DOI: https://doi.org/10.24113/ijoscience.v6i11.327

Application of DNN and Hadoop for Educational Big Data

Rishiram
M.Tech Scholar
Department of CSE
VITS
Bhopal (M.P), India
rishirammaskare8@gmail.com

Sumit Sharma
Professor
Department of CSE
VITS
Bhopal (M.P), India
sumit_sharma782022@yahoo.co.in

Abstract: **— In current advancement in technology such as big data, has proved promising advantages in every field. In Educational field it has showed its efficiency for deployment of new advancement in education. It can utilized for educational and personality development of students and promoted the better development of education. In this work DNN is used for assessment of students and their development.**

Keywords*:* Data mining, Big Data, Educational Big Data, performance, Hadoop.

I. INTRODUCTION

Over a few decades, the emergence of big data analytics helps entrepreneurs to explore the data manually to carry out useful patterns in the market. Big data analytics has derived various opportunities for the institutions, policy makers, educationalists, administrators and learners. The opportunities are enhanced knowledge flow and learning success over the organization, cross collaboration over the institutions become comfortable and learning effectiveness would be enhanced, cost reduction over organizing financial performance become possible and academic risk would be lowered. Through traditional application software, big data won't be processed [1]. Hence, it requires cloud based technologies like Hadoop and Spark to mine huge amount of data. This big data approach offer organizations with effective way to stay strong and active in the business. In addition to this, Hadoop platform has received attentions as it renders various advantages to the institutions and learners. This study aims at the influence of big data in the education and how the education system will be enhanced by using big data analytics [2][3].

Nowadays big data analytics has been used in the education. Besides various opportunities the educationalist experience some challenges to deploy big data analytics. The challenges are enunciating data flow, training practitioners and decision making and actions. Retaining data for the analysis is significant challenge for the deployment of educational analytics. It is difficult to access required data from the incorporated database system and hard to create data warehouse for all institutions [4][5]. Unstructured data and lack of quality can leads to essential issues. In order to create a understanding of the system among the educators, the trainees need to involve in learning the system and takes more time. It would be difficult for educators and learners to offer information in an informative way. However, the big data influence the education sector in an effective. To sort out these challenges, this study will be proposed.

Big data analytics provide effective assistance to the organization in order to use the data to determine new field in the business. This mining will develop new opportunities and enhance smart business. This data analytics has been resulting in profits, effective operations and customer relationship. Enterprise can able to obtain cost advantages as this cloud-based analytics focus on particular issues. It is significant to notice that usage of Hadoop in organizations work faster and make effective decisions as this platform has capability to determine source of data [6][7][8]. As per the customer needs, the new products and services can be produced by using analytics. Hence companies are focusing on the needs by enhancing services to fulfill the customer needs.

A. Shaping the education sector

Educational institutes like universities, colleges, schools has carried huge amount of data. It can be determined to focus on which enhance the operational effectiveness of the educational institutions [9]. Student exam results and development of educational needs is highly relied on changing educational requirements will be computed by using statistical analysis.

B. Career prediction

Big analytics helps to determine the student performance report will enhance the authority to know about the student strength and

weakness. Such report will implicates some solutions to student about the areas to be focused in future.

C. Learning analytics

Learning analytics has received enhanced attention as it provides various advantages to the institutions of higher education and enhance student retention, student success and provide accountability. Learning analytics focus on managing the capacity of the analytics such as acting on predictions and forecast behaviour. The objective of learning analytics is to enhance the prediction over time.

D. Slow progress of big data in education

In the education sector, big data is considered as the game changer in the academic performance. The learning company is helping this analytics in an efficient way.

The usage of big data in the education is increased with privacy and security concerns. As the big data focused on digitalizing data, there is no roadblocks on how to process, store and access the student learning data when preventing such data from being misused or abused [10][11]. The student learning data are stored and collected in online learning system, mobile devices and school district offices. The disconnection among these aspects leads to data security as the influence of data security breach in one database will affects the whole data system. It may also cause blocking the linkage of various databases. In order to make a significant balance between sharing and securing data, the data security protocols makes the linkage and reduce insignificant data facts. The other significant challenges in implementation of big data in education are the private information and student learning data could be used against the students even they move in educational system or workplace [12]. It allows the students to learn by offering learning materials which are enhanced by the adaptive learning algorithms and the academic performance of the students are digitally tracked. In order to resolve these problems, our proposed system will be used. The education sector may use this Hadoop platform to enhance the reliability and scalability process.

II. Proposed Methodology

In this research work a framework is designed on the basis of machine learning approach on educational big data for prediction and analysis of student performance. The proposed work is focused on following features:

- Student Performance Prediction
- Student Attendance Shortfall
- Appreciation providing to deserving students

All above discussed features can be included in one common frame work for that hybrid machine learning application is used along with concept of MapReduce because educational big data contains large amount of data and it will cost more computational time. So, by applying MapReduce the overall processing time is reduced as parallel processing is performed. The proposed flow chart is given below:

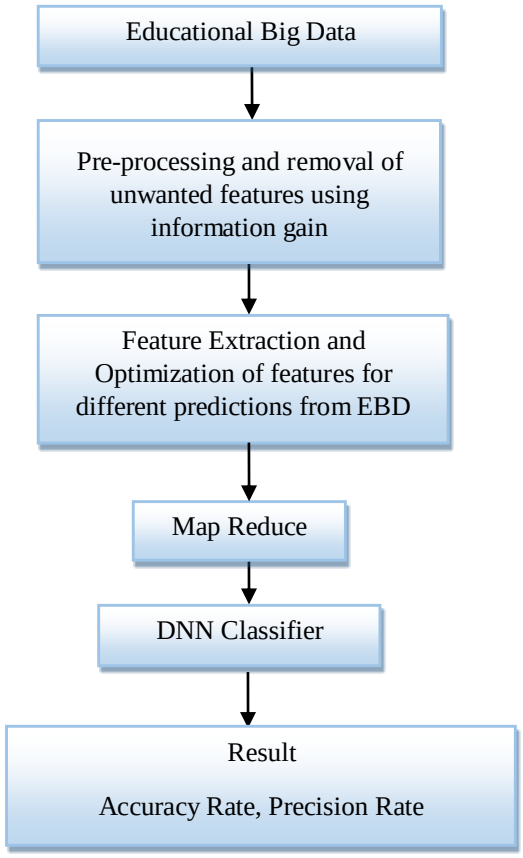

Fig.1. Proposed Flow Diagram

The selection of proper initialization techniques significantly decreases the time required to converge when trained. Data feature matrix are initialized that are arbitrarily projected over the feature vector space without any training.

1) Regularization of network

DNNs, when trained with a large number of parameters, confront the problem of overfitting. To overcome the problem of overfitting, a simple and powerful technique called dropout regularization is used. In dropout regularization, certain units from the neural network are selected randomly and ignored during training. Initially, for each hidden layer, the probability of the dropout rate is fixed to 1.0; further, it is tuned between 0.5 and 1.0 over the validation set. Because dropped units are shared within the network, the same units of the hidden layer are dropped at each node. The hidden layers in CNN do not suffer from the problem of overfitting as they are not directly involved in the process of prediction. However, due to the large size of the datasets, keeping dropout on the fully-connected layers are beneficial.

2) Activation functions

The common choices of activation functions for empirical study suggest sigmoid function, tanh function, or ReLu. However, for a given labeled input, the objective is to learn representations for fine-grained (multi-class) sentiment polarities. To attain this, a standard softmax activation is used for the output layer that takes a node's vector z(v) as input and produces prediction y(v). The output of the softmax function is equivalent to a categorical probability distribution, i.e., the probability that any of the classes are true.

$$f(x) = \frac{e^x}{\sum_{n=1}^{N} e^x N}$$

A. Training the networks

DNN are often trained with optimization techniques that need a loss function to estimate the model error. The network parameters are optimized to minimize the loss in accordance with the output of a loss function using various optimization techniques. The loss is typically measured in terms of negative log-likelihood or residual sum of squares depending upon the learning task.

Gradient descent is the most popularly used algorithm to perform the optimization of neural networks. It provides a way to minimize an objective function f(θ) parameterized by a model's parameters through updating the parameters in opposite direction of the gradient of the objective function ∇θf(θ) with respect to the parameters. The variants of gradient descent have been tried depending upon the size of data to compute the gradient of the objective function. For example, SGD performs parameter updates for each training sample x_i and label y_i.

1) Loss functions

Errors on a validation set are measured during training and stopped (early stopping) if validation error does not get improved. It is calculated using loss function by matching the target value with predicted value returned by a neural network. For multi-class problems, the most suitable loss function can be a categorical cross-entropy (CCE). In CCE, there must be the same number of output values as that of classes. The result of the final layer is passed through a softmax function to obtain each node's output as a probability between 0 and 1.

III. Result Analysis

To evaluate the performance of methodology, the proposed algorithm is simulated in following configuration:

1. Pentium Core I5-2430M CPU @ 2.40 GHz
2. 4GB RAM
3. 64-bit Operating System
4. MATLAB Platform

For simulation result, the research is focused towards implementation of feature co-relation using information gain method and DNN. For executing this simulation, a dataset of students is created on excel and then imported for analysis.

A. Performance Parameters

1) Accuracy

The result analysis is performed to find accuracy of the proposed methodology and to decide the performance rate of proposed methodology.

Accuracy = (TP+TN)/(TP +TN+FP+FN)

Where,

TP = True Positive, that means if student performance is good and predicted label also stands for good performance.

TN = True Negative, that means if student performance is poor and predicted label also stands for poor performance.

FP = False Positive, that means if student performance is poor and predicted label also stands for good performance.

FN = False Negative that means if student performance is good and predicted label also stands for poor performance.

2) Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE) is a measure of the predictive accuracy of a forecasting method in statistics, for example in estimating the trend. It usually expresses the precision in percentage and is defined by the formula:

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{Target_{value} - Obtained_{value}}{Target_{value}}$$

3) R-squared (R^2)

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

For this the research methodology is designed using information gain based feature relation identification and DNN for predicting either good or poor performance and behavior of the students in an university/school. These evaluations are performed on testing dataset. The training testing dataset is divided into 85:15 ratio, 80:20 ratio, 70:30 ratio and 60:40 ratio.

Table I: Performance Evaluation for Proposed Algorithm

No. of Students (out of 1500)	Accuracy (in %)	MAPE (in %)	R^2
225	96.27	7.6	0.02

300	90.85	6.8	0.1
450	89	8.9	0.04
600	84.2	10.8	0.1

A. Appreciation Providing

Then in next section it finds the students which may be considered for appreciation due to their good performance, co-curricular activities and their good behaviour towards teachers and students. This methodology gives an innovative direction for student's growth and development which ultimately get into direction for enhancement of growth rate of any organization and ultimately to growth of a country. The table II gives a sample of selected students out of 600 students testing sample of students for providing bonus/ Promotion. The sorted list of selected students ID of students are given in the table.

Table II: Selected Students for Bonus/Promotion

No. of Students	Selected Students Id for bonus				
10	21	43	54	65	87
	103	112	314	415	595
20	21	43	54	65	87
	103	112	314	415	595
	84	63	39	372	527
	167	259	478	503	529
30	21	43	54	65	87
	103	112	314	415	595
	84	63	39	372	527
	167	259	478	503	529
	24	165	150	76	37
	486	361	490	537	592

B. Attendance Shortfall Analysis

Table III represents the shortfall of attendance of students in each department of a university. This proposed model shows its efficiency in all respect either it is required for decision for promotion or to forecast the future shortfall of students in any department. So, that bunk nature of each student will be forecasted in prior and helps in deciding or preparing decisions to track such students and to enhance their performance.

Table III: Student Shortfall Analysis

Groups	Number of Student Shortfall in Each Department		
	CSE	ME	EX
10 Students Each Dept.	1	1	2
20 Students Each Dept	1	1	2
30 Students Each Dept.	3	3	2
40 Students Each Dept.	4	5	3
50 Students Each Dept.	4	5	5
All Students Each Dept.	6	68	39

Table IV: Comparative Performance Evaluation

Methodology	Average Accuracy (Approx.)
DNN	89
K. Abe [1]	78

The table IV and figure 2 gives a comparative result analysis of proposed work with existing work. The result shows the enhancement of proposed work with approx. 14% and the work is also extended towards the finding students which can be selected for appreciation which was not discussed in existing work.

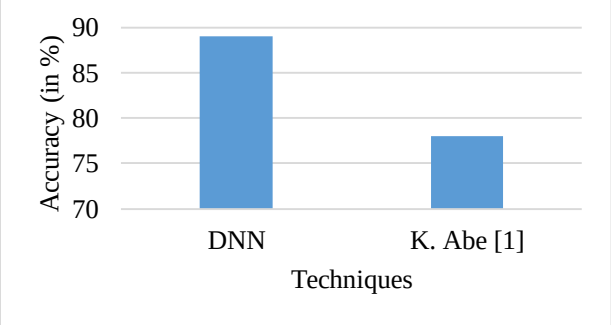

Figure 2: Comparative Accuracy Evaluation

IV. CONCLUSION

Big data technology is now the forefront of scientific and technological development, and is an indispensable technology in education data mining. cloud storage is used for storing the data and Hadoop platform for analyzing the educational data in an efficient manner. By using this framework, the education sector can enhance the student retention, improve teaching effectiveness, transform into effective decision thinking and actions and student acquisition optimization. Some of the important facts analyzed and concluded in this work are stated as below:

i. In this research work, DNN is designed to predict performance/behavior of the student. The result shows that the accuracy of model is approx. 89% and shows improved performance with existing methods by approx. 14% in terms of accuracy.

ii. This model is quite efficient for finding eligible students for finding and sorting the best and deserving students.
iii. This model also gives motivational message to existing student to improve their performance.
iv. This model can also decide to give forecasting for shortage of attendance of students with respect to department.
v. This decision support system also efficient with respect to time.

REFERENCES

[1] K. Abe, "Data Mining and Machine Learning Applications for Educational Big Data in the University," 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Fukuoka, Japan, 2019, pp. 350-355.

[2] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute; San Francisco. 2011.

[3] Zaslavsky A, Perera C, Georgakopoulos D. Sensing as a service and big data. arXiv preprint. 2013;1301.0159.

[4] Zikopoulos P, Eaton C. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media; New York. 2011.

[5] Chen H, Chiang RH, Storey VC. Business intelligence and analytics: from Big Data to Big Impact. MIS Quarterly. 2012;36(4):1165–88.

[6] Baker RS, Inventado PS. Educational data mining and learning analytics. In: Larusson A. J., White B. editors. Learning Analytics, Springer, New York. 2014; 61–75.

[7] Romero C, Ventura S. Educational data mining: a survey from 1995 to 2005. Expert Systems with Applications. 2007;33(1):135–46.

[8] West DM. Big data for education: data mining, data analytics, and web dashboards. Governance Studies at Brookings. 2012;4:1–0.

[9] Siemens G, Long P. Penetrating the Fog: Analytics in Learning and Education. EDU- CAUSE Review. 2011;46(5):30.

[10] Picciano AG. The Evolution of Big Data and Learning Analytics in American Higher Education. Journal of Asynchronous Learning Networks. 2012;16(3):9–20.

[11] C Ghosh, C Cordeiro, DP Agrawal, M Bhaskara Rao, Markov chain existence and hidden Markov models in spectrum sensing, in Proceedings of the IEEE International Conference on Pervasive Computing & Communications (PERCOM) (Galveston, 2009), pp. 1–6 122.

[12] Z. Shao, H. Sun, X. Wang and Z. Sun, "An Optimized Mining Algorithm for Analyzing Students' Learning Degree Based on Dynamic Data," in IEEE Access, Vol. 8, pp. 113543-113556, 2020.