International Conference on Emerging Innovation in Engineering and Technology

ICEIET-2017

A Bigdata Analytic Approach For Finding Instructor's Performance Based On Students Outcome

Dr.M.S. Anbarasi[1], M.Mohana[2], S.JansyRani[3], V.DivyaPriya[4] , M.Aysha[5]

[1]Assistant Professor,[2]Student Member/IT, Pondicherry Engineering College

[1]anbarasims@pec.edu [2]mohanam@pec.edu [3]mejansysaran@gmail.com [4]divyapriyav@pec.edu [5]aysham@pec.edu

Abstract

In education institutions, analyzing the student dataset is performed using the data mining techniques. Based on the academic marks of the student, predicting the tutor performance will be helpful for the institutions to develop their education system. The existing methodologies are mainly performed using the decision tree algorithm which takes more time. In this paper, predicting the mentor performance using K-means algorithm with the MapReduce programming model in an efficient way. The experimental setup is carried out in Hadoop framework with MapReduce programming model. The result analysis is evaluated for accuracy, precision, recall, specificity by comparing with the existing classification schemes. Our proposed technique improvises the prediction accuracy and reduces the time.

Keywords— Decision tree algorithm, Hadoop, K-means, MapReduce programming, Performance evaluation.

I. INTRODUCTION

Today, one of the greatest difficulties of advanced education establishments is the expansion of information and how to utilize them to enhance nature of scholarly projects and benefits and the administrative choices [1],[3]. An assortment of "formal and casual" techniques in view of "qualitative and quantitative" strategies is utilized by advanced education foundations to settle issues, which keep them far from accomplishing their quality targets [1],[2]. Regardless, systems used as a piece of cutting edge training for quality articles are fundamentally in light of predefined inquiries and frameworks to separate the data. Also, these strategies do not have the capacity to uncover valuable shrouded data.

Masked information in larger datasets is best processed with data mining techniques. Data mining (at times called information revelation) is the way toward finding "hidden message" examples and learning inside vast sums of information and procedure of making forecasts for results then again practices. Data mining can be best defined as the robotized procedure of extricating helpful learning and data counting designs, affiliations, changes, patterns, oddities, and significant structures that are obscure from vast or complex datasets.

Big Data is eminent not in view of its size, but rather due to its relationality to other information.Due to the methods used to store the data, Big Data is fundamentally networked (threaded with connections). But these connections are not useful directly. The actual value comes from the patterns that can be derived from the related pieces of data about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself [3]. Other than this, Big Data has tremendous volume, high speed, much assortment and variety.These features of Big Data present the main challenges in analyzing Big data which are:

(1) Efficient and effective handling of large data,

(2) Processing time and accuracy of results trade – off; and

(3) Filtering important relevant data from all the data collected.

To meet the scalability and performance requirements in very large datasets, efficient and parallel implementation of algorithms plays very important role. Using experimental results, they have demonstrated that the proposed algorithm is efficient and can scale well large datasets on commodity hardware [3]. Hadoop MapReduce framework to implement K-Means algorithm to make it applicable to very large data. It can be executed efficiently, in parallel, by applying proper <key, value> pairs. The proposed work contains the data collection of students mark list along with the instructor handled for each subjects. Then the collected data is classified using the K-means clustering algorithm. The mapreduce programming model is utilized to process the large amount of data in hadoop framework. The rest of the paper is organized as follows: Section II gives the clear idea about the related paper. Section III contains the K-means algorithm with the mapreduce programming model. Section IV shows the experimental setup of the hadoop framework. Section V represents the results and the discussion Section VI concluded the proposed work

.

RELATED WORKS

In paper [2], proposed the buildup which is suitable typologies for the 15,000 understudies, analysts utilized both Two Step and K-means, two capable clustering algorithms. They initially connected the calculations to the general groupings distinguished above, with blended outcomes. The limits among bunches were hazy and scattered, and even after rehashed testing on holdout datasets, and additionally the evacuation of suspected anomalies (cases that don't seem to have a place with any gathering), the outcomes did not enhance essentially. It's conceivable that the understudies' underlying revelation of objectives did not direct their scholastic conduct.

In paper [3], Newly developed Web-based educational technologies offer researchers unique opportunities to study how students learn and what approaches to learning lead to success. Web-based systems routinely collect vast quantities of data on user patterns, and data mining methods can be applied to these databases. This paper presents an approach to classifying students in order to predict their final grade based on features extracted from logged data in an education Web-based system. We design, implement, and evaluate a series of pattern classifiers and compare their performance on an online course dataset. A combination of multiple classifiers leads to a significant improvement in classification performance. Furthermore, by learning an appropriate weighting of the features used via a genetic algorithm (GA), we further improve prediction accuracy. The GA is demonstrated to successfully improve the accuracy of combined classifier performance, about 10 to 12% when comparing to non-GA classifier. This method may be of considerable usefulness in identifying students at risk early, especially in very large classes, and allow the instructor to provide appropriate advising in a timely manner.

In paper [4], The graduate admissions process is crucial for controlling the quality of higher education, yet, rules-ofthumb and domain-specific experiences often dominate evidence-based approaches. The goal of the present study is to dissect the predictive power of undergraduate performance indicators and their aggregates. We analyze 81 variables in 171 student records from a Bachelor's and a Master's program in Computer Science and employ state-of-the-art methods suitable for high-dimensional data-settings. We consider regression models in combination with variable selection and variable aggregation embedded in a double-layered cross-validation loop. Moreover, bootstrapping is employed to identify.

importance of explanatory variables. Critically, the data is not confounded by an admission-induced selection bias, which allows us to obtain an unbiased estimate of the predictive value of undergraduate level indicators for subsequent performance at the graduate level. Our results show that undergraduate level performance can explain 54% of the variance in graduate-level performance. Significantly, we unexpectedly identified the third-year grade point average as The most significant explanatory variable, whose influence exceeds the one of grades earned in challenging first-year courses. Analyzing the structure of the undergraduate program shows that it primarily assesses a single set of student abilities. Finally, our results provide a methodological basis for deriving principled guidelines for admissions committees.

II. PROPOSED SYSTEM

The proposed work contains the data collection of students mark list along with the instructor handled for each subjects. Then the collected data is classified using the K-means clustering algorithm. The mapreduce programming model is utilized to process the large amount of data in hadoop framework.

K-Means Clustering algorithm proceeds as follows

a) The required number of clusters (K) must be chosen beforehand.

b) Then, it randomly selects K initial cluster centers.

c) The third step is to choose each data object of the input data set having n data objects and compare its distance To all the centers of the K clusters. The data object is added to the cluster whose centre is closest to the data object.

d) The cluster centres are re-calculated after each iteration.

e) This process iterates until the criterion function converges.

Hadoop Framework:

Hadoop MapReduce is a programming model of Hadoop framework [8]. It is developed to write applications which process huge amounts of data in parallel manner on large cluster of machines of commodity hardware. Hadoop

MapReduce works in a reliable, fault-tolerant manner. MapReduce allows for distributed processing of the map and reduction operations. As its name depicts, MapReduce has two phases.

a. Map phase

b. Reduce phase

When the input dataset is provided to a MapReduce job, it is into independent data chunks which are, then processed by the map tasks in parallel. After processing these independent chunks, MapReduce sorts the outputs of the map tasks, which are then provided to the reduce tasks as input. The input and the output of the Map tasks and reduce tasks are stored in a file-system. MapReduce also takes care of scheduling tasks, monitors every task and if any tasks fails in between the execution, it re-executes the failed task. The MapReduce framework operates these computation on a set of key/value pairs of input, and provides a set of key/value pairs as output conceivably of different types.

For example: (Input) <k1, v1> -> map -> <k2, v2> -> combine -> <k2, v2> -> reduce -> <k3, v3> (output)

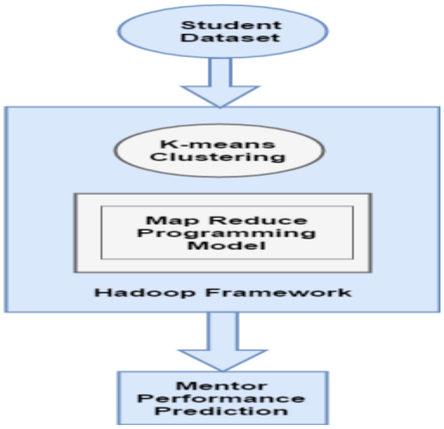
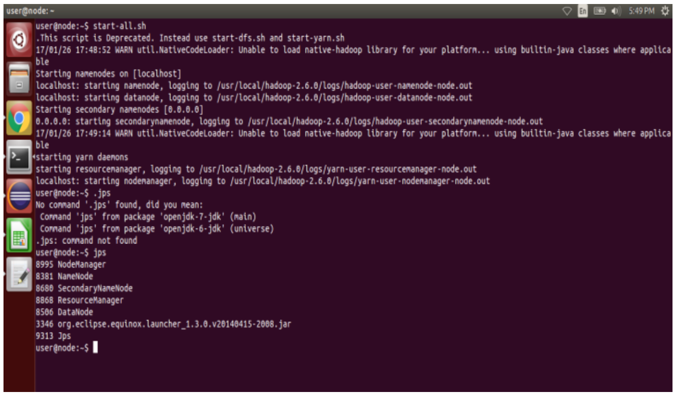


Fig 1. The System Model for predicting the mentor performance from student result

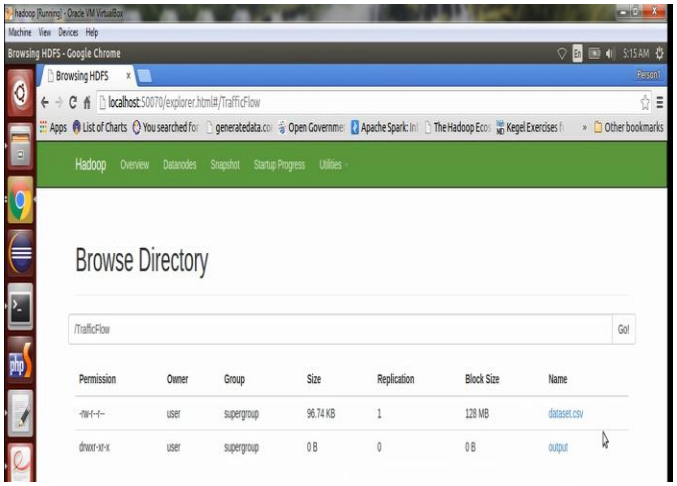
III. EXPERIMENTAL SETUP

The implementation is carried out in the Ubuntu 14.04 LTS with Hadoop 2.6.0 and the dataset is collected from various sources to predict the mentor performance.



```
user@node:~$ start-all.sh
.This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
17/01/26 17:48:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applica
ble
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-user-namenode-node.out
localhost: starting datanode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-user-datanode-node.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-user-secondarynamenode-node.out
17/01/26 17:49:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applica
ble
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-user-resourcemanager-node.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-user-nodemanager-node.out
user@node:~$ .jps
No command '.jps' found, did you mean:
 Command 'jps' from package 'openjdk-7-jdk' (main)
 Command 'jps' from package 'openjdk-6-jdk' (universe)
.jps: command not found
user@node:~$ jps
8995 NodeManager
8381 NameNode
8680 SecondaryNameNode
8868 ResourceManager
8506 DataNode
3346 org.eclipse.equinox.launcher_1.3.0.v20140415-2008.jar
9313 Jps
user@node:~$
```

Fig 2. Represent the jps service node startup in the hadoop framework.



Browse Directory

/TrafficFlow

Permission	Owner	Group	Size	Replication	Block Size	Name
-rw-r--r--	user	supergroup	96.74 KB	1	128 MB	dataset.csv
drwxr-xr-x	user	supergroup	0 B	0	0 B	output

Fig 3. Shows the Hadoop Directory for the input dataset processing.

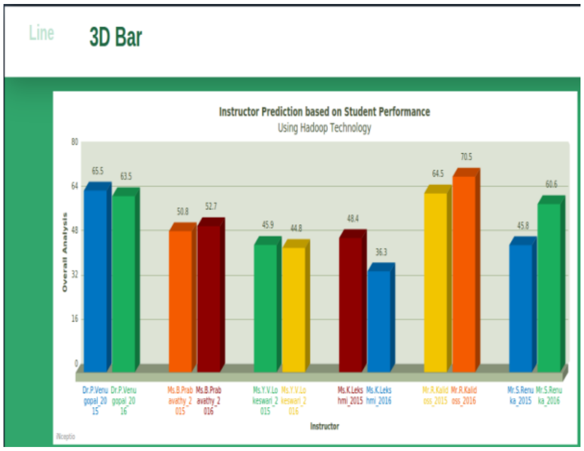


Fig 4 Represent the 3D Bar graph of the prediction using the Hadoop Framework.

IV. RESULT & PREDICTION

The dataset is collected from the various sources regarding the student marklist along with the mentor handled for each subjects. The dataset includes 10000 student for testing process to evaluate the performance of the K-means with mapreduce programming model in Hadoop framework.

Table 1: Represent the confusion matrix

	Predicted Negative	Predicted Positive	Total
Negative Cases	TP: 8760	FN: 400	P
Positive Cases	FP: 240	TN: 600	N
Total	P’	N’	P+N

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{P}}$$

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{P + N}}$$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{N}}$$

Table 2. Represent the calculated result of the accuracy, precision, recall and specificity

Model	**Accuracy**	**Precision (%)**	**Recall**	**Specificity**
K-Means	93.1	94.9	93.1	92.6
DA	90.5	94.1	90.6	90.4

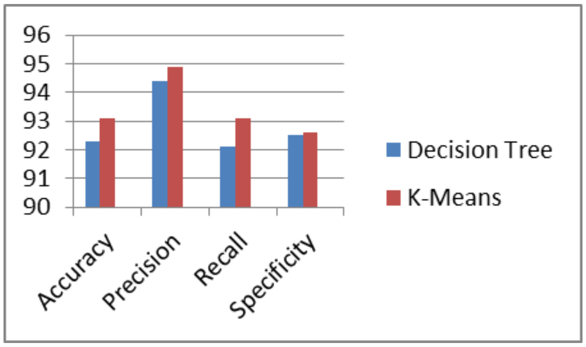


Fig 5. Comparison between the Decision tree and the K-means for parameters such as accuracy, precision, recall, specificity.

V. CONCLUSION

In recent years, educational institutional field is the raising area to deploy the bigdata approaches. The study of the related paper gives a key idea to implement the K-means algorithm with the Hadoop framework. Consequence of this work is improvising the accuracy of classification with reduced computational time by using the K-means algorithm with the Mapreduce programming model in hadoop framework to predict the mentor performance. From the result and prediction part, the calculated parameter values such as accuracy, precision, recall and specificity shows the improvised percentage than the previous algorithm.

REFERENCE

[1] M. Abaidullah, N. Ahmed, and E. Ali, ``Identifying hidden patterns in students' feedback through cluster analysis," Int. J. Comput. Theory Eng., vol. 7, no. 1, pp. 16_20, 2015.

[2] N. Delavari, S. Phon-Amnuaisuk, and M. R. Beikzadeh, ``Data mining application in higher learning institutions," Inform. Edu.-Int. J., vol. 7, no. 1, pp. 31_54, 2007.

[3] M. Goyal and R. Vohra, ``Applications of data mining in higher education," Int. J. Comput. Sci. Issue, vol. 9, no. 2, pp. 113_120, 2012.

[4] Mustafa Agaoglu, “Predicting Instructor Performance Using Data Mining Techniques in Higher Education”, IEEE Access, 2016.

[5] Jing Luan, “Data Mining Applications in Higher Education”, in www.spss.com/worldwide, 2004.

[6] B. Minaei-Bidgoli ; D.A. Kashy ; G. Kortemeyer ; W.F. Punch, “Predicting student performance: an application of data mining methods with an educational Web-based system” in IEEE access, 2003

[7] B. K. Baradwaj and S. Pal, ``Mining educational data to analyze students' performance," Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 6, pp. 63_69, 2011.

[8] S. Calkins and M. Micari, ``Less-than-perfect judges: Evaluating student evaluations," NEA Higher Edu. J., pp. 7_22, Fall 2010.

[9] J. Sojka, A. K. Gupta, and D. R. Deeter-Schmelz, ``Student and faculty perceptions of student evaluations of teaching: A study of

similarities and differences," College Teach., vol. 20, no. 2, pp.44_49, 2002.

[10] L. Coburn. (1984). Student Evaluation of Teacher Performance, ERIC/TME Update Series. [Online]. Available: http://ericae.net/edo/ED289887.htm

[11] S. A. Radmacher and D. J. Martin, ``Identifying signi_cant predictors of student evaluations of faculty through hierarchical regression analysis," J. Psychol., vol. 135, no. 3, pp. 259_269, 2001.

[12] S. M. Hobson and D. M. Talbot, ``Understanding student evaluations: What all faculty should know," College Teach., vol. 49, no. 1, pp. 26_32, 2001.

[13] D. L. Crumbley and E. Fliedner, ``Accounting administrators' perceptions of student evaluation of teaching (SET) information," Quality Assurance Edu., vol. 10, no. 4, pp. 213_222, 2002.

[14] S.-H. Liaw and K.-L. Goh, ``Evidence and control of biases in student evaluations of teaching," Int. J. Edu. Manage., vol. 17, no. 1, pp. 37_43, 2003.

[15] H. C. Koh and T. M. Tan, ``Empirical investigation of the factors affecting SET results," Int. J. Edu. Manage., vol. 11, no. 4, pp. 170_178, 1997.

[16] K. Mckinney, ``What do student ratings mean?" Nat. Teach. Learn. Forum, vol. 7, no. 1, pp. 1_4, 1997.

[17] W. W. Timpson and D. Andrew, ``Rethinking student evaluations and the improvement of teaching: Instruments for change at the University of Queensland," Stud. Higher Edu., vol. 22, no. 1, pp. 55_66, 1997.

[18] J. E. Whitworth, B. A. Price, and C. H. Randall, ``Factors that affect college of business student opinion of teaching and learning," J. Edu. Bus., vol. 77, no. 5, pp. 282_289, 2002.

[19] M. Ahmadi, M. M. Helms, and F. Raiszadeh, ``Business students' perceptions of faculty evaluations," Int. J. Edu. Manage., vol. 15, no. 1, pp. 12_22, 2001.

[20] C. R. Emery, T. R. Kramer, and R. G. Tian, ``Return to academic standards

[21] : A critique of student evaluations of teaching effectiveness," Quality Assurance Edu., vol. 11, no. 1, pp. 37_46,200.