Predicting Student Performance using KNN Classification in Bigdata Environment

A Seetharam Nagesh[1], Ch V S Satyamurty[2] and K Akhila[3]
[1] Senior Asst. Professor, CVR College of Engineering, IT Department, Hyderabad, India
Email: nageshf25@gmail.com
[2] Assoc. Professor, CVR College of Engineering, IT Department, Hyderabad, India
Email: satyamurty@cvr.ac.in
[3]B.Tech Student, CVR College of Engineering, IT Department, Hyderabad, India
Email: kathariakhila@gmail.com

Abstract: **The data mining techniques are widely used to run business in profits and also for analyzing the data for different business needs. The analysis of data has many advantages. The extraction of hidden trends in the large data is much helpful to make decisions. The data mining system also called as knowledge discovery from databases, helps to find useful predictions from the data. The patterns extracted from data mining techniques are beneficial to provide quality education in educational institutions. In this paper, the student performance is predicted by applying one of the classification algorithms KNN classifiers using the Hadoop platform, which aids to take decisions of knowledge workers such as academic council.**

Index Terms: **KNN Classifier, Data mining, Student, Big data, Academic performance.**

I. Introduction

The Datamining techniques are applied in various fields to extract the previously unknown patterns from the data. The patterns are helpful to gain knowledge by the decision makers. The decisions taken by knowledge worker are useful to improve the business process. The data mining algorithms have proved their usefulness in various application domains, a few to mention, Credit Card frauds, Sports, Health Care, Banking, and Insurance. The data mining techniques are used by researchers in the Educational domain [1] and it is known as Educational data mining. The classification or clustering techniques are used by most of the researchers to know the implicit patterns from the educational data. The Prediction of student performance [2] - [16] or predicting the grades in the subjects is helpful for stakeholders. The above-mentioned techniques are applied in the educational domain to know hidden patterns. Using the patterns, the academic committee can take decisions and implement new methods to improve the quality of education, as well as implement the best practices in the niche areas. The measures taken to improve the quality in academics will be helpful for the institute as well as the industry. The data mining consists of several algorithms for classification and clustering. Classification can be done by using the decision tree, Naive Bayes, support vector machine, neural networks, KNN algorithms. Similarly, clustering consists of different algorithms like Partition based, Density based, Hierarchical, Model-based [20], etc. In this paper, the authors used KNN classification technique for the student data implemented in big data environment using Hadoop and MapReduce.

The main motivation of the paper is to predict the student performance by knowing the grade of the student in Choice-Based Credit System (CBCS) using the KNN classification technique applied in Hadoop MapReduce Platform [22]. The model created by the authors will be of great help in the educational institutions to take appropriate measures to improve the quality of the institute.

Section II consists of Classification models, Section III includes Related work, Section IV consists of the Methodology, Section V consists of Experimental Results and Discussion and Section VI deals with conclusions.

II. Classification Models

The data mining [17] algorithms are broadly classified as supervisory and un-supervisory methods. The supervisory methods are also termed as classification. The classification is a two-step process. The first step is the learning phase of classifier model. The dataset used in the learning phase is called as training data. The training data consists of class label. The model will be built using the training data and one of the classifier algorithms. The model is applied to the test data to assign the class label. There are different classification methods used to compare the classifier accuracy. There are different classification algorithms used to build the model and classify the unknown sample. A few popular algorithms are decision tree induction, Naïve Bayes, Artificial Neural Network, Support Vector Machine, Bayesian Belief Networks, KNN, Rule-based and Regression Analysis.

The Decision tree classification algorithm can be constructed using several algorithms. One of the algorithms used to construct decision tree is C4.5. The C4.5 algorithm is an enhancement of Iterative Dichotomiser (ID3) algorithm to build the decision tree. The most popular Decision tree algorithm C4.5 divides the dataset recursively into subgroups until no subgroup can be divided further or remaining data samples belong to one group. This is the terminating criterion of the algorithm. For dividing the dataset into different subgroups, information gain is used as the measure. The information gain is the most popularly used measure. The other measures used to construct the decision tree are GINI index, Gain ratio, and Chi-square.

Unlike the Decision Tree, Support Vector Machine (SVM) divides [12], [21] the data using hyperplane to minimize the error and optimize the margin. So far, the above two techniques are linear classifier techniques, and there are nonlinear classifier methods, like the brain. Here, the researchers simulate the brain, because of this, it is called as Artificial Neural Networks (ANN). It consists of the learning phase and classification phase using weights, feedback, activation function.

K-Nearest Neighbor algorithm [18], [22] is a non-parametric method of classification. It is also called as instance based or lazy learning algorithm. The data sample can be assigned a class label by most of the nearest neighbors.

The Rule based classification uses IF-Then rules to make the classification. This is also called as antecedent and precedent.

Discriminant Analysis [19] is used to predict the categorical variable. It is useful for classifying the categorical variable. Here linear discriminant analysis is similar to principle component analysis.

III. Related Work

There is an increasing use of classification techniques in Education domain [1] in the last six years. Researchers are applying classification techniques to know the student performance. This study helps to compare the performance of the classification methods and the attribute measures they considered.

Ramaswami M, et.al [2] applied statistical techniques such as F-measure and Receiver Operating Characteristic (ROC) to find the minimum cardinality and high predictive probability by using six filter techniques and comparative study to find features which are more important.

Affendey L.S, et.al [3] collected 2427 data records and applied preprocessing, feature selection of 254 attributes and used different classification algorithms in WEKA data mining tool. They measured the accuracy of different classifier algorithms for the dataset they have considered.

Bhardwaj, K., Pal. S, [4] collected the data from the previous years and employed Bayesian classification algorithm on the student results and predicted student outcome in the examinations. The outcome of the prediction will help the institute to improve the pass percentage and also students' grades.

Bekele, R., et.al, [5] applied the Bayesian approach to predict the student performance of Ethiopian colleges. They have collected 574 samples by performing the survey, and finally, they considered 514 samples by taking the probability factor greater than 0.64. They formed Bayesian belief network to predict performance to improvise learning process.

Osmanbegovic E, et.al [6] applied three supervised algorithms such as Decision Tree, Naïve Bayes, Neural Network based algorithms to predict the student grades in the upcoming examinations and learning accuracy of the students. This helps to improve the results and help the faculty to take appropriate methodology to enhance the student performance. They have implemented the Decision tree algorithm for prediction of student performance such as grades or CGPA. Mladen D et.al [7] built another model to know about the student who completes the graduation.

Ogunde A.O., Ajibade D.A [8] used Iterative ID3 algorithm for student grades of the University of Nigeria. The accuracy of the algorithm is 79.556.

Romero, C, Ventura, S [9], have employed decision tree to find the academic performance and used Chi-square automatic interaction detection (CHAID) to perform the classification.

IV. Methodology

The students academic performance is given considerable importance by various committees inspecting the college, especially in technological institutions. The academic performance of the student is determined by the marks in the internal examinations and the end semester examination. The student performance in each semester is determined by the total marks obtained in each subject and, the cumulative total of subjects and labs in that semester. In the curriculum, there are two internal examinations. The internal examination marks are a combination of three components. These are mid marks, assignment, and attendance. There are two internal examinations and average of the two examinations is awarded as internal marks. The end semester examinations are conducted, and minimum marks must be obtained in the end examination to pass, but the overall marks for passing in each course is 40(internal plus external).

A. Data Preparation and Preprocessing

The data is of three academic years in a subject in the second year I semester of Department of Information Technology of CVR College of Engineering. The data consists of absents marked as AB but modified as zero for experimentation purpose. The data may contain an erroneous value. The incorrect value will be converted to either upper bound or lower bound as preprocessing step.

B. Data Selection

In Engineering, during each semester, the curriculum supports mostly six subjects and two labs with approximately 28 credits. The Object-Oriented programming through Java subject is considered out of the six different subjects. The subject is considered because of its familiarity. The above said subject internal marks are considered with Mid1, Assignment1, Mid2 and Assignment2. The sample training dataset is as shown in Table I.

TABLE I
STUDENT_TRAIN DATASET

Mid1	Mid2	Assign1	Assign2	Grade
12	3	10	10	C
13	13	10	10	B+
5	6	10	10	B+
1	7	10	10	B
9	13	10	10	B
14	13	10	10	A+
3	13	10	10	B
3	4	10	10	F
13	13	10	10	F
2	5	10	10	B
14	17	10	10	S
15	13	10	10	A
15	18	10	10	S
1	1	10	10	F
3	4	10	10	P
1	4	10	10	P
16	15	10	10	S
15	5	10	10	A
15	18	10	10	S
12	4	10	10	F
3	13	10	10	C
4	13	10	10	A
4	13	10	10	A
14	16	10	10	A+
18	16	10	10	S

The above data is sample data of student_train data. The data is useful to build the model. The test data does not contain class label, so we should assign the class label using the model built in the learning or training phase. Table II is the test data without class labels.

TABLE II
STUDENT_TEST DATASET

Mid1	Mid2	Assign1	Assign2
19	19	10	10
19	19	10	8
17	15	8	8
18	14	10	8
16	16	10	9
14	13	10	10
20	19	10	10
19	20	10	10
5	12	10	8
13	15	10	7

C. KNN Classification

The algorithm works by considering number of the class labels in the train data. The dataset that is considered has nine class labels namely, S, A^+, A, B^+, B, C, P, F, AB. So, 'K' is Nine. The nine class labels of the training dataset are used to build the model. The class labels are the grades of the students. The grades are assigned based on the percentage of marks obtained by the student in that subject. The grades are defined as follows:

Grade: – Final Marks obtained in the subject are split into 9 classes : S is > 80% , A+ is >75% and < 80% , A is >70% and <75% , B+ is >65% and <70% , B is >60% and <65% , C is > 50% and <60% , P is >40% and <50%, F is <40%, and AB for Absent.

Now, Euclidian distance is considered as a measure to calculate the distance of the test label with the centers and assign the class label to the test sample by majority voting or nearest neighbor. The test data sample will be assigned the class label by determining which center is the nearest one.

Generalizing Test Phase

1. Determine the value of 'k' (input)
2. Consider the training dataset by storing the model and class labels of the data points.
3. Load the sample of test data points from the testing dataset.
4. Calculate the Euclidian distance amongst the 'k' closest neighbors of the testing data point from the training dataset based on a distance metric.
5. Assign the class label which is nearest to the training dataset to new data point from the test data.
6. Repeat steps 3, 4, 5 until all the data points in the testing data are classified.

The formula to calculate Euclidean distance is as shown in (1),

$$\sqrt{\sum_{i=1}^{n} (xi - yi)^2} \qquad (1)$$

In this manner, it will assign the class label to all the test samples and measure the accuracy of the model.

The above algorithm is implemented in Hadoop Map Reduce environment.

The <key, value> pair of the dataset is assigned to mapper function. The mapper function computes the distance with each class and lists it out. Next in the reducer phase, the first 'k' neighbors in ascending order of the distance measured are considered and the majority vote is taken and the class label is assigned to the test data sample according to the majority vote.

The input and output directories are organized and they are named as traindatafile and testdatafile. Having the files, apply the K-Nearest Neighbor method in a distributed environment of Hadoop by following the algorithms given below to design the Map and Reduce methods for K-Nearest Neighbor.

Algorithm 1 for Mapper Function

- **Procedure** KNN MAPPER Design
- Create key-value pair list to maintain test dataset
- testlist = new testlist
- Load file containing test data file

```
•  load testfile
•  Update key-value pair in test dataset
•  testlist <= testdatafile
•  Open train dataset file
•  Open traindatafile
•  Load training data points one at a time and calculate distance measure with every test data point
•  Distance (traindata, testdata)
•  Write the distance of test data point from all the train data samples with their respective class labels in ascending order
•  Compare testdatafile with testdata(dist., label)
•  Call Reducer
•  End procedure
```

Algorithm 2 for Reducer Function

```
•  Procedure for KNN Reducer Design
•  Load the center value of "k"
•  Load testdatafile
•  Open testdatafile
•  Load testdatafile points one at a time
•  Read testdatafile
•  Initialize counters for all class label
•  Set counters to zero
•  Look through top 'k' distances to each data point and increment counter for each class label it belongs
•  for i = 0 to k
       o  counter++
       o  Assign the class label to testdatafile point depending on the highest value
       o  testdatapoint =classlabel($counter_{max}$)
•  Update outputfile with class label for testdatapoint
•  outputfile = outputfile + testdatapoint
•  End procedure
```

Thus, the above algorithms, KNN mapper and KNN reducer will assign the class label to each test data point.

V. Results And Discussion

The K-Nearest Neighbor algorithm was implemented on Hadoop cluster Setup, consisting of 4 nodes in CVR College of Engineering lab. One node acts as Name Node and Job Tracker and other three nodes act as Task Trackers and Data Nodes. All the nodes are of Intel Pentium Dual Core G3240 4th Generation, 8GB DDR3 RAM. The operating system is Ubuntu 14.04.4 LTS. The programming for coding the MapReduce is JAVA 1.8.0. Apache Hadoop 2.7.1 and all the nodes are created as a cluster.

The training data points are stored in student_train, and test data points are stored in student_testdata. The data files are copied into the Hadoop Distributed File System (HDFS). The Hadoop MapReduce runs iteratively until all the data points of test file are classified. The process is started with the small dataset and increased to larger dataset and runs multiple times, so as to classify the data points more accurately.

TABLE III
STUDENT_OUTFILE

Mid1	Mid2	Assign1	Assign2	Grade
19	19	10	10	S
19	19	10	8	A+
17	15	8	8	A
18	14	10	8	A+
16	16	10	9	A+
14	13	10	10	A
20	19	10	10	S
19	20	10	10	S
5	12	10	8	F
13	15	10	7	C
13	17	10	8	A
14	19	8	8	A+
13	16	10	8	B
13	15	10	8	A

Table III indicates the student grades obtained in the final examination by running the algorithm. Suppose a student gets a grade C or P or F as the class label. The academic council can take appropriate measure to improve the student's academic performance before the examination is really conducted.

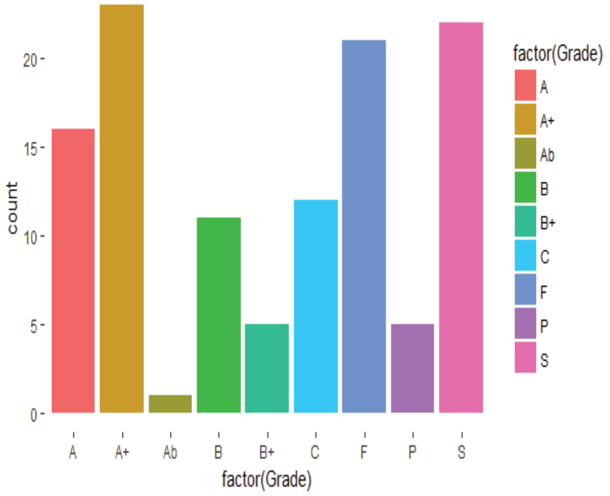

Figure 1. Grades and number of students in each grade

Figure1. indicates the student count in each grade. This also helps the academic council to assess the performance of a class. In the Figure1. the grade 'F' count is 22 and 'C' are 13 students. These grade students can be identified and implement appropriate methods to reduce the count and improve the students' performance.

VI. Conclusions

KNN classification technique is implemented in Hadoop MapReduce environment to classify or predict the class labels of the students appearing for the End Semester Examinations in one subject. By analyzing the result, we can improve the performance of the weak students by mentoring and conducting individual sessions, to pass in the examination, which in turn will help us to take appropriate steps to improve the quality of education in the Institution. The experiment is conducted for only one course of the Second year I semester OOPS through Java subject. But the analysis can be carried out for various other courses where there are more number of failures, as well as, to know the end semester cumulative grade obtained by the student.

References

[1] A. Peña-Ayala, ``Review: Educational data mining: A survey and a datamining-based analysis of recent works". Expert Syst. Appl. Volume 41 (4), 1432-1462, 2014.

[2] Ramaswami M, and Bhaskaran R, "A Study on Feature Selection Techniques in Educational Data Mining". Journal of Computing. Volume 1(1), 2009.

[3] Affendey L.S., Paris I.H.M., Mustapha N., Sulaiman M. N. and Muda Z. "Ranking of Influence Factors in Predicting Student's Academic Performance". Information Technology Journal Volume 9 (4), 832-837, 2010.

[4] Bhardwaj, K., Pal. S "Data Mining: A prediction for performance improvement using classification". International Journal of Computer Science and Information Security. Volume 9(4), 2011.

[5] Bekele, R., Menzel, W. "A Bayesian approach to predict performance of a student (BAPPS): A Case with Ethiopian Students". Proc. IASTED International Conference on Artificial Intelligence and Applications, 2005.

[6] Osmanbegovic E., Suljic M. "Data mining approach for predicting student performance". Economic Review-Journal of Economics and Business. Volume 10(1), 2012.

[7] Mladen D., Mirjana P. B., Vanja Š., "Improving University Operations with Data Mining: Predicting Student Performance". International Journal of Social, Behavioral, Educational, Economic, and Management Engineering Volume 8(4), 2014.

[8] Ogunde A.O., Ajibade D.A. "A data Mining System for Predicting University Students F=Graduation Grade Using ID3 Decision Tree approach", Journal of Computer Science and Information Technology, Volume 2(1), 2014.

[9] Romero, C, Ventura, S. "Educational Data Mining: A Review of the State-of-the-Art. IEEE Transaction on Systems, Man, and Cybernetics, Part C" Applications and Reviews. Volume 40(6), 2012.

[10] Kovacic, Z. "Early prediction of student success: Mining student enrollment data". Proceedings of Informing Science & IT Education Conference, 2010.

[11] Cortez P, Silva A. Using data mining to predict Secondary school student performance. Journal of information science Volume 2(6),2013.

[12] F. D. Kentli and Y. Sahin, ``An SVM approach to predict student performance in manufacturing processes course," *Energy, Edu., Sci. Technol*. Volume 3(4) pp. 535-544, 2011.

[13] V. Ramesh, P. Parkavi, K. Ramar, Predicting student performance: a statistical and data mining approach, International Journal of Computer Applications Volume 63 (8), 2013.

[14] R. S. Bichkar "Predicting Students Academic Performance Using Education Data Mining", World Journal of Computer Application and Technology, Volume 2(2) 43-47, 2014.

[15] T. M. Christian, M. Ayub, Exploration of classification using nbtree for predicting students' performance, in: Data and Software Engineering (ICODSE), 2014 International Conference on, IEEE, pp. 1–6, 2014.

[16] K. F. Li, D. Rusk, F. Song,"Predicting student academic performance". Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, 2013.

[17] S. Singhal, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering". International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume. 2(6), 2013.

[18] Mustafa Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education". IEEE. Translations and content mining are permitted for academic research only. Volume 4 ,2169-3536, 2016.

[19] R. W. Klecka, "Discriminant Analysis". Sage Publications, 1980.

[20] J. Zimmerman, K. H. Brodersen, H. R. Heinimann, and J. M. Buhmann, "A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance". Journal of Educational. Data Mining. Volume 7(3), pp. 151_176, 2015.

[21] S. T. Jishan, R. I. Rashu, N. Haque, R. M. Rahman, "Improving accuracy of student's final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique". Decision Analytics Volume 2 (1), 2015.

[22] Prajesh P Anchalia, Kaushik Roy, "The K-Nearest Neighbor Algorithm Using MapReduce Paradigm". Fifth International Conference on Intelligent Systems, Modelling and Simulation, IEEE Explorer, pp 513-518, 2015.