

Performance Analysis of Classification Algorithms in Big Data Environments

1. Problem Statement

The accelerating development of big data in various fields, including social media, smart cities, healthcare, and network security, has led to the fact that the efficient classification algorithms are required which can work with large and complicated datasets. Different machine learning algorithms like Decision Trees, Random Forest, and Support Vector Machines have been extensively applied to classification problems because they are easy to understand and have comparatively less computational demands. Nevertheless, as the volume, velocity and quantity of big data increases, these conventional methods might have restrictions on scalability and processing speed.

Recently, the research on the usage of advanced machine learning and deep learning models to enhance the quality of classification in a large data setting has been examined. Indicatively, Samson (2024) contrasted classical machine learning algorithms with deep learning models and discovered that in highly nonlinear and unstructured data, deep learning models are more effective and use large amounts of computing power and hardware acceleration. On the same note, a study by Punia et al. (2021) has established that machine learning models can efficiently handle big amounts of unstructured data gathered with the help of social media platforms, though their computing capabilities are highly reliant on feature choices and preprocessing strategies.

In addition, computing systems like the Apache Spark that are distributed have also been developed to enhance effectiveness and scale to big data analytics. Junaid et al. (2022) demonstrated that Spark ML-lib is a better platform in terms of scalability than traditional machine learning platforms based on single nodes when dealing with large amounts of data. Although such improvements are made, there are still some challenges. Most studies being conducted currently are mainly concerned with the accuracy of classification and lack other

important considerations including scalability, computational cost, resource consumption, and interpretability of the model.

Moreover, other studies also emphasize that no single classification algorithm can fit best any kind of data and area of application (Ba'abbad et al., 2021). According to other works, ensemble models like Random Forest and Gradient Boosting are effective in processing complex and imbalanced data (Bulavas et al., 2021). Nevertheless, no detailed analysis systems are yet available to contrast various classification algorithms in the circumstances of the real big data with similar standards.

Thus, there is an obvious necessity of additional studies that will conduct a systematic comparison of the various classification algorithms in the context of big data considering not only the accuracy of prediction but also factors such as scalability, efficiency, and the amount of used resources. The analysis will enable researchers and practitioners to choose the most suitable algorithms to be used in the large-scale data processing application and enhance decision-making in big data systems.

2. Research Objectives

This study intends

- to provide the analysis and evaluation of the performance of various classification algorithms within the big data settings.

The objectives of the study in particular are:

- To examine the nature and issues of the big data environments when running a machine learning classification task.
- To compare the performance of various classification algorithms Decision Trees, Random Forest, Support Vector Machines, Neural Networks.
- To assess the scalability and the computational efficiency of classification algorithms using large-scale datasets.
- To determine the most appropriate classification algorithm to be used in processing structured and unstructured big data.

- To suggest suggestions as to the choice of effective classification models in big data analytics applications.

3. Literature Review

Paper 1: Comparative Performance Analysis of Regression and Classification Algorithms in Big Data Environments (Samson, 2024)

In this research paper, the comparative performance of regression and classification algorithms in big data set-ups is drawn through a holistic comparison of scalability, predictive accuracy and computational efficiency. The study compares the classic machine learning models, including Linear Regression, Random Forest and Support Vector Machines with deep-learning models, including Convolutional Neural Network (CNN) and Deep Neural Network (DNN). The simulation of big data conditions in reality on high-dimensional datasets with different computational conditions was performed. RMSE, MAE, Accuracy and F1-score were used as performance measures in order to evaluate the effectiveness of the model. The results show that classic algorithms are very efficient in the case of structured data with medium dimensionality, whereas deep learning models are much more efficient in nonstructured and highly nonlinear contexts. Nevertheless, deep learning applications demand high memory and graphics card acceleration systems like TensorFlow. The paper highlights the significance of explainable and hybrid AI models to achieve the trade-off between accuracy, scalability, and interpretability in industrial applications that require data in large volumes.

Paper 2: Performance Analysis of Machine Learning Algorithms for Big Data Classification (Punia et al., 2021)

The study examines the performance of different machine learning classification algorithms in processing and analyzing unstructured big data in large-scale that is collected through the use of social media like Twitter. The experiment applies various supervised and unsupervised learning methods that are Decision Trees, Neural Networks, Support Vector Machines, Naive Bayes, Linear Regression, and K-Nearest Neighbors. The main aim is to convert unstructured data into structured data that has the capacity to support predictive analytics. As the experimental findings indicate, machine learning algorithms are very effective in structuring data and detecting patterns

where conventional data processing methods fail. Nonetheless, the performance of the algorithm differs in relation to feature selection, preprocessing methods, and size of the datasets. Although the study can be useful in a comparative analysis of classification models, it lacks sufficient evaluation of the scalability of distributed computing or hardware resource consumption, which are important in the context of real big data ecosystems. The results emphasize the need to choose algorithms considering the properties of the data and not basing it on the accuracy metrics.

Paper 3: Performance Evaluation of Data-driven Intelligent Algorithms for Big Data Ecosystem (Junaid et al., 2022)

This paper measures the performance of machine learning algorithms on the distributed big data setup with Apache Spark ML-lib and compares it with RapidMiner and Scikit-learn. This study compares classification algorithms, which are Logistic Classifier, Decision Tree Classifier, Random Forest Classifier, and Gradient Boosted Tree Classifier, and regression models, which are evaluated on large benchmark datasets, e.g. SUSY, HIGGS, and HEPMASS. The results show that Apache Spark ML-lib has better scalability and distributed processing features in comparison with one-node machine learning platforms. Random Forest and Gradient Boosted Trees were the best ensemble-based methods, with a high level of prediction performance on any dataset. Nonetheless, the use of computational resources and the time to execute the data also went up with the size of the data. The paper has highlighted the significance of distributed systems to manage big datasets of analytics and real-time data streams. Although the research has good scalability evaluation, it does not discuss much on the model interpretability and the optimization of energy efficient computation in big data ecosystems.

Paper 4: A Comparative Study of Traffic Classification Techniques for Smart City Networks (AlZoman & Alenazi, 2021)

This paper examines the use of supervised machine learning algorithms to identify traffic in smart city network to improve Quality of Service (QoS) management. The proposed study also compares the Support Vector Machines, random forest, K-Nearest Neighbors, and decision tree algorithms with that of the classical port-based classification techniques. The experimental outcomes show

that, the decision tree algorithm had the highest average accuracy of 99.18 and better than other models and by far higher than conventional port-based methods. The methods based on machine learning showed a better ability to adapt to encrypted and dynamically assigned traffic in the network. Nevertheless, the research concentrates on accuracy and does not analyze much the performance in extreme big data scale situation and distributed computing frameworks. Also, the efficiency of real-time processing and resource utilization was not properly examined. The results indicate that machine learning is highly effective in improving the classification in network settings, but still needs to be assessed on large-scale infrastructures of the big data.

Paper 5: A Short Review of Classification Algorithms Accuracy for Data Prediction in Data Mining Applications (Ba'abbad et al., 2021)

The paper is a review of twenty research papers discussing the accuracy of classification algorithms in different fields of application such as business, agriculture, health, and education. The algorithms compared in the study include Random Forest, Naive Bayes, Decision Trees (C4.5) and Feed-Forward Backpropagation Neural Networks (FFBPN). Findings suggest that the performance of Algorithms is very heavily contextual with the Random Forest having better results in online social network data, Naive Bayes being effective in data classification of agriculture and C4.5 Decision Trees being effective in predicting data in an educational record. The paper finds that there is no best classification algorithm that can be used in all big data scenarios. Rather, the choice of algorithm must be informed by the structure of the dataset, the complexity of the features and requirements of the domain. Although the review gives important information on comparative accuracy, it fails to discuss the issue of scalability of computations, distributed processing models, and performance in big data situations of large scale under real-time conditions.

Paper 6: Study of Multi-Class Classification Algorithms' Performance on Highly Imbalanced Network Intrusion Datasets (Bulavas et al., 2021)

This paper explores the behaviour of multi-classes of classification algorithms on highly imbalanced network intrusion data; in this case, the difficulty lies in identifying low-rate malicious traffic classes. The relative ratio of benign traffic to malicious traffic in actual network scenarios

is vastly greater than what is achievable in real-world classifier training conditions, and this leads to extreme problems of class imbalance, which are undesirable in terms of classifier training. Unlike standard methods that combine rare classes of attacks into one binary classification problem, the research study adheres to the multi-class design and seeks to improve rare class detection. The authors used the SMOTE over-sampling algorithm to equalize the classes with low frequencies and performed experiments on three benchmark datasets CIC-IDS2017, CSE-CIC-IDS2018 and LITNET-2020. Other machine learning algorithms were tested and much hyperparameter tuning and feature selection methods used. The performance measures were Precision, Balanced Accuracy Score, G-mean and biasvariance decomposition. Findings show that the ensemble-based decision tree algorithms, especially AdaBoost, Random Forest, and Gradient Boosting, were more effective in detecting the rare intrusion classes and showed high levels of robustness in imbalanced big data classification tasks.

Critical Analysis

The analyzed articles indicate that the efficiency of classification algorithms in big data settings is predetermined by a number of key factors:

1. Data Characteristics:

Traditional machine learning models are effective with structured datasets, whereas deep learning architectures are effective with unstructured or high-dimensional datasets.

2. Scalability:

Apache Spark is a distributed platform that enhances the performance of processing big data ecosystems.

3. Computational Efficiency:

Deep learning models are more accurate but need the acceleration of the graphics card and the large amount of computational resources.

4. Selection Trade-off: There are trade-offs in the selection of algorithms.

- Random Forest and Gradient Boosting are robust and stable.

- Decision Trees are very interpretable with competitive accuracy.
- Deep Neural Networks are better at nonlinear problems but make the process less explainable.

One of the consistent weaknesses in the studies is lack of a standardized benchmarking and absence of real time experimentation of big data.

Relations amongst Studies.

1. The need to have scalable classification models in big data environment is highlighted across all studies.
2. Random Forest and Gradient Boosting are two major models with strong baseline performance as mentioned in most studies.
3. Distributed computing systems (Spark, GPU acceleration) are important to enhance performance.
4. Deep learning algorithms always prove to be significantly better than traditional algorithms in nonlinear or unstructured data sets.
5. The speed of an algorithm greatly relies on the field and the data.

Conflicts Between Findings

1. Conventional vs Deep Learning Models:

Other reports suggest that traditional algorithms are as accurate as deep learning models on structured data, whereas others state that deep learning is much more accurate.

2. Precision vs Irrelevance:

Although deep learning is more accurate, multiple studies contend that interpretable models such as Decision Trees are more important in practice.

3. Single vs Distributed Platforms:

Other studies are concerned with the accuracy of algorithms, but not on distributed scalability, and other studies emphasize scalability as the main evaluation criterion.

4. Domain Dependency:

Some researchers believe that there is no algorithm that is universally better than the others, but some of them suggest certain ones as generally high-performing.

Synthesis Matrix for Literature Review

Study	Authors	Year	Research Focus	Algorithms Evaluated	Datasets / Environment	Evaluation Metrics	Key Findings	Limitations
Comparative Performance Analysis of Regression and Classification Algorithms in Big Data Environments	Samsan	2024	Comparative performance of regression & classification models in big data	LR, RF, SVM, CNN, DNN	High-dimensional big data environments, GPU-supported frameworks	RMSE, MAE, Accuracy, F1-score	Deep learning models outperform in nonlinear/unstructured data; traditional ML efficient for structured datasets	High computational cost; limited interpretability of deep models
Performance Analysis of ML	Punia et al.	2021	Classification of unstructured	DT, NN, SVM, NB, LR, KNN	Twitter-based large-	Accuracy comparison	ML improves structuring of unstructured data;	Limited scalability analysis;

Algorithms for Big Data Classification			social media big data		scale datasets		performance varies by preprocessing	lacks distributed evaluation
Performance Evaluation of Data-driven Intelligent Algorithms for Big Data Ecosystem	Junaid et al.	2022	Evaluation of ML algorithms on distributed platforms	LC, DT, RF, GBTC, LR, RFR	Apache Spark ML-lib vs RapidMiner vs Sklearn; SUSY, HIGGS, HEPMASS datasets	Accuracy, execution time, scalability	Spark ML-lib shows superior scalability; ensemble methods perform strongly	Increased resource consumption with larger datasets
Traffic Classification Techniques for Smart City Networks	AlZoman & Alenazi	2021	ML-based traffic classification for smart city networks	SVM, RF, KNN, DT	Smart city network traffic datasets	Accuracy	Decision Tree achieved highest accuracy (99.18%); ML outperforms port-based methods	Limited big data scalability evaluation

Classification Algorithms Accuracy for Data Prediction	Ba'abbad et al.	2021	Review of classification algorithms across domains	RF, NB, C4.5, FFBN	Multi-domain datasets (business, agriculture, health, education)	Accuracy	Algorithm performance is domain-dependent; no universal best model	Focuses mainly on accuracy; lacks scalability analysis
Multi-Class Classification on Imbalanced Intrusion Datasets	Bulavas et al.	2021	Rare class detection in imbalanced network intrusion datasets	AdaBoost, RF, Gradient Boosting, others	CIC-IDS2017, CSE-CIC-IDS2018, LITNET-2020	Precision, Balanced Accuracy, G-mean, Bias-Variance	Ensemble decision tree models perform best for rare class detection	Requires oversampling (SMOTE); computational complexity

References

Ba'abbad, I., Althubiti, T., Alharbi, A., Alfarsi, K., & Rasheed, S. (2021). A short review of classification algorithms accuracy for data prediction in data mining applications. *Journal of Data Analysis and Information Processing*, 9(3).

Bulavas, V., Marcinkevičius, V., & Rumiński, J. (2021). Study of multi-class classification algorithms' performance on highly imbalanced network intrusion datasets. *Informatica*, 32(3).

Junaid, M., Ali, S., Siddiqui, I. F., Nam, C., Qureshi, N. M. F., Kim, J., & Shin, D. R. (2022). Performance evaluation of data-driven intelligent algorithms for big data ecosystem. *Wireless Personal Communications*, 126(3), 2403–2423.

Punia, S. K., Kumar, M., Stephan, T., Deverajan, G. G., & others. (2021). Performance analysis of machine learning algorithms for big data classification: ML and AI-based algorithms for big data analysis. *International Journal of E-Health and Medical Communications*, 12(4), 60–75.

Samson, F. (2024). Comparative performance analysis of regression and classification algorithms in big data environments. *University of Ibadan*.

AlZoman, R. M., & Alenazi, M. J. F. (2021). A comparative study of traffic classification techniques for smart city networks. *Sensors*, 21(14), 4677.