Journal of Ubiquitous Systems & Pervasive Networks
Volume 13, No. 1 (2020) pp. 21-26

Student orientation using machine learning under MapReduce with Hadoop

F.Ouatik[a]*, M.Erritali [b], F.Ouatik[c], M.Jourhmane[d]

[a]*Department of Computers Sciences, Sultan Moulay Slimane University,Beni Mellal, Morocco*
[b]*Department of Computers Sciences, Sultan Moulay Slimane University, Beni Mellal, Morocco*
[c] *Department of physic, Cadi Ayad University, Marrakech, Morocco*
[d] *Department of Mathematics , Sultan Moulay Slimane University Beni Mellal, Morocco*

Abstract

Academic orientation is a procedure which consists in helping students to succeed in their educational path and to know the ideal path. the choice, of course, allows the student to prepare for professional life and find the right path to the future. It is therefore essential to choose the sector well to avoid any pitfall and risk hazardous choices and also reorientation. There are establishments which organize orientation sessions with the students, but the number of students is very high, which pushed us to carry out an orientation system to make this service accessible by all the students and also to make the procedure very powerful orientation and in a lapse of time. Decision making related to orientation is very complicated by the fact that it is linked to several factors (prerequisite, the student's previous academic career, marks and the number of absences by subject) as the learner's marks give us information about his abilities and the extent of his mastery of the specialization subjects. As for the student's tendencies and desires, they appear through his keenness on these subjects, and he is not absent from them, And it is evidenced by the number of absences by subject of the learner.These data are useful, but they pose problems in storage and processing, so the solution is to use Big Data technology. In this article we used the notes and the number of student absences by subject. These data are stored in Hadoop Distributed File System (HDFS) and processed by MapReduce using Hadoop framework. We compared the classification accuracy and speed up of Neural Networks, Naive Bayes and k-nearest-neighbors classification algorithms to make the decision and we found that Naive Bayes is the most suitable for this procedure.

Keywords: *Machine learning, Big Data, MapReduce, K-Nearest-Neighbors, Neural Networks and Naive Bayes.*

1.Introduction

Machine learning is a science for discovering models and making predictions from data based on statistics, pattern recognition and predictive analysis. it is very effective in situations where the decision must be made taken from large, diverse and changing data sets, that is: Big Data. For the analysis of such data, it proves to be much more efficient than traditional methods in terms of accuracy and speed. Traditional analytical tools are not efficient enough to fully exploit the value of Big Data. . The volume of data is too large for comprehensively analyzes, and the correlations and relationships between this data are too large for analysts to be able to test all hypotheses in order to derive value from this data. There are open source tools for storing and processing massive data, and among them is a tool that is widely used in most important work, it is Hadoop, which is used by [1] to compare Fitting Speed and the Predictive Accuracy of SVM and decision. another comparison of SVM classifiers based on modified PSO and SVM classifiers based on the traditional PSO algorithm made by [2] using the classification accuracy. Some work used Hadoop and Weka to compare accuracies of Naïve Bayes, SVM, and j48, using Knowledge flow of Weka before and after normalization [3], and [4] use InterIMAGE Cloud Platform to compare Naïve Bayes, SVM, Decision Trees and Random Forest. however the quality and security of data are important. Therefore [5] present a global solution to evaluate the data's quality of Big Data without impacting the security of data. in another work [6] propose a conversion tool from a relational to graph database like a solution of Big Data. There is little research on student's orientation, It relies on filling out a form or oral interviews or portfolio, but this takes a lot of time ,so to made this process automatic ,[7] adopt e-portfolio for students orientation and [8] make a chatbot of educational and professional guidance based on John Holland 's theory and the RIASEC questionnaire ,but to process numerous data we must use Big Data and this is our study. This paper presents a method to guide students based on big data and machine learning algorithms ,It is organized as follows: Section1 presents Big Data and Hadoop, mapreduce and HDFS, Section2 presents Classification algorithms with mapreduce (Neural Networks, Naive Bayes and K-nearest-neighbors)

* Corresponding author. Tel.: +212653793053
E-mail address: farouk.ouatike@gmail.com

DOI: DOI: 10.5383/JUSPN.13.01.003

and Section3 evaluates the Performance of Neural Networks, Naive Bayes and k-nearest-neighbors using accuracy and execution time of these algorithms and Section 4 concludes the paper.

2.Big data handling tools

The orientation of the students requires collecting gigantic amounts of data. This poses a major challenge: it is no longer just a question of collecting and storing these volumes of data, it is also a question of processing and analyzing them in real time. However, conventional data management tools have become unsuitable for processing, either for technical reasons, or for economic reasons, or for both. To resolve this problem, we decided to use Big Data. Big data refers to data sets which, due to their Variety, Speed or Volume cannot be easily stored, manipulated or analyzed with traditional methods, such as spreadsheets, relational databases or ordinary statistical tools. There are new tools have been developed to overcome the problems of collecting, storing and processing large volumes of data. Among them, Hadoop.

2.1. Hadoop

Hadoop [7] is an open source IT platform, It is part of the Apache foundation. It is capable of handling gigantic volumes of data, structured and unstructured, within the framework of a distributed system. It offers great flexibility. Its performance changes almost linearly depending on the number of machines making up the cluster. The higher number of nodes, the shorter the execution time of jobs. It works on the principle of calculation grids consisting in distributing the execution of an intensive data processing on several nodes or clusters of servers. Java is the preferred language for writing native Hadoop programs. However, it is possible to use python, bash, ruby, Perl....

The benefits of Hadoop

- Economical: Hadoop allows companies to unleash the full value of their data by using inexpensive servers.
- Flexible: Hadoop allows you to store all types of data in an extensible manner. The data can be unstructured and not follow any structured schema (PDF, MP3, database, etc.) thanks to its file system.
- Users can transfer their data to Hadoop, without needing to reformat it.
- Tolerates failures: data is replicated across the cluster so that it is easily recoverable following a failure of the disk, node or block.

The components of Hadoop

Hadoop is mainly made up of two components:

- The distributed file management system, HDFS.
- The MapReduce framework (version 1 of Hadoop) / YARN (version 2 of Hadoop).

More specifically, the Hadoop ecosystem includes many other tools covering data storage and distribution, distributed processing, data warehouse, workflow, programming, not to mention the coordination of all components. We are talking about tools like Hive, Pig, Hbase, Flume, etc.

2.2. MapReduce method

MapReduce [8] is a programming model (or structure) available in Hadoop environments that is used to access big data stored in the Hadoop File System (HDFS). MapReduce is an essential element and an integral part of the functioning of the Hadoop environment. With MapReduce , rather than sending the data to the location of the application or algorithms, the algorithms are executed on the server where the data are already located, which has the effect of speeding up processing. Data access and storage is done on disk - inputs are usually stored as files containing structured, semi-structured or unstructured data, and the output is also stored in files. MapReduce plays a major role in processing large amounts of data. The distribution of data within numerous servers allows the parallel processing of several tasks, each relating to pieces of files. The Map function performs a specific operation on each element. The Reduce operation combines the elements according to a particular algorithm, and provides the result. Note that the principle of delegation can be recursive: the nodes to which tasks are entrusted can also delegate operations to other nodes.

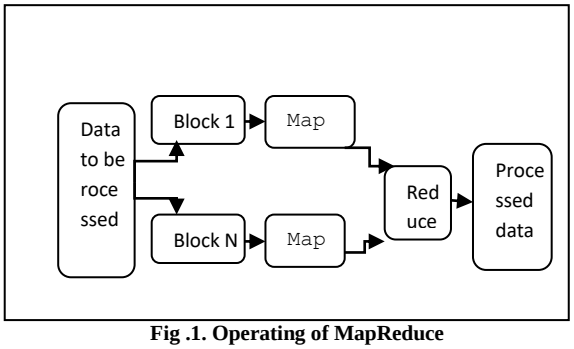


Fig .1. Operating of MapReduce

This figurine presents Operating of MapReduce
At the heart of MapReduce are two functions, Map and Reduce, which are sequenced one after the other.

Map
The input data is divided into smaller blocks. Each block is then assigned to a Mapper for processing. The Hadoop framework decides the number of Mappers to use depending on the volume of data to be processed and the size of the memory blocks available on each Mapper server.

Reduce
When all the Mappers have finished their processing, the framework mixes and sorts the results before transmitting them to the Reducers (Reducers cannot start until all the Mappers have finished their processing). Map output values assigned the same key value are assigned to a single Reducer, which aggregates the map values for that key.

Combine and Score

There are two intermediate steps between Map and Reduce: Combine and Partition.

The combine is an optional process. A Combiner is in fact an additional Reducer, which works separately on each Mapper server. The Reducer continues to reduce the data of each Mapper in a simplified form before transmitting it downstream - which facilitates and accelerates the mixing and sorting operations, since the volume of data to be processed has been reduced. In many cases, due to the cumulative and associative actions performed by the Reduce function, the class of the Combiner is that of the Reducer (if necessary, the Combiner can be associated with a separate class).

Partition is the process which translates the key, value> pairs generated by the Mappers into another key, value> pairs which are injected into the Reducer. Partition defines the mode of presentation of the data to the Reducer and assigns this mode to a given Reducer.The default Partitioner determines the key hash value transmitted by the Mapper and assigns a partition based on this value. There are as many scores as there are Reducers. When the partitioning is finished, the data of each partition are transmitted to a given Reducer. MapReduce implements the following features:

- Automatic parallelization of Hadoop programs.
    - HDFS is responsible for the distribution and replication of data;
    - The master divides the work into parallel jobs and distributes them;
    - The master collects the results and manages the breakdowns of the nodes.
- Transparent management of distributed mode.
- Tolerance for breakdowns.

A MapReduce program consists of the following three parts

- The driver, which runs on a client machine, is responsible for initializing the job and then submitting it for execution.
- The mapper is responsible for reading and processing the data stored on disk.
- The reducer is responsible for consolidating the results from the map and then writing them to disk.

2.3. HDFS

Hadoop Distributed File system (HDFS)[9]is a distributed file system. This means that it uses the network to manipulate access to the system. Each component of the network can therefore access each resource of other computers making up this network. In addition, HDFS is a system suitable for working with large volumes of data (1 GB and more). The great positive point of this system is universality. It does not need a very powerful machine to function properly. It is designed to work on ordinary machines.

The objectives of HDFS are:

- Guarantee data integrity despite a system failure (machine crash)
- Allow rapid processing of large files (GB and above);
- Bring the reading closer to the location of the files rather than bringing the files closer to the reading;
- HDFS integrates interfaces that allow applications to locate requested resources in the cluster more quickly.

3. Classification algorithms

Machine learning puts artificial intelligence at the service of Big Data. These are systems that use algorithms to learn from receiving data. One of the most representative examples of this technology is the Amazon.com recommendation engine.

3.1. Neural Network

Neural Networks [10] are based schematically on the functioning of biological neurons. They are generally equated with artificial intelligence.

The principle of the Neural Network is to schematically reproduce the human brain and its mode of association in order to automate it in a machine. Like our brain, the Neural Network begins with a learning phase, with the association of input data (image, sound, text, etc.) with specific categories as references. Once this learning is done, when new data inputs will be presented to it, the Neural Network will associate them with this or that category depending on what it has learned. The Neural Network works by statistical analysis, as opposed to other systems capable of learning which make decisions by comparison with a set of deductive rules (algorithmic logic...).

The principle of the Neural Network is to create statistical reasoning. Based on what he has learned in the past, through a learning "base", he will make decisions about the input data he receives. The choice of classification will be based on the probability of resemblance to a class that it already knows. This is also called experiential learning, with the Neural Network acting as a statistical decision aid.

Neural Network Algorithm:

- gi = {b1, b2, b3,…, be}, gi ∈ G, where
- gi : an instance;
- G : a dataset;
- q : the length of gi; it and the number of inputs of a neural network;
- The inputs have this format <instk, targk, type>;
-Instk represents gi,: a neural network input;
- Targk represents the output desirable if instk is a training instance;
- type= “train” or “test,” :type of instk; if the value “test” is set, targk field left empty.
- D test data
- O output
- γ : weight
- β : bias
- Er : represent the error-sensitivity
- I : the input of neurons

input :
G,D
output:
O
q mappers , l reduce
1-one back-propagation contains inputs q inputs, l outputs, d neurons in hidden layers .
2 - initialize γ_{ij} =random$_{1ij}$ ϵ (-1,1)
β_{2j} =random$_{1ij}$ ϵ (-1,1)
3 -v $g_i \epsilon$ G, gi =$\{b_1, b_2, \dots, b_p\}$
input $b_i \rightarrow q_i$,neuron j in hidden layer
$Ijd = \sum_{i=1}^{q} b_i \cdot \gamma_{ij} + \beta_j$
Ljd = $1/(1 + e\text{^}Ijs)$
4- input lj $\rightarrow$outi ,neuron j in output layer
$Ijl = \sum_{i=1}^{d} L_{js} \cdot \gamma_{ij} + \beta_j$
Ljl = $1/(1 + e\text{^}Ijl)$
5- in each output
ERjl= Ljl (1- Ljl)(target j- Ljl)
6- in hidden layer
$ERjd = Ljd(1- Ljd) \sum_{i=1}^{l} ER_i \cdot \gamma_{il}$
7 - update
$\gamma_{ij} = \gamma_{ij} + \mu \cdot ERj \cdot Lj$

$\beta_j = \beta_j + \mu \cdot ERj$
repeat 3,4,5,6,7
until min $(E[e^2])$= min $(E[(target_j - Ljl)^2])$

8-devide D (D_1,D_2,...D_q)
9- execute(3),(4) for each mapper
10-mapper outputs(gj,Lj)
11- reducer collects
do (9),(10),(11)
until D is traversed
12- reducer outputs O
end

3.2. K-nearest neighbors

The k nearest neighbors (KNN)[11] algorithm is a simple and easy to implement supervised machine learning algorithm that can be used to solve classification and regression problems. To make a prediction, the K-NN algorithm will be based on the entire dataset. Indeed, for an observation, which is not part of the dataset, which we wish to predict, the algorithm will look for the K instances of the dataset closest to our observation. Then for these neighbors, the algorithm will be based on their output variables (output variable) to calculate the value of the variable of the observation that we want to predict. Otherwise:• If K-NN is used for the regression, the mean (or median) of the variables of the closest observations will be used for the prediction• If K-NN is used for the classification, it is the mode of the variability of the closest observations which will be used for the prediction. It needs a distance calculation function between two observations. The closer two points are to each other, the more similar they are and vice versa.There are several distance calculation functions, in particular, the Euclidean distance, the distance from Manhattan, the distance from Minkowski, that from Jaccard, the distance from Hamming… etc. We choose the distance function according to the types of data we are handling. So for quantitative data (example: weight, wages, size, the amount of electronic basket, etc.) and of the same type, Euclidean distance is a good candidate. As for the distance from Manhattan, it is a good measure to use when the data (input variables) are not of the same type (example: age, sex, length, weight, etc.).

3.3. Naive Bayes

The Naive Bayes classifier [12] is a probabilistic, supervised, simple classifier based on the application of the Bayes theorem with the naive hypothesis, that is to say that the descriptors (Xi) are assumed to be conditionally independent of the predict variable (Y). Despite this strong assumption, the Naive Bayes classifier is very effective in many applications and is often applied for supervised classification. The Naive Bayes classifier acquired as input the estimation of conditional probabilities by variable P (Xij | Y).

Bayes' theorem.

$$p(c|f1,f2,...,fn)=\frac{p(c)p(f1,f2,...,fn|c)}{p(f1,f2,...,fn)} \quad (1)$$

There are many works that concern Naive Bayes with map reduce like [13],[14] and [15] , this is Naive Bayes algorithm:

Input:
Training data T.
D= (d1, d2, d3,.., dn) // the predictor value's variable in testing dataset.
Output:
O //Class of testing data .
Step :
Read T;
- Calculate mean and variance of predictor variables;
- Calculate di;
- Calculate the probability for each class;
- Get the highest probability;
- Return O;
end.

4.Performance evaluation

We used a Haddop Cluster consisting of 3 computers are Datanodes and one computer is Namenode. we configured The Hadoop cluster with 19 mappers and 1 reducer.For neural networks, we use 3 layers, the hidden

layer consists of 19 neurons.for k- nearest neighbors we take k=3.
In order to direct students to specializations: Science, Literature, Technology, Original Education, we calculated the rates and total absences of each student according to the subjects for the three years of preparatory school.The datasets used consist of 19 features (student_id, Math_rate, Physic_rate, Life_and_earth_sciences _rate, Arabic_rate, English_rate, Frenc_ rate, Islamic_rate, Philosophy_rate, Sport_rate; N-H-A- Maths; N_H_A_Physic; N_H_A_Life_and_earth_sciences; N_H_A_Arabic; N_H_A_English; N_H_A_French; N_H_A_Islamic; N_H_A_Philosophy N_H_A_Sport).

The size of the datasets was varied from 1 MB to 1 GB
We evaluate accuracy and execution time of Naive Bayes, KNN and Neural Network

4.1. Algorithms execution time

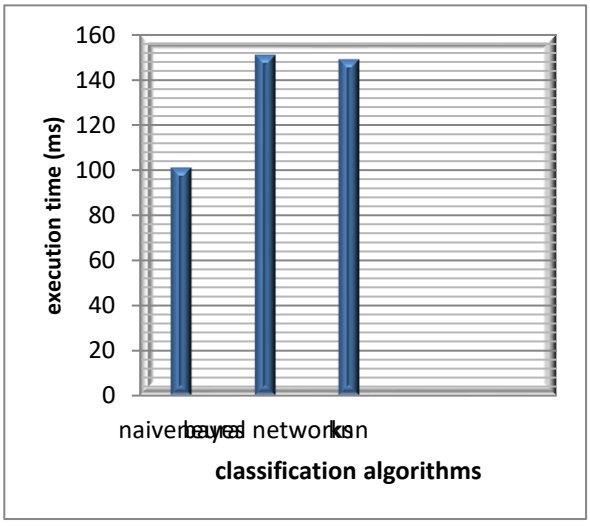
This figure presents Neural Network, K-nearest neighbors and Naive Bayes execution times.

Fig. 2. Neural Networks, K-nearest neighbors and Naive Bayes execution time.

The figure Fig. 1, show that the execution time of Naive Bayes is short compared to the others algorithms, but Neural Network execution time is near to k-nearest neighbors execution time.

4.2. Algorithms accuracy

This figure presents the Neural Networks accuracy, K-nearest neighbors accuracy and Naive Bayes accuracy.



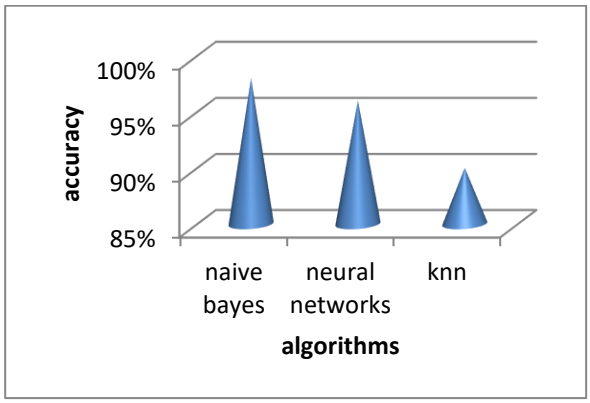


Fig. 3. Neural Networks, K-nearest neighbors and Naive Bayes accuracies.

From this figure , Naive Bayes accuracy is very high compared to the others algorithms.

5.Conclusion

KNN is very slow and its accuracy too low because it stores the entire data set to make a prediction, KNN makes predictions by calculating the similarity between an input observation and the different observations of the data set, on the contrary the execution time of Naive Bayes is low while its accuracy is very high.
After comparing Neural Networks, Naive Bayes and K-nearest neighbors by the time of execution and accuracy it becomes clear that Bayes is the most adequate to our data ,where it is characterized by accuracy and speed and this is necessary so that we can take the decision with high quality in a time lapse.

References

[1] Prajesh P Anchalia , Kaushik Roy ,"The k-Nearest Neighbor Algorithm Using MapReduce Paradigm" 2014 https://doi.org/10.1109/ISMS.2014.94

[2] Liliya Demidova , Evgeny Nikulchev , Yulia Sokolova “Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles” ,2016. https://doi.org/10.14569/IJACSA.2016.070541

[3] Anuja Jain , Varsha Sharma , Vivek Sharma “Big data mining using supervised machine learning approaches for Hadoop with Weka distribution” ,2017

[4] V. A. Ayma , R. S. Ferreira , P. Happ , D. Oliveira , R. Feitosa , G. Costa , A. Plaza , P. Gamba “classification algorithms for big data analysis, a map reduce approach”,2015. https://doi.org/10.5194/isprsarchives-XL-3-W2-17-2015

[5] Mohamed Talha, Nabil Elmarzouqi, Anas Abou el kalam, "Quality and Security in Big Data: Challenges as opportunities to build a powerful wrap-up solution",2020. https://doi.org/10.5383/JUSPN.12.01.002

[6] Yemna Sayeb, Radhouane Ayari, Sarra Naceur, Henda Ben Ghezala, "From Relational Database to Big Data: Converting Relational to graph database, MOOC database as example" 2010.

[7]GuerssFatima zahra, AitdaoudMohammed, DouziKhadija, TalbiMohammed, NamirAbdelouahed, "Implementation of a computerized system for the orientation of the Moroccan student in the university", 4th world conferenceon educational technology researches, wcetr- 2014.

[8] Omar Zahour, El Habib Benlahmar, Ahmed Eddaoui, Hafsa Ouchra,OumaimaHourrane,"Towards a Chatbot for educational and vocational guidance in Morocco: Chatbot E-Orientation",International Journal of Advanced Trends in Computer Science and Engineering, 2020. https://doi.org/10.1016/j.procs.2020.07.079

[9] Song Gao, LinnaLi, WenwenLi, KrzysztofJanowicz, Yue Zhang "Constructing Gazetteers from Volunteered Big Geo-Data based on Hadoop"2014.

[10]Ibrahim Abaker Targio Hashem,Nor Badrul Anuar,Abdullah Gani,Ibrar Yaqoob,Feng Xia,Samee Ullah Khan,"MapReduce: Review and open challenges",2016.

[11] Dinh-Mao Bui, Shujaat Hussain, Eui-Nam Huh, Sungyoung Lee ,"Adaptive Replication Management in HDFS Based on Supervised Learning ",2016.

[12] GrégoireMontavon,WojciechSamek,Klaus-Robert Müller ,"Methods for interpreting and understanding deep Neural Networks",2017

[13]Panagiotis Moutafis, George Mavrommatis, Michael Vassilakopoulos,SpyrosSioutas, " Efficient processing of all-k-nearest-neighbor queries in the MapReduce programming framework",2019 https://doi.org/10.1016/j.datak.2019.04.003

[14] LIU Peng, ZHAO Hui-han,TENG Jia-yu, YANG Yan-yan, LIU Ya-feng, ZHU Zong-wei " Parallel Naive Bayes algorithm for large scale Chinese text classification based on spark"2019 https://doi.org/10.1007/s11771-019-3978-x

[15] Sunil Kumar and Maninder Singh," Diabetes Data Analysis UsingMapReduce with Hadoop",2018

[16] Bingwei Liu, Erik Blasch, Yu Chen ,Dan Shen, Genshe Chen," Scalable Sentiment Classification for Big DataAnalysis Using Na¨ıve Bayes Classifier",2013 . https://doi.org/10.1007/978-981-13-1642-5_15

[17] Mulualem Mheretu Temesgen,Dereje Teferi Lemma," A Scalable Text Classification Using NaiveBayes with Hadoop Framework",2019. https://doi.org/10.1007/978-3-030-26630-1_25