International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

Comparative Study of Binary Classification Algorithms to Analyze the Students' Performance on Virtual Machine

Rajathi .M[1], Dr. Ramaswami Murugesh[2]

[1]Ph.D Scholar, Department of Computer Application, Madurai Kamaraj University, Madurai, rajisadhu05[at]gmail.com
[2]Professor, Department of Computer Application, Madurai Kamaraj University, Madurai, mrswami123[at]gmail.com

Abstract: ***In recent times, the utilization of learning management systems in education has been increased to a great extent. In order to access online contents, students have started to use mobile phones, especially smart phones. Student's online activities generate immeasurable amount of data that cannot be processed by traditional tools and techniques. This has resulted in the invasion of Big Data technologies and tools into education, to process the enormous amount of data involved. The large volume of data collected for many years contains hidden knowledge, which helps to improve the students' performance. The big data collected from institutions are analyzed in a distributed environment in order to maintain the efficiency and reduce the computational complexity. Machine learning algorithms are used to find meaningful patterns that are hidden in the data. In this paper, machine learning algorithms such as Decision Tree, Naïve Bayes and NBTree are analyzed in both local machine as well as virtual machine on higher secondary school students' data. The data set consists of 17 features and 115328 records. The analysis is based on the parameters such as accuracy, execution time and memory usage. The outcome of our study shows that the NBTree algorithm gives the highest accuracy of 98% with execution time of 0.234 seconds and 49.42 MB of memory usage in local mode. Similarly, the highest accuracy of 99% was achieved with execution time of 0.184 seconds, 45.42 MB of memory usage in virtual mode.***

Keywords: Big Data, Decision Tree, Naïve Bayes, NBTree, Binary Classification, Students Performance Prediction, AWS.

1. Introduction

Educational Data Mining (EDM), is an area that is emerged not only for making findings within the data generated from educational settings but also used to understand students and settings effectively. EDM has developed as an important field of active research in the recent past [1]. One of the key areas of applications of EDM is to improve the student's academic performances in schools, colleges and other educational institutions through the predictions obtained from the models. Basically, the prediction of student's performance with high accuracy is helpful to identify slow learners and to distinguish slow academic achieves or weak students. The developed model is favourable to the teachers, parents and educational planners to improve their student's performance level. It also provides substantial solutions to solve their future problems. The academic results of higher secondary school education are a turning point for a student, as it bridges high school and higher secondary education. But then factors such as demographic, academic and socio-economic restrict the students' performance [2]. It necessitates a need for predicting the academic performance of students at plus two examinations.

The growth of academic data size in school education is being increased rapidly. The knowledge is hidden in the huge volume of data collected for many years and has a huge potential to improve the quality of education and student's performance. In the information era, a huge quantity of data is in hand for decision makers where big data analysis plays an important role. Big data refers to data file that is high in volume, variety and velocity that is tedious to handle using traditional tools and techniques. Due to the rapid growing data set, extracting value and knowledge provides suitable solutions for these datasets. Furthermore, decision makers required to gain valuable insights from these varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. The value provided by big data analytics is the application of advanced techniques [3].

Large datasets can be evaluated by establishing physical distributed environment or by hiring cloud based distributed environment. The cloud-based environment favors scalable virtual resources on demand and thus makes it suitable for handling high volume of data. Due to the large volume of data and the complexity of analysis of the data, the analysis of huge data in the distributed environment is necessary [4]. Distributed computing environment processes massive amount of data by partitioning it, processes it in parallel and by merging the results. Cloud computing services provide storage and computing resources those can be hired as per the requirements. An interactive tool is required to demonstrate data mining results to the user and receive user feedback. Data visualization tools facilitate better understanding of the analysis given by machine learning tools. Web services of Amazon provide scalable computational facilities. The work of creating and maintaining complex infrastructure is handled by Amazon Web Services (AWS). Thus, user or developer can concentrate on design part of the application. Cloud service provided by AWS allows user to reduce the cost incurred for physical infrastructure. Virtual environment can be established with the help of AWS as per the computational needs [5].

Machine learning algorithms [6] summarize large datasets and create machine learning models or classifiers from the data, and thus discover implausible about the data that is used for prediction. The dataset used in this research study is higher secondary school data from Madurai district for three years (2016, 2017, and 2018). The dataset contains 17

attributes of 1,15,328 records. This dataset contains the output attribute as a binary class. The supervised machine learning algorithms such as decision trees and Naïve Bayes are considered suitable for the analysis. In this paper, comparative study of classification algorithms such as Decision Tree [7], Naïve Bayes [8] and NBTree [9] are applied on massive dataset.

2. Literature Study

In fields such as customer service, fraud detection, and business intelligence, artificial intelligence and machine learning are increasingly finding their way into enterprise applications. Machine learning, however, needs a lot of hardware and software infrastructure. All has been made much simpler by the success and development of cloud services.

Aijila and Bankole [10] attempted to implement three machine learning algorithms – neural network, support vector machine and linear regression for cloud-based prediction models for the Amazon EC2 TCP-W benchmark framework. They evaluated the performance with two measures – response time and through put. Their results revealed that SVM offers the best predictive model to provide a more robust scaling decision option for the client.

Wu et al. [11] developed a novel approach to machinery prognostics using a cloud-based parallel machine learning algorithm to predict tool wear in dry milling operations. A parallel random forest algorithm was developed using the MapReduce framework implemented on Amazon EC2. Higher speed was achieved via this parallel algorithm.

In the financial industry, the application of machine learning is growing day by day in order to predict the financial aspects of the organization, such as income based on the assets and resources that they carry from time to time within the organization. Sriramakrishnan Chandrasekaran [12] proposed various regression models to forecast the income of the company and implemented on AWS EC2 and set up an instance S3 to store financial data.

Cai et al. [13] deployed various machine learning algorithms on four different cloud platforms with a very large data collection. Using running time, tuning specifications and ease-of-programming on Amazon EC2, they concluded that Spark+Python were the most appealing and simpler for large applications.Hafez et al. [14] conducted research study to compare the efficacy of various machine learning algorithms with varying data size on Apache-Spark platform. An accuracy and training time were considered as parameters for the comparison. The outcome of the study showed that for marketing dataset, decision tree as the most efficient algorithm. The maximum prediction accuracy for packing and statistical dataset is provided by logistic regression algorithm.

Various advantages can be granted when cloud computing technology is implemented in educational institutions to provide scalable and versatile IT services. The main advantages of cloud computing in education can be to track student learning skills in wider student enrolments by implementing machine learning algorithms.

3. Big Data

Big data [15] is information having high volume, velocity and variety that demand cost-effective, innovative forms of information processing for intensified insight and decision making. Three main aspects which characterize big data are volume, variety, and velocity, or the three V's. The volume of the data is its size and how huge it is. Velocity refers to the rate with which data is changing or how often it is created. Finally, variety includes the different formats and types of data, as well as the various kinds of uses and ways of analyzing the data. The central attribute of big data is its volume. Big data can be evaluated by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Moreover, data from different sources including logs, click streams and social media make big data really big. The structured data is now fasten with unstructured data, such as extensible Markup Language (XML) or Rich Site Summary (RSS) feeds. The data which is received from audio, video and other devices are hard for analytics purpose. Additionally, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, big data, also holds variety as big as volume.

Moreover, speed or velocity is one of the main property through which big data can be explained. This is primarily the amount of data generation or the amount of data delivery. The major edge of big data is streaming data, which is collected in real-time from the websites. Some researchers and systems have discussed the addition of a fourth V or veracity. Veracity focuses on the quality of the data. This identifies the quality of the collected big data as good, bad or undefined due to data inconsistency, incompleteness, ambiguity, latency and approximations.

Big Data refers to any data that is huge and critical that conventional tools and techniques are not adequate to process them. Examples of Big Data include the amount of streaming data on the online resources such as YouTube videos viewed, twitter feeds and mobile phone location data. The main challenges to be addressed by big data are Storage, analysis and reporting.

a) *Storage*
   Nowadays the quantity of data produced from the internet is in the order of exabytes, but the volume that hard disks can hold is in the range of terabytes. Comparatively the data generated in education is much less to internet data, it may increase in future. The traditional techniques like SQL based queries are unable to store or process Big Data and thus non-SQL approach is followed in our study.
b) *Analysis*
   Different Online learning sites produce data with various structure and size. The data analytics may consume a lot of time and resources. To overcome this, scaled down architectures are employed to process the data in a distributed manner. Big data is chunked into smaller pieces and processed in a wide number of computers

connected throughout the network and the processed data is aggregated.

c) *Reporting*
Traditional reports involve display of statistical data in the form of numbers. When the size of data is large, traditional reports become more difficult to interpret by human beings. In such cases, the format used to view should be easily recognizable. To handle these challenges, innovative technologies and techniques will assist individuals and organizations to integrate, analyze and visualize different types of data.

Distributed Computing together with management and parallel processing principle allow acquiring and analyzing Big Data and makes the Big Data Analytics a reality. Different aspects of the distributed computing paradigm resolve different types of challenges involved in Big Data Analytics. The mechanisms related to data storage, data access, data transfer, visualization and predictive modeling using distributed processing in multiple low cost machines are the key considerations that make big data analytics possible within stipulated cost and time practical for consumption by human and machines.

4. Big data with Machine Learning

Nowadays people just don't want to collect data, they need to understand the meaning and importance of the data and use it to aid them in decision making process. Data analytics is the process of applying algorithms in order to analyze sets of data and discover unknown patterns and extract useful information. Classification is one of the most important data analytics method. In this work, classification algorithms such as Decision tree, naïve bayes and NBTree algorithms are used to find the performance of students in their higher secondary exam results. There are two main steps in data classification, namely learning step and classification step. In learning step, a classification model is built using an algorithm on a training set. Training set used for learning step must have class labels for given data. After a classifier model is built, it is utilized for predicting class labels for test data.

Decision tree is a flow-chart-like tree structure. The internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution. A decision tree consists of nodes. Each node represents specific information. Decision tree learning is started from root node and discrete values are produced at each node by testing the values of attribute. These discrete values act as target function. Then by using target function, value of attribute for next node is evaluated. This process is repeated for each new node. The learned tree is represented by if-then rules. Decision tree algorithms can be applied on large amount of data and valuable predictions can be produced. These predictions evaluate future behavior of problem.

Naive Bayes is a simple multiclass classification algorithm based on the application of Bayes' theorem. Each instance of the problem is represented as a feature vector, and it is assumed that the value of each feature is independent of the value of any other feature. One of the advantages of this algorithm is that it can be trained very efficiently as it needs only a single pass to the training data. Initially, the conditional probability distribution of each feature given class is computed, and then Bayes' theorem is applied to predict the class label of an instance.

BTree [16] is one of classification methods that was introduced by Ron Kohavi. It is a hybrid algorithm from Naïve Bayes and decision tree combined. This algorithm is similar with decision tree except in its leaf. The decision tree has a branch from recursive process, and the leaves are from the Naïve Bayes classifier, not a node that contains final result from a class. For continuous attributes, a threshold is chosen so as to limit the entropy measure. The utility of a node is evaluated by discretizing the data and computing the fivefold cross-validation accuracy estimation using Naive Bayes at the node. The utility of the split is the weighted sum of utility of the nodes and this depends on the number of instances that go through that node. The NBTree algorithm tries to approximate whether the generalization accuracy of Naive Bayes at each leaf is higher than a single Naive Bayes classifier at the current node. For discrete valued attributes, the Naive Bayes method performs quite well. With the increase in data size, the performance also improves.

5. Data Source

The data source is collected from the office of the Chief Educational Officer (CEO) in Madurai District. The data is compiled by the demographic, geographic and academic factors which influence the students' performance in their studies. A total of 1, 30,408 records were collected for three years (2016, 2017 & 2018) for Madurai District. Before the classification models were constructed the data was gone through the preprocessing stage to make it ready for the analysis. The dataset contained valueless attributes, missing instances, inadequate attributes data types and other problems that necessitated the need for preprocessing stage before feeding it into the analysis phase. Therefore, the dataset was cleaned, encoded and the missing values were imputed. After cleaning, we had 1, 15,328 records for model construction. An important step in constructing a model was selecting the features for classification. Here 17 features were considered for classification model construction. Year, Taluk name, School location, Type of school, Category of school, Age, Community, Sex, Group, Marks in 6 subjects, Total and Result are the attributes used in this classification.

The class variable is Result – it depends upon the marks obtained by students in their higher secondary exam. Performance prediction models are studied by varying class variable as two (Pass, Fail). In this two-case problem, labeling the class values is fixed based on the marks obtained by the students. Pass is for students with 40% and above and Fail is students for students with below 40%. The model is constructed after encoding the categorical values of all predictor variables into numeric values.

6. Research Experiments and Results

The virtual machine is created on AWS cloud [17] and machine learning algorithms decision tree, naïve bayes and

NBTree are executed in python. Amazon Elastic Compute Cloud (EC2) [18] has been used here to create virtual machine. Amazon Ubuntu AMI (Amazon Machine Image) based SSD (Solid State Driver) volume type is also created to support the better I/O performance. T2.micro instance is chosen for this ubuntu machine having 1 vCPU and 1024 MiB memory. A keypair is created to login into the created instance using puttygen. Putty software is used to connect to the AWS instance using the private key generated.

The data used in this paper is higher secondary school result data of Madurai District. It consists of 17 attributes and 1,15,328 instances. The class attribute is binary (Pass or Fail). The analysis is done on ubuntu machine using the machine learning algorithms in standalone mode and in virtual cloud environment.

The classification performance of Decision Tree, Naïve Bayes and NBTree are compared on the basis of parameters such as predictive accuracy, execution time and memory used. The predictive accuracy describes whether the predicted values match the actual values of the target field within the incertitude due to statistical fluctuations and noise in the input data values. The execution time or CPU time of a given task is defined as the time spent by the system executing that task, including the time spent executing run-time or system services on its behalf. Program memory usage typically refers to flash memory when it is used to hold the program (instructions). Program memory may also refer to a hard drive or solid-state drive (SSD). The Table I shows the values of these parameters in standalone environment.

Table 1: Comparison of algorithms in standalone environment

Algorithm	Predictive Accuracy	Exe. Time (sec)	Memory used (Mb)	Precision	Recall	F-Measure
Decision Tree	0.9686	0.9286	49.41	0.97	0.99	0.98
Naïve Bayes	0.9112	0.2342	48.42	0.99	0.91	0.95
NBTree	0.9812	0.3442	46.42	0.98	0.90	0.96

The NBtree algorithm shows a high predictive accuracy (0.9812) among the three algorithms. The execution time is the runtime of the algorithm. NBTree algorithm shows minimum execution time (0.2342 sec) than the other algorithms. The memory used (46.42 Mb) for the execution of the NBTree algorithm is low than the other two algorithms. Overall, the hybrid NBTree algorithm shows a better performance than Decision Tree and Naïve Bayes algorithm in standalone mode.

Table II: Comparison of algorithms in cloud environment

Algorithm	Predictive Accuracy	Exe. Time (sec)	Memory used (Mb)	Precision	Recall	F-Measure
Decision Tree	0.9686	0.6259	47.41	0.97	0.99	0.98
Naïve Bayes	0.9212	0.1726	49.42	0.99	0.91	0.95
NBTree	0.9882	0.1826	45.42	0.98	0.90	0.96

The Table II shows the comparison of algorithms in AWS cloud environment. The hybrid NBTree algorithm shows a better predictive accuracy (0.9882) than the other algorithms. The execution time (0.1726 sec) is less in Naïve Bayes algorithm and the memory usage (45.42 Mb) is less in hybrid NBTree algorithm.

The execution time shows a great difference between the standalone and cloud environments. The time is reduced half for execution of algorithms in cloud environment due to the distributed nature which can be seen from both tables.

7. Conclusion

Distributed computing environment allows massive data analysis by processing the data parallelly. The cloud based virtual environment provides scalability and optimized performance. The study is performed to compare performances of Decision Tree algorithm, Naïve Bayes algorithm and NBTree algorithm used for classifying higher secondary school dataset. This study analyses the relationship between the demographic and other academic factors with students' academic performance at higher secondary level. Classifier models with two class values are used for predicting student performance. The predictions of student‘s academic performance can be useful in many contexts. Mainly the higher secondary school result is a turning point for the students, so it is important to be able to identify excellent students by predicting their performance for getting their desired groups at the college level.

This study demonstrates that hybrid NBTree algorithm is the best tree among all three algorithms as it shows high accuracy in both the environment. A minimum time is taken by Naive Bayes algorithm to perform the classification in the both the environments. We noted that the memory used to perform the classification is less in hybrid NBTree algorithm. Overall, the hybrid algorithm NBTree shows a better performance in predicting the students' performance.

References

[1] Jacob, J., Jha, K., Kotak, P., & Puthran, S. (2015). Educational Data Mining techniques and their applications. In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) (pp. 1344-1348). IEEE.\
[2] Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924.
[3] Sin, K., & Muthu, L. (2015). Application of Big Data In Education Data Mining And Learning Analytics--A Literature Review. ICTACT journal on soft computing, 5(4).
[4] Naik, N., & Purohit, S. (2017). Comparative study of binary classification methods to analyze a massive dataset on virtual machine. Procedia computer science, 112, 1863-1870.
[5] Swensson, E., & Dame, E. (2014). Big Data Analytics Options on AWS. In Technical Report. Technical report, 12 2018. vi vi, 12, 13.
[6] Abdualgalil, B., & Abraham, S. (2020, February). Applications of Machine Learning Algorithms and Performance Comparison: A Review. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-6). IEEE.

[7] Lakshmi, D., Arundathi, S., & Jagadeesh, D. (2014). Data Mining: A prediction for student's performance using decision tree ID3 method. International Journal of Scientific & Engineering Research, 5(7), 1329-1335.
[8] Ramaswami, M., & Rathinasabapathy, R. (2012). Student performance prediction. International Journal of Computational Intelligence and Informatics, 1(4).
[9] Christian, T. M., & Ayub, M. (2014, November). Exploration of classification using NBTree for predicting students' performance. In 2014 International Conference on Data and Software Engineering (ICODSE) (pp. 1-6). IEEE.
[10] Ajila, S. A., & Bankole, A. A. (2013). Cloud client prediction models using machine learning techniques. In 2013 IEEE 37th Annual Computer Software and Applications Conference (pp. 134-142). IEEE.
[11] Wu, D., Jennings, C., Terpenny, J., & Kumara, S. (2016, December). Cloud-based machine learning for predictive analytics: Tool wear prediction in milling. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 2062-2069). IEEE.
[12] Chandrasekaran, S. (2018). A Machine Learning Implementation of Predicting the Real Time Scenarios in a better way. International Journal of Pure and Applied Mathematics, 119(15), 1301-1311.
[13] Cai, Z., Gao, Z. J., Luo, S., Perez, L. L., Vagena, Z., & Jermaine, C. (2014, June). A comparison of platforms for implementing and running very large scale machine learning algorithms. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data (pp. 1371-1382).
[14] Hafez, M. M., Shehab, M. E., & El Fakharany, E. (2016, October). Effective selection of machine learning algorithms for big data analytics using apache spark. In International Conference on Advanced Intelligent Systems and Informatics (pp. 692-704). Springer, Cham.
[15] Elgendy, N., & Elragal, A. (2014, July). Big data analytics: a literature review paper. In Industrial conference on data mining (pp. 214-227). Springer, Cham.
[16] Pani, S. K., & Nayak, M. (2016). The decision tree learning algorithm using nbtree for classification of data. ICSTM.
[17] Suyog, B. (2018). Cloud Computing Using Amazon Web Services (AWS). International Journal of Trend in Scientific Research and Development, pp. 2156-2157.
[18] Kulkarni, G., Sutar, R., & Gambhir, J. (2012). Cloud computing-Infrastructure as service-Amazon EC2. International journal of Engineering research and applications, 2(1), 117-125.