Students' Orientation Using Machine Learning and Big Data

https://doi.org/10.3991/ijoe.v17i01.18037

Farouk Ouatik (✉), Mohammed Erritali
Sultan Moulay Slimane University, Beni-Mellal, Morocco
farouk.ouatike@gmail.com

Fahd Ouatik
Cadi Ayad University, Marrakech, Morocco

Mostafa Jourhmane
Sultan Moulay Slimane University, Beni-Mellal, Morocco

Abstract—Students' orientation in public institutions and choosing their academic paths or their appropriate specialization is important to students to continue their studies Easily in their school career. Therefore, we decided to make the student's orientation process automatic and individual, relying on an information system that works on Big Data technology, that enables us to process the information collected for each student (Student's points and number of absences in each subject and also their tendencies). Then we used the algorithms of machine learning, that enable us to give the appropriate specialization to each student. In this paper, we compared the accuracy and execution time of the following algorithms (Naïve Bayes, SVM, Random Forest Tree and Neural Network), where we found that Naïve Bayes is the best for this system.

Keywords—Big data, Classification, Naïve Bayes, SVM, Random Forest Tree, Neural Network

1 Introduction

Student's orientation is an important and difficult process because, taking the decision regarding the human being is very complicated, but it is a process that depends mainly on the learner's points in the first place and the desire in the second degree because if the student likes a specialty, but he does not have sufficient capabilities for him, he will not be able to keep pace with the study program for this specialty. This is why Educational orientation considered a crucial step in the curriculum of each student, more specifically high school students. Unfortunately, the students of the secondary school always find themselves facing this orientation problem, because when they are in the secondary school they cannot yet decide on their choices of orientation, which prevents everyone from going straight to his preferences orientation, then causes a feeling of injustice, which may result in dropping out of studies [1].

Multiple factors influence the orientation of students, mainly social data which does not influence academic results and are far from taking into account the characteristics of all students. Second, families know or hear that, depending on the specialty of the baccalaureate obtained, the possibilities for further studies and access to higher education and professional integration differ, especially when the labor market is tight. To solve this problem, we try to create a system to help the student achieve this. Considering the number of students and the need for time, we decided to use Big Data.

In literature some related work have compared general machine learning such as Sunita B. Aher and Lobo L.M.R.J. [2]compared seven Classification Algorithm (Naive Bays, Simple Cart, ZeroR, J48,Decision Table, ADTree and Random Forest) using weka, then they finded that ADTree works better for Moodle database then Seyed Reza Pakiz and Abolfazl Gandomi [3] compared four classification algorithms but using Mapreduce model with traditional models, they conclude that classification algorithms based on Mapreduce model work better in large datasets. Similarly, Wael Etaiwi, Mariam Biltawi and Ghazi Naymat [4] compared two machines learning classifiers Naïve Bayes and Support Vector Machine (SVM) classifiers using MLlib, of the Apache Spark, and They concluded that Naïve Bayes is more powerful than SVM for Big Data, in another work, Amine Rghioui, Jaime Lloret and Abedlmajid Oumnad [5] compare J48, Bayes Net, ZeroR and Naïve Bayes using healthcare data, and conclude that j48 better than the other classifier algorithm with 99.21% accuracy,another study [6] F.Ouatik ,M.Erritali,F.Ouatik and M. Jourhmane compare Naive Bayes, Neural Networks, and k-nearest-neighbors, by classification accuracy and speed up. They find that Naive Bayes algorithm work better, but in e-learning [7][8][9] and exactly student orientation, there are no studies that could facilitate this process. For this we try to use big data to help students in their orientation.

2 Big Data

The term Big Data describes collections of very large volumes of data - both structured and unstructured - that can be processed and exploited to generate intelligible and relevant information [10].

Big Data is characterized by the "3V" rule: Volume (Big Data refers to very large volumes of raw data), Variety (a Big Data set is typically composed of heterogeneous data, structured or not) and Speed (or Velocity, Big Data is generated at "high speed" or even continuously, which also means processing it quickly, even in real time).

2.1 Hadoop

Hadoop can be considered as a scalable data processing system for the storage and batch processing of very large amounts of data. Its principle is based on multi-node distributed processing to drastically increase the computing and storage capacities in order to process very large amounts of data [11]. Hadoop's business benefits are numerous. With this software framework, it is possible to store and process vast

amounts of data quickly. Faced with the increase in the volume of data and their diversification, mainly related to social networks and the Internet of Things, this is a significant advantage:

The distributed computing model of Hadoop allows you to quickly process Big Data. The greater the number of calculation nodes used, the higher the processing power. The processed data and applications are protected against hardware failures. If a node fails, the tasks are directly redirected to other nodes to ensure that the distributed computation does not fail. Multiple copies of all data are stored automatically.

Unlike traditional relational databases, you do not need to process the data before storing it. You can store as much data as you want and decide later how to use it. This groups unstructured data like text, images and videos.

The open-source framework is therefore free and relies on standard machines to store large amounts of data. Finally, it is possible to adapt the system to support more data by simply adding nodes. The required administration is minimal.

2.2 Mapreduce

Mapreduce [12] is a patron of IT development architecture, invented by Google1, in which parallel, and often distributed, calculations of potentially very large data are carried out. MapReduce runs on a large machine cluster and is highly scalable. It can be implemented in several forms thanks to different programming languages like Java, C # and C++. For novice developers, the Framework is useful because library routines can be used to create parallel programs without worrying about infra-cluster communications, task monitoring, or error handling. Programmers with no experience in parallel and distributed systems can easily use large system resources distributed. In order to distribute the input data and weld the results, it operates in parallel on massive clusters. The size of a cluster has no impact on data processing. In fact, tasks can be spread across any number of servers. That's why MapReduce and Hadoop simplify software development. It is available in several languages including C, C ++, Java, Ruby, Pearl and Python. Programmers can use MapReduce libraries especially based on Java 8 to create tasks without worrying about communication or coordination between nodes.

This is the representation of MapReduce.

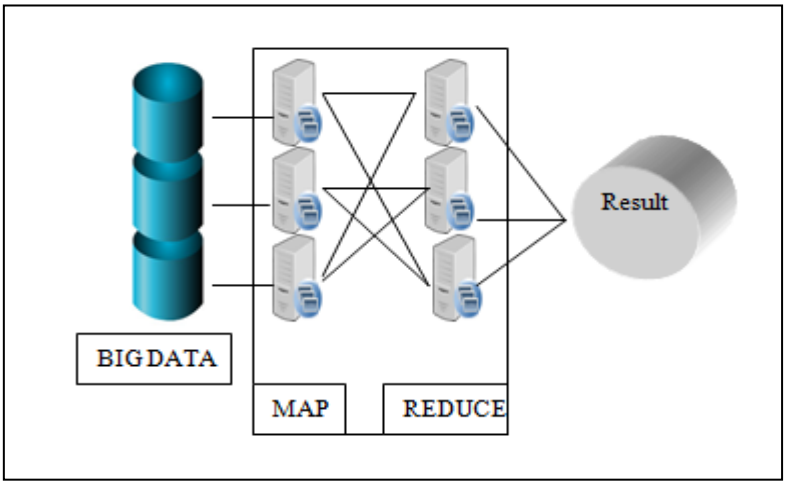

Fig. 1. Representation of MapReduce

This figurine represents MapReduce work with map step and reduce step.

2.3 HDFS

HDFS (Hadoop Distributed File System) [13] is the Hadoop component responsible for storing data in a Hadoop cluster. HDFS can be launched on commodity hardware, which makes it very tolerant to errors. Each piece of data is stored in several places, and can be retrieved under any circumstances. In the same way, this replication makes it possible to fight against the potential corruption of the data. However, HDFS stands out from a typical file system for the following main reasons:

- HDFS is optimized to maximize data rates. The size of a data block is thus 64 MB in HDFS against 512 bytes at 4 KB in most traditional file systems, which reduces the seek time. (It is possible, however, to increase the size of a block to 128 MB or 256 MB as needed).
- HDFS is a file management system, type of Write Once Read Many (WORM) file management system, then it is accessed several times.
- HDFS provides a block replication system with a configurable number of replications (3 by default). During the writing phase, each block corresponding to the file is replicated to separate nodes in the cluster, which helps to ensure reliability and readability when reading data. If a block is unavailable on one node, copies of that block will be available on other nodes.
- HDFS relies on the native OS of the file system to present a unified storage system based on a set of heterogeneous disk and file systems.

3 Classification

This data analysis method brings together supervised learning algorithms adapted to qualitative data. The objective is to learn (in other words, to find) the relation which links a variable of interest, of qualitative type, to the other observed variables, possibly for the purpose of prediction [14]. We use classification when the variable of interest is qualitative, i.e. it takes its values in a space that does not have natural metrics. For example, we can try to predict the literary genre of a book; this variable is discrete (genre "detective", genre "science fiction", etc.) and there is no relation between the genres, it is difficult to define a distance between them. The simplest classification algorithms are logistic regression, the k-nearest neighbor, the most complex are the neural networks, the vector machine supports, the mixture model (mixture models), the Bayesian classifier, Random Forest Tree, OneR, etc.

3.1 Naïve bayes

Naive bayes [15], commonly used in machine learning, is a collection of classification algorithms based on Bayes' theorem. It is not a single algorithm, but a family of algorithms. All these algorithms share a common principle, namely that each classified characteristic is independent of the value of any other characteristic. Even though it is a relatively simple concept, Naive Bayes can often outperform the most complex algorithms and is extremely useful in common applications such as spam detection and classification. They allow us to predict the probability of an event occurring based on the conditions we know for the events in question. The name comes from Bayes' theorem.

$$P(A|B) = P(B|A)\ P\ (A)/\ P(B) \quad (1)$$

3.2 SVM

Support Vector Machine (SVM): SVMs [16] are a family of machine learning algorithms that solve classification, regression, and anomaly detection problems. They are known for their solid theoretical guarantees, their great flexibility and their ease of use even without great knowledge of data mining. Their principle is simple: its purpose is to separate data into classes using a border as "simple" as possible, so that the distance between the different groups of data and the border that separates them is maximum. This distance is also called "margin" and the SVMs are thus qualified as "wide margin separators", the "support vectors" being the data closest to the border.

3.3 Random forest tree

Random forest tree [17] is a supervised machine learning method that can perform both classification and regression tasks. His principle is to use many decision trees, each one is constructed with a different subsample of the training set, and for each construction of a tree, the decision at a node is made according to a subset of variables

drawn randomly. Then, we use all the decision trees produced to make the prediction, with a majority vote for the classification (the predicted variable is of factor type), or an average for the regression (the predicted variable of numeric type).

3.4 Neural network

A Neural network [18] is a calculation model whose design is very schematically inspired by the functioning of real neurons (human or not). Neural networks are generally optimized by statistical learning methods thanks to their capacity for classification and generalization, such as automatic classification of postal codes or decision-making regarding a stock purchase according to the evolution of Classes. They enrich with a set of paradigms allowing to generate vast functional, flexible and partially structured spaces. They belong on the other hand to the family of the methods of artificial intelligence which they enrich by allowing to make decisions based more on perception than on formal logical reasoning.

There are many tools use these algorithms, in this work we used Weka to compare the Performance of these classifiers based on accuracy and execution time in order to choose the best.

3.5 Weka

Weka (Waikato Environment for Knowledge Analysis) [19] is a set of tools for manipulating and analyzing data files, implementing most artificial intelligence algorithms, inter alia, decision trees and neural networks. It is written in Java. It mainly consists of:

- Java classes for loading and manipulating data.
- Classes for the main supervised or unsupervised classification algorithms.
- Attribute selection tools, statistics on these attributes.
- Classes allowing to visualize the results.

Large data volumes linked to Big data quickly lead to memory saturation problems when using data mining software. Weka implements a set of techniques and architecture allowing to circumvent these limits and to successfully manage these Big Data such as MOA, [20] that is an open-source framework, which contains a set of learning algorithms and assessment tools. it is used for big data flow exploration and also boosts bidirectional interaction with Weka.

4 Results

As Figure 2 shows, the Naïve Bayes classifier is the most precise, with an accuracy of 92.10%, then Neural Network with 90.37%, SVM gives 88.13% of accuracy, followed by Random Forest lends an accuracy of 86.22%.

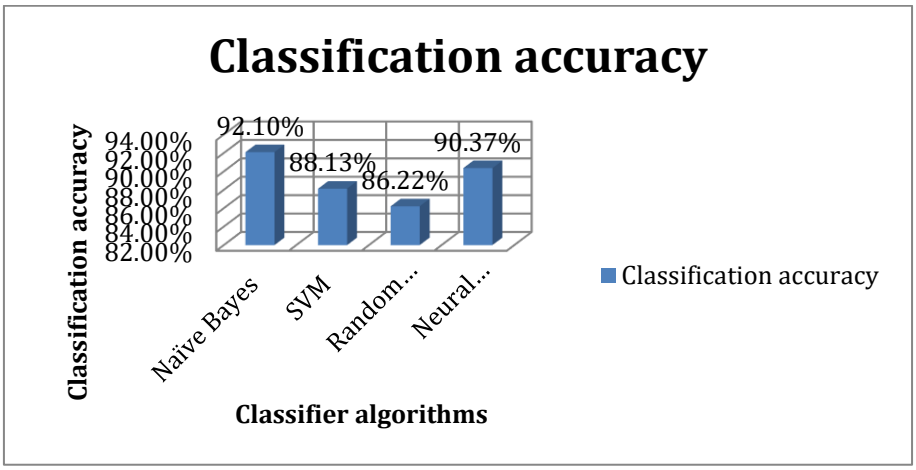

Fig. 2. SVM, Random Forest Tree, Naive Bayes and Neural Network classification accuracy

Figure 3 illustrates the data processing time for the classification algorithms. The Naïve Bayes classifier was found the most accurate between all the classifiers used in this article. Here too, we can see that the execution time of Naïve Bayes is adequate for this use.

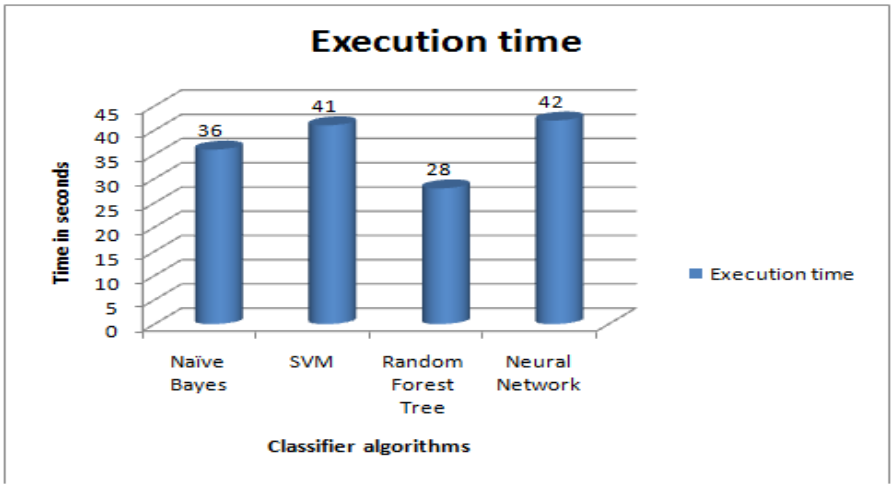

Fig. 3. The execution time of SVM, Random Forest Tree, Naive Bayes and Neural Network

5 Conclusion

In this study, we compare four classification algorithms, to find the right algorithm for student orientation, using student grades and also the number of absences for each

subject These four classification algorithms, are Neural Network, Naïve Bayes, SVM, Random Forest Tree. We use Weka with MOA package to test the result. After the test we find that Naïve Bayes is better for students' orientation.

6 References

[1] M. Phanupong, et al., "Prediction of Student droput using personel profile and data mining aproch," Intelligent and Evolutionary Systems, pp. 143-155.

[2] B.Sunita et al., "Selecting the Best Supervised Learning Algorithm for Recommending the Course in E-Learning System" March 2012.

[3] Seyed Reza Pakize, et al., "Comparative Study of Classification Algorithms Based on MapReduce Model"2014.

[4] Etaiwi, Maria et al., " Evaluation of classification algorithms for banking customer's behavior under Apache Spark Data Processing System" 2017. https://doi.org/10.1016/j.procs.2017.08.280

[5] Amine Rghioui et al., " Big Data Classification and Internet of Things in Healthcare " April-June 2020.

[6] F.Ouatik , et al., " Comparative study of MapReduce classification algorithms for students orientation".," Procedia Computer Science Volume 170, 2020, Pages 1192-1197. https://doi.org/10.1016/j.procs.2020.03.030

[7] F. Ouatik, et al., "The EOLES project remote labs across the Mediterranean: an example of a successful experience," Proceedings of the 2017 International Conference on Smart Digital Environment, 2017.

[8] R. Conjin, et al., "Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS," IEEE Transactions on Learning Technology.

[9] Sfenrianto Sfenrianto, et al., "E-Learning Effectiveness Analysis in Developing Countries: East Nusa Tenggara, Indonesia Perspective" Bulletin of Electrical Engineering and Informatics Vol 7, No 3 September 2018 https://doi.org/10.11591/eei.v7i3.849

[10] Gunasekar Thangarasu, Kayalvizhi Subramanian, " Big Data Analytics for Improved Care Delivery in the Healthcare Industry " International Journal of Online and Biomedical Engineering, Vol. 15, No. 10, 2019. https://doi.org/10.3991/ijoe.v15i10.10875

[11] Du Juan, Wu Fenfen," Distributed Data Mining in Wireless Sensor Networks " International Journal of Online and Biomedical Engineering, Volume 12, Issue 11, 2016.

[12] Zhang Feng, Xue Hui-Feng, Xu Dong-Sheng, Zhang Yong-Heng, You Fei," Big Data Cleaning Algorithms in Cloud Computing " International Journal of Online and Biomedical Engineering, Volume 9, Issue 3, July 2013.

[13] SUN Ya-ni, CHEN Xinhua," Application and Realization of an Improved Apriori Algorism in a Hadoop Simulation Platform for Mass Data Processing ", International Journal of Online and Biomedical Engineering, Volume 12, Issue 2, 2016.

[14] Oluwakemi Christiana Abikoye, Benjamin Aruwa Gyunka," Android Malware Detection through Machine Learning Techniques: A Review " International Journal of Online and Biomedical Engineering, Vol. 16, No. 2, 2020. https://doi.org/10.3991/ijoe.v16i02.11549

[15] Wan Fairos Wan Yaacob, et al, " Supervised data mining approach for predicting student performance" Indonesian Journal of Electrical Engineering and Computer Science Vol. 16, No. 3, December 2019, pp. 1584~1592. https://doi.org/10.11591/ijeecs.v16.i3.pp1584-1592

[16] Mohammad Karimi Moridani, Shahrzad Marjani," A Review of the Methods for Sudden Cardiac Death Detection " International Journal of Online and Biomedical Engineering, Vol. 16, No. 9, 2020.
[17] Mostafa A.Salama, Ghada Hassan," A Novel Feature Selection Measure Partnership-Gain",International Journal of Online and Biomedical Engineering, Vol. 15 No. 4, 2019 .
[18] Mohammad Karimi Moridani, Shahrzad Marjani," A Review of the Methods for Sudden Cardiac Death Detection " International Journal of Online and Biomedical Engineering, Vol. 16, No. 9, 2020. https://doi.org/10.3991/ijoe.v16i09.14485
[19] Farhad Alam et al., "Comparative Study of J48, Naive Bayes and One-RClassification Technique for Credit Card Fraud Detection using WEKA"2017.
[20] Albert Bifet et al., " MOA: Massive Online Analysis", 2010.

7 Authors

Farouk Ouatik works in Sultan Moulay Slimane University, Beni-Mellal, Morocco. Email: farouk.ouatike@gmail.com

Mohammed Erritali works in Sultan Moulay Slimane University, Beni-Mellal, Morocco.

Fahd Ouatik works in Cadi Ayad University, Marrakech, Morocco

Mostafa Jourhmane works in Sultan Moulay Slimane University, Beni-Mellal, Morocco.

Article submitted 2020-08-26. Resubmitted 2020-10-09. Final acceptance 2020-10-13. Final version published as submitted by the authors.