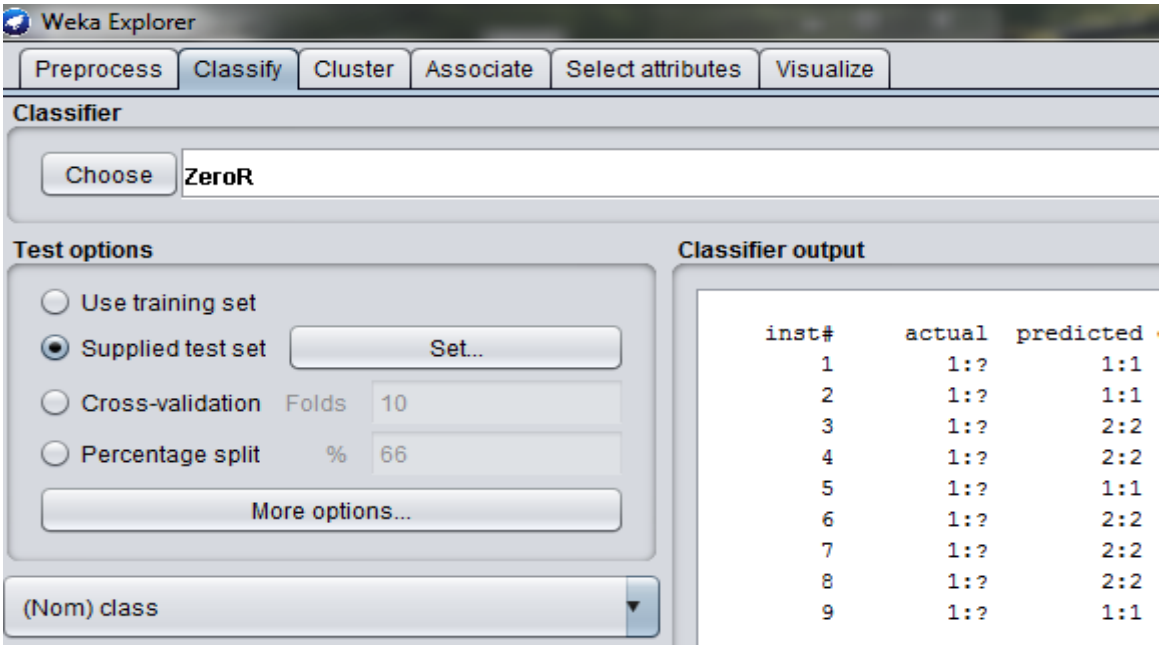

Figure 6. Analysis of the Predicted Classes

The anticipated classes can be seen in the correct sideboard of Weka after re-assessment of the model. As in the figure portrayed above, there are various characteristics including occurrence number, genuine and anticipated. In the anticipated characteristic, the obscure class (spoke to as "?") is alloted with the decided class on a similar calculation of AI. For instance, in first example, the anticipated class is 1 which was obscure in the test dataset. In comparative manner, different qualities can be anticipated.

WekaDeepLearning4j is the devoted bundle for the execution of profound learning in various applications. This library is discharged so the highlights and exactness of profound learning can be

utilized with the information of investigation and prescient mining-based applications. The forces of Java writing computer programs are related at the back end of Weka with profound learning module.

The key highlights and layers are coordinated in the Weka with profound learning incorporates the, Convolution Layer, Dense Layer, Subsampling Layer, Batch Normalization, Long Short Term Memory (LSTM), Global Pooling Layer, Output Layer and these can be utilized for the pernicious records ID.

3. Conclusion

The approaches of AI, machine learning, information science and information disclosure are intently related spaces for the logical and designing applications. These calculations can be utilized for the advancement of new calculations and taking care of the issues of the improvement of the designs in various spaces for the social just as logical areas. These calculations are having custom capacities which can be refreshed according to the necessities of the dynamic datasets to accomplish a higher level of exactness and execution with related components of adequacy.

References

[1] K. Bayoude, Y. Ouassit, S. Ardchir, and M. Azouazi, "How Machine Learning Potentials are transforming the Practice of Digital Marketing: State of the Art," Period. Eng. Nat. Sci., vol. 6, no. 2, pp. 373–379, 2018.

[2] [A. S. Abdullah, M. A. Abed, and I. Al Barazanchi, "Improving face recognition by elman neural network using curvelet transform and HSI color space," *Period. Eng. Nat. Sci.*, vol. 7, no. 2, pp. 430–437, 2019.

[3] I. A. Witten, E. Frank , M. A. Hall, & C.J. Pal . Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann,2016.

[4] V. V. Thendral Tharmalingam, "An Efficient Convolutional Neural Network Based Classifier to Predict Tamil Writer," *Period. Eng. Nat. Sci.*, vol. 6, no. 1, pp. 285–295, 2018.

[5] X. Meng, J. Bradley, B. Yavuz , E. Sparks, S. Venkataraman , D. Liu, et al. Mllib: Machine learning in apache spark. The Journal of Machine Learning Research, 17(1), 1235-1241,2016.

[6] S. Rashid, A. Ahmed, I. Al Barazanchi, and Z. A. Jaaz, "Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set," *Period. Eng. Nat. Sci.*, vol. 7, no. 2, pp. 448–457, 2019.

[7] M. I. Jordan, & T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260,2015.

[8] S. Sra, S. Nowozin, & S. J. Wright. (Eds.). Optimization for machine learning. Mit Press,2012.

[9] T. C. Smith, & E. Frank. Introducing machine learning concepts with WEKA. In Statistical genomics (pp. 353-378). Humana Press, New York, NY,2016.

[10] L. Kotthoff , C. Thornton , H. H. Hoos, F. Hutter, & K. Leyton-Brown. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. The Journal of Machine Learning Research, 18(1), 826-830,2017.

[11] M. Feurer, A. Klein , K. Eggensperger, J. Springenberg, M. Blum, & F. Hutter. Efficient and robust automated machine learning. In Advances in neural information processing systems , pp. 2962-2970,2015.

[12] J. Read, P. Reutemann, B. Pfahringer, & G. Holmes. Meka: a multi-label/multi-target extension to weka. The Journal of Machine Learning Research, 17(1), 667-671,2016.

[13] R.R. Curtin, J. R. Cline, N. P. Slagle, W. B. March, P. Ram, N.A. Mehta, & A.G. Gray. MLPACK: A scalable C++ machine learning library. Journal of Machine Learning Research, 14(Mar), 801-805,2013.

[14] J. Brownlee. Machine learning mastery. URL: http://machinelearningmastery. com/discover-feature-engineering-howtoengineer-features-and-how-to-getgood-at-it,2014.

[15] R. Arora . Comparative analysis of classification algorithms on different datasets using WEKA. International Journal of Computer Applications, 54(13),2012.

[16] T.C. Sharma, & M. Jain. WEKA approach for comparative study of classification algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2(4), 1925-1931,2013.

[17] P. M. Domingos. A few useful things to know about machine learning. Commun. acm, 55(10), 78-87,2012.

[18] A. Desai, & R. Sunil. Analysis of machine learning algorithms using WEKA. International Journal of Computer Applications, 975, 8887,2012.

[19] S. Drazin, S., & M. Montag. Decision tree analysis using weka. Machine Learning-Project II, University of Miami, 1-3,2012.