Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 5, pp. 1009–1019
DOI : 10.47974/JSMS-1299

Exploring the efficiency : A comprehensive analysis of machine learning algorithms in WEKA software

Nikhil Kamble *
Department of Computer Engineering & Information Technology
COEP Technological University
Shivajinagar
Pune 411005
Maharashtra
India

Atharva Phatak †
Department of Biotechnology
Indian Institute of Technology
Madras 600036
Chennai
India

Aaryan Joshi §
Department of Computer Science
Progressive Education Society's
Modern College of Arts, Science and Commerce
Shivajinagar
Pune 411005
Maharashtra
India

Rahul Adhao ‡
School of Computer Engineering
MIT Academy of Engineering
Alandi
Pune 412105
Maharashtra
India

* *E-mail:* kamblenb22.comp@coep.ac.in (Corresponding Author)
† *E-mail:* be21b009@smail.iitm.ac.in
§ *E-mail:* S-2101100@moderncollegepune.ac.in
‡ *E-mail:* rba.comp@coep.ac.in

Vinod Pachghare @
Department of Computer Engineering & Information Technology
COEP Technological University
Shivajinagar
Pune 411005
Maharashtra
India

Abstract

In data science, choosing the right machine learning algorithms is essential for getting the best possible predicted performance. To investigate the effectiveness of different machine learning algorithms, a thorough analysis is conducted within the WEKA software framework. A suitable venue for this investigation is WEKA, a popular platform for machine learning and data mining activities that offers a wide range of algorithms. Our study compares and assesses the effectiveness of several machine learning algorithms on various datasets, taking into account variables like scalability, accuracy, and computing economy. Using a strict approach, analyses for patterns and trends provide insight into the advantages and disadvantages of particular algorithms in different contexts. The research attempted to utilize many machine learning algorithms to determine the accuracy of the dataset after deleting particular fields.

Subject Classification: *68W40 Analysis of algorithms.*

Keywords: *Machine learning, WEKA, Accuracy, Machine learning classifiers, Feature selection, Internet of Things.*

1. Introduction

The selection of algorithms is a critical factor in influencing the outcome of predictive modeling activities as machine learning advances. Given the variety of algorithms that are now available, a careful analysis of their effectiveness and suitability for a range of datasets is required. Such studies benefit greatly from the software package WEKA, which incorporates a wide range of machine-learning methods [1].

The Internet of Things (IoT) era has both potential and problems for machine learning applications due to the huge volume of data generated by an increasing number of networked devices. In this work, the performance of machine learning algorithms on IoT datasets is highlighted through an efficiency analysis conducted within the WEKA software framework. The Internet of Things (IoT) is becoming common in many

@ *E-mail:* **vkp.comp@coep.ac.in**

fields, including smart cities, industrial automation, and healthcare. Therefore, it is critical to know how well machine learning algorithms interpret IoT data.

This research aims to present the comparative evaluation of WEKA classifiers in the context of the IoT Device Recognition dataset. The rapid expansion of Internet of Things (IoT) devices necessitates effective and precise device recognition to maintain network security and optimization. Classification Techniques as a core idea of data mining is the main emphasis of the study. The research uses a numerical dataset with 1000 instances and 298 characteristics.

The research explores the use of several classification algorithms, including J48, LMT, Random Forest, Random Tree, Bayes NET, IBk, LogiBoost, Bagging, Random Committee and PART. A thorough evaluation of each algorithm's performance indicators, including accuracy, recall, and F1-score, is conducted to give a general idea of its advantages and disadvantages. These algorithms are contrasted using the WEKA outcomes of their computations [2].

2. Literature Review

Many studies have been conducted on this topic in the past. WEKA is a machine learning tool that has gained popularity. Several research has done on performance comparison of different classifiers [3]. To expand on the importance of machine learning techniques in supporting medical diagnosis, this study explored the classification effects of several machine learning algorithms on medical datasets. S. Asha Kiranmai and A. Jaya Laxmi analyzed the classification of power quality problems using WEKA and the effect of attributes on classification accuracy in data mining [4].

Their main focus is on classifying power quality issues such as voltage sag, swell, imbalance, and interruption by employing the Random Forest decision trees, Random Tree, and J48 algorithms. Mohd Fauzi bin Othman and Thomas Moh Shan Yau [5], improved the prediction results of evaluating performance in breast cancer. Machine learning techniques were used to improve the prediction outcomes. Specifically, these researchers have used a few methods, which include the trending and efficient machine learning algorithms. Similar research was conducted by Sushil Kumar Kalmegh, who evaluated REPTree, Simple Cart, and Random Tree classification algorithms on the Indian News dataset [6].

The dataset was categorized into 'Business,' 'Criminal,' and 'Education.' 'Medical' and 'Politics' were processed for stop word removal, tokenization, and stemming. The Random Tree algorithm was best as others could classify only the News Table correctly. .The fast development of networking technologies and IoT devices over the past several years led to significant growth in IoT networks. As there are many parts in network data, the machine learning model gives less accurate results. The authors work on reducing the time by reducing features [7]. By calculating Information gain and Gain ratio with Pearson coefficient values for all features, the Author chooses the top features. When compared with every attribute, the proposed approach showed a decrease in computing time with equal accuracy. Sayali A. Dolas, Shitalkumar A. Jain, Avinash N. Bhute, who created the IoT and Q-learning mechanisms to create a safety management architecture for workers in small-scale candy manufacturing businesses [8].

Additionally, it offers the low-cost benefits associated with automation. The different parts of the system are the flame sensor, temperature, humidity, and gas. Utilizing data mining tools and algorithms for association rules mining, this research endeavors to establish a methodology for generating novel rules from an extensive collection of identified rules derived over a four-year period from traffic accident data in Alghat Province, Saudi Arabia. The study conducted by Abdelmoneim Ali Mohamed Hamed and Faisal Mohammed Nafie Ali employs the WEKA software for their investigation [9].The authors highlight two crucial facets of smart health care: communication and monitoring, and they describe how an alarm is sent off to notify emergency personnel and family members of the situation [10]

3. Data Preprocessing

Preparing the data is the first step in the pipeline for machine learning and analysis. To prepare raw data for analysis or machine learning model training, it must be cleaned and transformed. Effective preparation of data can have a major effect on the accuracy and performance of analytical models. First, checked the dataset and added the appropriate missing values. The dataset is already in numeric form, so there is no need to convert it into numeric form.

4. WEKA : A Machine Learning Tool

As a machine learning tool, WEKA provides a base for data mining [11]. In 1997, experts from the University of Waikato in New Zealand created it [3]. This collection of data mining methods was made available to the public under a GPL license. Calculations for association rules, information preprocessing, data preprocessing, visualization, regression, and grouping are carried out by WEKA. It consists of about 76 order classifications, three association rules, and 49 data preprocessing steps for such computations. Modules namely Knowledge Flow, a interface based on java for conducting machine learning tests, and Experimenter, a tool for creating analyses and arranging various systems for testing, are part of the WEKA suite. Additionally, WEKA supports several file formats, including .csv, .arff, and data, to help extract necessary details from unprocessed data.

WEKA often uses ARFF file format for its data files, which are made up of unique tags that identify various elements in the file (most notably, attribute names, attribute types, attribute values, and data) [12]. The Explorer is the primary WEKA interface. It has several panels, each of which has a specific function. After loading a dataset, more analysis is possibly done using one of the Explorer's other panels.

5. Machine Learning Classifications

5.1 *J48*

J48 is a linear method for constantly and categorically examining data. Using a set of training data, this method builds decision trees using the concept of information entropy. Recursively partitioning data based on attribute values, the method generates decision trees by determining the appropriate attribute for splitting based on information gain or gain ratio.

5.2 *LMT*

With the goal of achieving high prediction accuracy and interpretability, Logistic Model Tree integrates the concepts of Decision Tree Learning and Logistic Regression. A Logistic Variant of decision trees, in this case split using C 4.5 after a linear regression model generated at each node by the LogiBoost method. Without the need for any tuning settings, LMT generates a single tree with logistic regression models at the leaves, multiway splits on nominal characteristics, and binary splits on numerical ones.

5.3 *Random Forest*

The output of several decision trees is combined in this ensemble learning technique to produce a single outcome.

An extension of the bagging method, each tree has a unique attribute, variety, and features concerning other trees. It is based on the concept of decision trees and combines the predictions from multiple decision trees to improve accuracy and generalization.

5.4 *Random Tree*

A single tree incorporates randomness in the feature selection process. It is a decision tree in which each leaf contains an optimized linear model for the local subspace that it described. At each node a random subset of features is considered for splitting instead of evaluating all features, thereby creating more diverse trees.

5.5 *Bayes NET*

Graphical Models that representing joint probabilistic relationship based on direct acyclic graphs enabling direct representation of causal relations between variables. The two sets that make up a DAG's structure are the set of directed edges and the set of nodes, or vertices, which stand for random variables.

Bayes Network that models sequences of variables are called dynamic Bayesian networks.

5.6 *IBk*

IBk, also known as Instance-Based k-Nearest Neighbors, works by figuring out which examples are most like the ones we are trying to classify. This algorithm does this by looking at the closest neighbors and seeing which class they belong to. It's handy when dealing with data that's a bit messy or doesn't follow a clear pattern.

5.7 *Logiboost*

Logiboost is a clever tool that helps weak classifiers do better. It gives more attention to the ones that are doing well and learns from them. It is a boosting algorithm that, based on their accuracy, enhances the performance of weak classifiers by assigning weights to them. It employs a logistic regression model to iteratively optimize the combination of weak classifiers, effectively improving predictive accuracy. Logiboost is known for its ability to handle complex relationships in the data.

5.8 *Bagging*

Bagging, also known as Bootstrap Aggregating. Bagging is an ensemble learning method that builds multiple instances of a base classifier by training them on different bootstrap samples of the dataset. The final prediction is determined by averaging or voting across these models, reducing overfitting and improving generalization. BAGGING enhances the stability and reliability of various base classifiers.

5.9 *Random Committee*

Random Committee is an ensemble technique that constructs a committee of diverse classifiers by training them on randomly sampled subsets of the data. The final prediction is often based on the highest vote or averaging of individual classifier outputs.

5.10 *PART*

PART recursively selects the most informative attributes for classification. Unlike traditional decision trees, PART does not always fully expand the tree, allowing it to handle incomplete or missing data effectively. This makes PART a suitable choice for datasets with varying levels of attribute completeness and supports decision-making in scenarios with partially available information.

6. Proposed Model Architecture

The suggested model makes use of the IoT dataset to improve the IoT environment's performance. Data preparation is necessary before using any machine learning approach. Preprocessing of data involves deleting characteristics from the dataset that are redundant or unnecessary, which helps to increase the accuracy as well as speed of the system.

Subsequently, a chosen feature list is used to calculate this outcome. Multiple classifiers were used to conduct the classification using the WEKA tool. The steps in the suggested structure have been displayed in Figure 1.

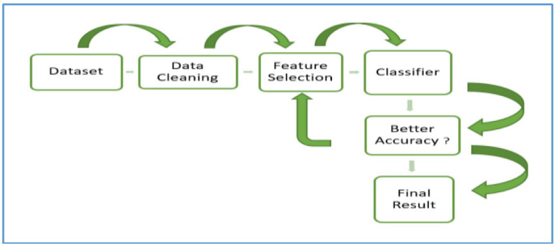

Figure 1

Architecture of the proposed model

7. Experimentation and Result Analysis

7.1 *Experimentations*

The IoT-Device Identification dataset was utilized in the proposed model research. The platform on which the model has been developed consists of Windows 10 Pro and 64-bit OS, 8 GB RAM, and a Core(TM) i5-6300U CPU running at 2.40GHz and 2.50GHz.

To evaluate and investigate the performance that the chosen classification algorithms, such as the J48, LMT, Random Forest, Random Tree, Bayes NET, IBk, LogiBoost, Bagging, Random Committee and PART, followed completely with WEKA's suggested experimental procedure. 75 percent is used for training, while the 25 percent is used for analyzing the result.

The results of the simulation with 297 features (All Features) are shown in Figure 2 and Figure 3 shows the result with 249 features. Accuracy remains constant that is 79.5% for the Bayes NET classification algorithm.

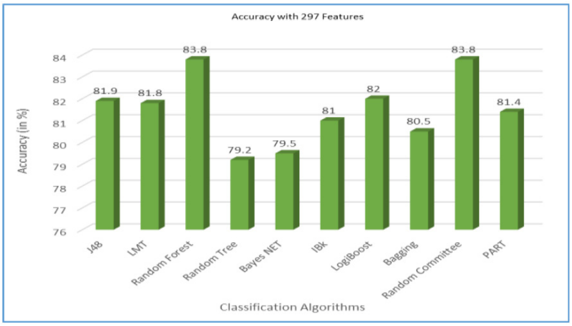

Figure 2

Accuracies of classifiers with 297 features

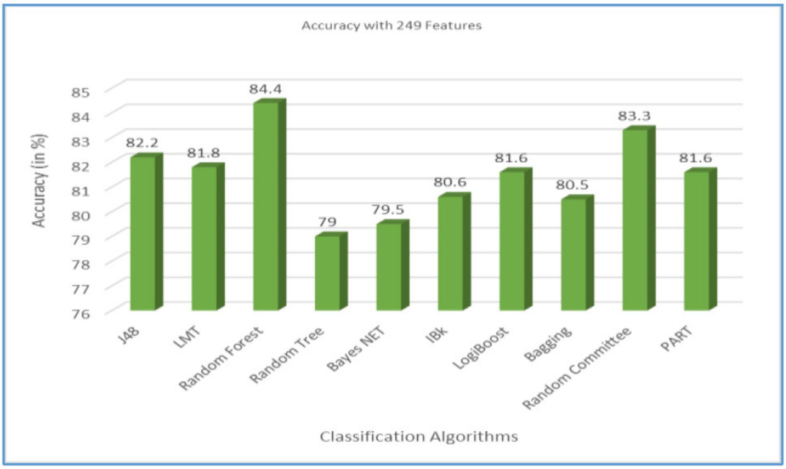

Figure 3

Accuracies of classifiers with 249 features

The research has done on 10 intervals of features from 297 to 248. For these nine-feature sets with an interval of six, simulation has been performed for all above classification algorithms. As features reduce accuracy will increase for most of the classification algorithms.

7.2 *Result Analysis*

The Data obtained from all the simulations is shown in Figure 1, Figure 2 and Table 1. The simulation has been performed for J48, LMT, Random Forest, Random Tree, Bayes NET, IBk, LogiBoost, Bagging, Random Committee and PART. Table I represents the accuracy of different algorithms concerning their feature sets. From the table, we see that the classification algorithm that performed best was Random Forest. It showed the highest accuracy, 85%, with the 6th attribute set (267 Features). The least-performing classification algorithm was Random Tree, showing an accuracy of 79 % with the ninth feature set (297 Features). The accuracy remains constant, which is 79.5% for the Bayes NET classification algorithm.

Maximum variation in accuracies was observed in the Random Committee classification algorithm. As features increase, accuracy increases in all the classifiers and time reduces for simulation. Figure 2 shows the accuracies with the highest number of features, and Figure 3 shows the accuracies with the lowest number of features.

Table 1

Accuracies of classifiers with different Features

	J48	LMT	Random Forest	Random Tree	Bayes NET	IBk	LogiBost	Bagging	Random Committee	Part
297	81.9	81.8	83.8	79.2	79.5	81	82	80.5	83.8	81.4
291	81.9	81.8	84.5	78.6	79.5	81	82	80.5	81.8	81.4
285	81.9	82.1	83.9	79.2	79.5	81	82	80.5	82.8	81.4
279	81.9	82.1	84.4	78.5	79.5	81	82	80.5	83.5	81.4
273	81.9	81.8	84.3	79.6	79.5	81	82	80.5	82.7	81.4
267	81.9	81.8	85	78.6	79.5	81	82	80.5	83.4	81.4.
261	81.9	81.9	84	79.9	79.5	81	81.6	80.5	83.8	81.6
255	82.2	81.7	83.8	79.1	79.5	81	81.6	80.5	83	81.4
249	82.2	81.8	84.4	79	79.5	80.6	81.6	80.5	83.3	81.6

5. Discussion and Conclusion

This work demonstrated an improvement in classification issue accuracy for IoT device identification. When working with enormous data sets, users can have lots of features that are far more complicated. In the field of computer security, feature engineering is valuable because it uses human expertise and domain-specific knowledge to make sure that the data has critical characteristics for IoT device identification classification challenges. The model's performance is evaluated when all features are taken into account and when fewer features are proposed. The proposed framework has demonstrated high accuracy while requiring less computing time. The Author has performed classification with eight classification algorithms with 10 intervals of features. In the future, the authors aim to determine the efficiency of their proposed approach using other IoT datasets that are available.

References

[1] G. Holmes, A. Donkin and I. H. Witten, "WEKA: a machine learning workbench," Proceedings of ANZIIS '94 - Australian New Zealand Intelligent Information Systems Conference, Brisbane, QLD, Australia, pp. 357-361 (1994), doi: 10.1109/ANZIIS.1994.396988.

[2] Desai, Aaditya & Rai, Sunil. Analysis of Machine Learning Algorithms using Weka (2013).

[3] Witten Ian, Hall Mark, Frank Eibe, Holmes Geoffrey, Pfahringer Bernhard, Reutemann Peter. "The WEKA data mining software: An update." SIGKDD Explorations. 11. 10-18 (2009). 10.1145/1656274.1656278.

[4] S. Asha Kiranmai and A. Jaya Laxmi, "Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy", Protection and Control of Modern Power Systems volume 3, Article number: 29 (2018).

[5] Mohd Fauzi bin Othman and Thomas Moh Shan Yau. "Comparison of Different Classification Techniques Using WEKA for Breast Cancer."IFMBE Proceedings book series (IFMBE,volume 15).

[6] Kalmegh, Sushilkumar R."Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News."

[7] Rahul B Adhao, Mayur Tayde, Vinod Pachghare. "Statistical feature selection based intrusion detection system for internet of things environment". AIP Conf. Proc. 2724, 020003 (2023).

[8] Kumar, Ankit, et al. "An enhanced quantum key distribution protocol for security authentication." *Journal of Discrete Mathematical Sciences and Cryptography* 22.4 : 499-507 (2019).

[9] Poonia, R. C., & Raja, L. Smart Technologies in Engineering. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 13(6), 1172-1172 (2020).

[10] Andrews, Leo John Baptist, Shanmugasundaram, Suresh, Kumar, V. Sampath & Alagappan, Annamalai. Application of Internet of Things - A smart healthcare concept for Botswana, *TARU Journal of Sustainable Technologies and Computing*, 1:2, 3 & 4, 93-103 (2019), DOI: 10.47974/2019.TJSTC.004

[11] Witten, Ian & Hall, Mark & Frank, Eibe & Holmes, Geoffrey & Pfahringer, Bernhard & Reutemann, Peter. The WEKA data mining software: An update. SIGKDD Explorations. 11. 10-18 (2009). 10.1145/1656274.1656278.

[12] Eshwari Girish Kulkarni and Raj B. Kulkarni, "WEKA Powerful Tool in Data Mining", *International Journal of Computer Applications* (0975 – 8887) National Seminar on Recent Trends in Data Mining (RTDM 2016).

Received May, 2024