January-February 2020
ISSN: 0193-4120 Page No. 12529 - 12535

Post Graduate Students' Performance Data in Competitive Examinations

[1] Ananth.Y.N,[2]Dr.Narahari.N.S,
[1] Senior Research Fellow and Ph.D student, Jain University, Bangalore
[2] Professor, Department of Industrial Engineering and Management, R.V.College of Engineering Autonomous institution affiliated to VTU Belagavi.

Article Info
Volume 82
Page Number: 12529 - 12535
Publication Issue:
January-February 2020

Abstract:
There are huge amount of data available with regard to competitive examinations for admission into Post Graduate Courses in the state of Karnataka. Typically the entrance examination for admission to the MCA course is based on Multiple Choice Question (MCQ) formats. The current research work is aimed at examining the suitability of the MCQ based formats in deriving information on the measurement of aptitude of the students for the course. The research involves analyzing students' performance data in competitive examinations. A novel attempt in this work is to establish the linkages between past performance results and the question paper as an instrument to measure the aptitude of the students. In order to analyze the student's performance, the power of the artificial neural networks is being attempted. Artificial neural networks(ANN) simulate the ability of the human brain to perceive underlying patterns in a given data set not perceptible using several other well-known statistical data analysis techniques. ANN have a set of inputs and outputs, the outputs being calculated through many iterations of transformations carried through a set of middle layers ,called as hidden layers. The number of these hidden layers can vary depending on the problem that is being tackled. For the transformation between the inputs and the outputs, the hidden layers of ANN use many triggering functions. In the current work, the students' results, taken from the public domain website are used as the data set. These are analyzed by using ANN technique. The output from the analysis helps in the formation of clusters based on the marks obtained by the students in the competitive examination. In the next step the relationship between the marks obtained by the students and the type of questions is established. The MCQ based questions are graded using the standard benchmark as in the Bloom's taxonomy. The ANN model can also help in predicting the results of the students. The output from the ANN model is then fed into an RDBMS platform for establishing relationships. Here, research concepts from data mining tools, clustering and document comparison have been used. The CART algorithm is then used to identify the clusters and groups of students based upon their performance and the questions that they have attempted. The computations in this research use the R programming environment for analysis using ANN and CART. The research study yielded the results in the formation of students clusters –Further research yielded the information on the type of the questions and benchmarking assigned to the questions. It was also possible for building association of the students' cluster with the questions types attempted. The insights gained helped in obtaining suggestions for modification and improvement of the questions types, so as to develop better measures of aptitudes.

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 24 February 2020*

Keywords: *ANN, CART, Decision Support System, Entrance Examination*.

1. INTRODUCTION

The entrance examination for MCA course is taken up by a large number of students. So it is imperative to study the nature of questions that come up in the question paper. The current work compares the question paper with the standard given by Bloom's taxonomy. It also studies the past performance of the students using artificial neural network- tries to get the mapping of the students vs the questions they have answered. Then by having quality bench marks,

January-February 2020
ISSN: 0193-4120 Page No. 12529 - 12535

the work suggests some questions in certain scenarios. This suggesting part has been done using RDBMS software and a front end

2. LITERATURE SURVEY:-

Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.[2]Artificial Neural Networks have been used in classifying the students into categories [1][3].Neural networks have high acceptance ability for noisy data and high accuracy and are preferable in data mining.[4].The use of neural networks is prominent as a data mining technique in education. Neural networks are being used to categorize the students as to falling into pass and fail categories.[5].Neural networks has been extensively used in analyzing data from Moodle logs[5].Neural networks have the ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by human beings or many other computational algorithms[2]

Classification methods are used in[8] to classify students' final grade of students. Brijesh Kumar Bharadwaj and Sourabh Pal[9] have used Classification task to analyze the performance of students. Decision tree algorithms are applied on students' internal marks data to predict the performance in the final exam.]Decision trees are used in data mining to study historical data and on the basis of data analysis and its rules, one can predict the results.[11].Classification algorithms like C$.5,ID3 and CART have been applied on engineering students' data to predict their performance in the final exam[12]

Bloom's taxonomy has been used in [13] using feature selection approaches for classifying teachers 'question into different cognitive levels. Bloom's taxonomy, along with k-means clustering techniques have been used in [15] to determine positive and negative cognitive skills with respect to reading comprehension tasks. A study has been conducted in [16] to find out whether examinations comply with the requirements of Bloom's taxonomy.

DATA: The data used in this work are the EXCEL files containing the results of MCA entrance examination for three years and also the question papers of the same entrance examination.

3. METHODOLOGY:

There are essentially two parts to this work. The first one is applying ANN algorithms to analyze the results and classify the students into categories based upon their marks. We will have four categories of students based upon this classification, namely Excellent, Good, Poor, Worst. We deduce, a rough mapping of what each of these clusters of students have answered in the question paper.After the first round of analysis, we make use of the libraries in R, to work with the CART algorithm to further classify the students clusters. The second part is using Bloom's taxonomy to benchmark the questions in the question paper. In addition, other categories in the bench marks would be created. Once this is done, the questions are going to be classified as belonging to three orders, ie high order, middle order and low order. Now the mapping between the students clusters and the questions would give us a relationship such as

St1→Q1[8],Q2[10], Q3[10]

Where st1 is a student cluster, and Q1, Q2, Q3 are question categories , the numbers 8, 10, 10 indicate how many questions in each of these categories have been answered by the cluster of the students St1.

Separately, we have a mapping of the questions, which should be the "ideal" kind of questions which should appear in the question paper. The above result is used to give supportive suggestions for the question paper setter in order to frame a "good" question paper. This is done through an RDBMS part and an UI using Visual Studio.

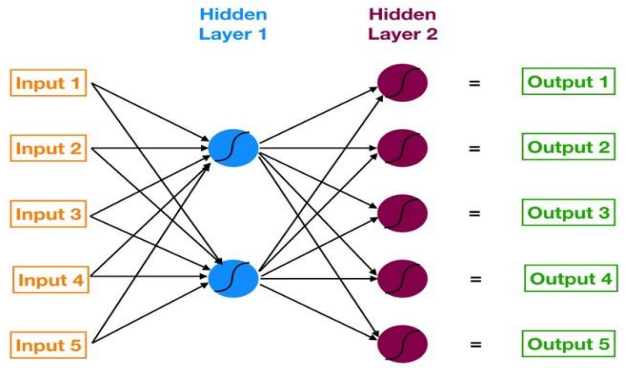
4. NEURAL NETWORKS:-

Neural networks are a system of analysis which works close to the way that the human brain works.

January-February 2020
ISSN: 0193-4120 Page No. 12529 - 12535

Neural networks contain a set of neurons connected by hidden layer of weights. These layers of weights together determine what the output for the next layer is. The working of neural networks can be depicted by the following diagram.

FIGURE 1:



The number of layers depends on the particular problem that we are trying to solve.

Neural networks work similar to the functioning of the human brain. These models are biologically inspired rather than an exact replica of how brain actually works. Neural networks have been very promising system in many forecasting applications and business classification systems due to their ability to "learn" from the data, their non parametric nature (ie no rigid assumptions), and their ability to generalize.

CART ALGORITHM: CART is a classification algorithm – the full expansion being Classification and Regression Trees. –This is a set of algorithm referred to as "decision trees", but on some platforms like R, this is being implemented as a set of libraries called "RPART"

BLOOM'S TAXONOMY:-

Educational psychologist Benjamin Bloom developed Bloom's taxonomy in 1956 to categorize intellectual skills, which are significant in the learning process. According to the taxonomy developed by Bloom, there are six cognitive levels. Bloom gave a system of methodology for the teaching and learning process by formulating six categories of verbs corresponding to these six levels. The categories are:

1. Knowledge
2. Understand
3. Apply
4. Analyze
5. Synthesis
6. Evaluate

Each of these verbs in the chart signifies certain level of the maturity of the student in the learning process. These verbs are guiding verbs and it is not necessary that the verbs should appear exactly as they are, in the question paper. Near equivalents can also be present.

5. APPLICATION OF THE METHODOLOGIES:APPLICATION OF NEURAL NETWORKS IN THE WORK USING R:-

In R, there is a set of libraries called neuralnet-which has got functions to perform analysis using neural network. These functions primarily implement neural networks with back propagation..

Our work, takes input the results obtained from the public domain website, which has got columns for the student regno and the rank .We preprocess the data in order to obtain the dependency of the rank on the marks. This will give us the clusters with a certain set of marks scored by the candidates. For example, if student reg.nos from A001 to A020 have got the ranks which are nearby, they fall into a cluster. To get such ranks we do the bifurcation of the clusters, during the data pre-processing stage as follows.

With this ranking order, we use the logic to know at least in some definite terms, as how many students would have answered a certain set of questions, which is again is random and probabilistic.

This method can be explained briefly as follows.

1. Since the question paper has got questions with four options, there can be 4 * 4 = 16 types of answering, in other words, sixteen clusters of students with the question answered in 16 different ways

January-February 2020
ISSN: 0193-4120 Page No. 12529 - 12535

So, n= N/16
Where n = no of groups
N = Total no of students in the rank list
2. The result sheet lists the ranks in the descending order of ranks
So, we start by assigning certain marks, nearest to the highest for the first rank student, divide the list into sixteen groups.
3. Within a group, there doesn't have to be equal no of students, but we create the groups in such a way that a certain rank prevails over a set of students and the no sixteen remains unchanged.
For example, if N=16000
Then n = 1000
So, we look at the first 1000 students, identifying each group as n1, n2, n3 …n16.
Then we look at the boundaries, where the rank is going to change, If a certain rank prevails over a certain number of students, then we stretch that grouped till the end of that rank. For example,

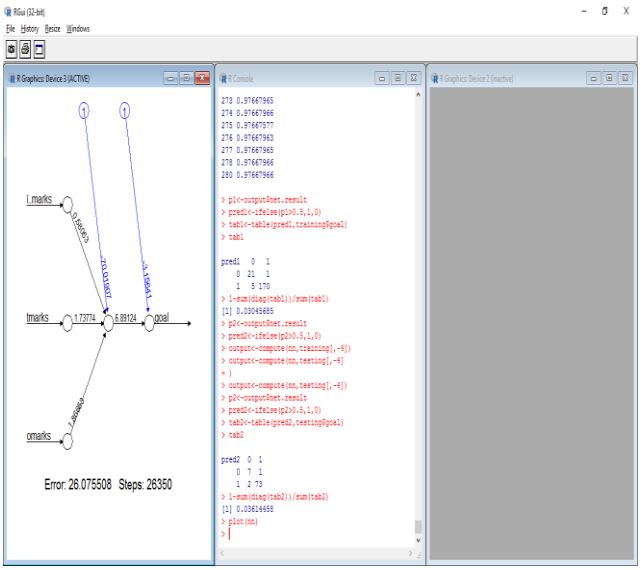
Rank	RegNo	Cluster no
1	A0001	1
2	A0002	1
3	A0003	1
4	A0004	1
4	A0005	1
4	A0006	1

.

.201	A220	2
202	A221	2
202	A222	2

Then we feed this data to the neural networks engine in R.
The result of this exercise will give us , in probabilistic measures , how many students would have answered a certain set of questions .For example , if the two clusters A001 to A200 , have a certain ranking, it is possible to ascertain how much of the 2 marks questions and how much of the one mark question they would have answered .For this purpose, we use ANN again, by creating three more columns, tmarks storing the number of two marks questions, omarks storing the number of one mark questions , goal, which is a binary variable storing whether the student got the seat or not and employing them in the analysis. The distribution of the number of two and one marks questions will be known by this analysis, by using ANN. The weights which get assigned to the nodes in the hidden layers will tell us, how much importance, that each of the type of the questions should be assigned to.
The following screen shows the application of neuralnet libraries to the data, where in a sample of 280 students has been taken and analyzed. The data has been analyzed by applying the Resilient Back Propogation-RProP algorithm of ANN.The run was done by having one hidden layer, five layers and also by having five repetitions to the algorithm. The following screenshot shows ANN algorithm running on a set of data.

SCREEN 1:



The data was divided into 70% training data and 30% testing data.The ANN algorithm was consistent in both the cases by generating misclassification error of 3% for both the partitions of the data set.
The confusion matrix for the above trial is given below:

January-February 2020
ISSN: 0193-4120 Page No. 12529 - 12535

Table 2:CONFUSION MATRIX FOR TRAINING DATA :

1	0	1
0	21	1
1	3	170

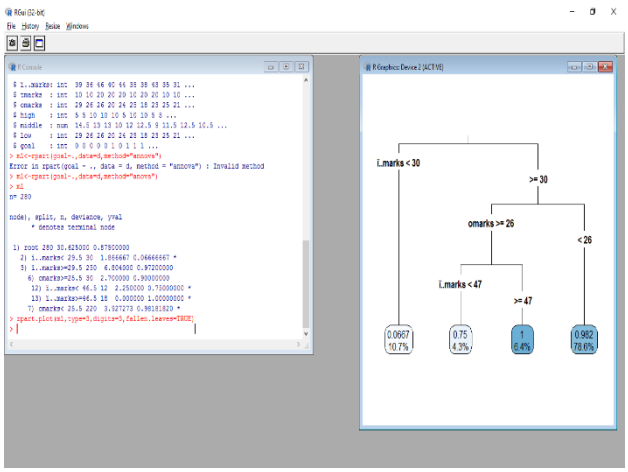
Table 3:CONFUSION MATRIX FOR TESTING DATA :

1	0	1
0	7	1
1	2	72

IMPLEMENTING REGRESSION TREES:

Regression trees have been generated for the data, using the R library rpart. In the following screenshot, a set of 280 samples of students have been used to run the rpart algorithm in R.

SCREEN 2:



The sample output shows that the decision tree has split the data into branches, depending on the marks that they have scored. The leaves of the tree depict the percentage of the students, who have scored the corresponding marks to obtain the seat

INTERPRETATION OF THE DATA ANALYSIS:

The output from this analysis gives us the number of student clusters, who have answered a certain portion of the two mark and one mark questions. From the above sample run, we can see that 170 students have answered the questions in a certain way. The weights that are generated in each of the hidden layers of the ANN output gives us the percentage of the two marks and one marks questions that the student clusters have answered. Combining the results of the analysis involving ANN and rpart , we can deduce the clusters of students who have answered a certain set of two and one mark questions. In the above screen showing the output from the classification tree , 10.7% of the students have scored marks less than 30,of those who have scored marks greater than 30 , 4.3% have scored marks less than 47 and 6.4 % have scored greater than 47 and so on. If we take this data and the weights from the ANN analysis output we can find out what percentage of the students with a certain marks have scored what percentage of two marks and one mark questions. The sample graph for the rpart analysis is not showing any two marks splits because of the data fed into it. Had there been sufficient data of two marks, the graph would have shown the splits on two marks questions also.

In the further part, the actual question paper is going to be analyzed, to find out the types of questions that exist. Then a comparison of the students' choice of answers is made to know the student aptitude. Then, the decision support system would suggest alternatives to the questions in the question paper.

THE DECISION SUPPORT SYSTEM:

In the analysis of the question paper, apart from the classifications given by Bloom's taxonomy, questions can be classified on the basis of the following parameters.

1. Mathematical:-
2. Logic based:-
3. Quantitative:-

These parameters can be given the ranks from 1 to 3, 1 being the highest and 3 being the lowest

At this point, text mining techniques are used to classify the questions. For this, we do the following.

1. Rank each question in the question paper. Based on the keyword contained in the question, the question is going to be compared with the verbs in Bloom's taxonomy or equivalents of such keywords. Also, the above classification is used to be further

January-February 2020
ISSN: 0193-4120 Page No. 12529 - 12535

classified and a suitable rank is given to the question. For example, a question containing the keyword "bus" and "sending" and "data" would be classified as a low order question. A question containing "probability" "dice" would be a "middle order "question. A question containing "database" "integrity" is considered as a "high order" question. Questions are actually compared like this by text mining libraries in R. Each question paper is being compared with a standard set of keywords, which contain keywords segregated on topic and syllabus. This is done in many iterations, changing the standard file of data, each time. Fort example, a question paper can be tested with standard keywords like "bus", "two's compliment", "how", "can", "he", "prime minister", to get the keywords for the low order questions, and with standard keywords like "database", "algorithm", "probability", "regression line", "analysis", to get the keywords for the high order questions. The text mining libraries in R would be able to get the equivalence as well as the number of repetitions of such words.

This comparative analysis will tell us which are the keywords that should be present in the question paper. After this, the analysis part done in R will also tell us which are all the keywords that are there in the question paper but not in the standard set of keywords. Those words are picked up again and they are analyzed further whether they should be present or not.

After the case by case analysis like this, cosine similarity of these corresponding documents is done.

By this kind of an analysis, it has been found that on an average, there are 30% of the "high order" questions,20% of "middle order" questions and 50%" low order" questions in the question papers.

The question paper for the future would then be designed based upon the following parameters:

1. No of students in clusters
2. Student ranks
3. Rankings given to the questions based on the types.
4. Predictive analysis involving all the above

The decision support would then come up using an user interface for integration:

The working of this part is as follows:-

1. The paper setter has to input the question to the system
2. Based on the analysis done hitherto, the system would respond with suggestions of alternate keywords.

For this to be possible, we evolve a database, which works based upon the ranks that we have generated. This can be elaborated as follows.

The database contains one table called Keywordrank :

This would store the rank assigned to each keyword in the question paper in a field called keyrank , in the prospective system. There is one more table called oldkeywordrank –This has got a field called oldkeyrank- which stores the rank for each of the keywords, which is already there in the question paper, the rank being assigned to it , from all the analysis explained earlier.

When a question paper setter starts inputting the questions – the rank from the oldkeywordrank of the keywords in the question paper are retrieved-If it is a suitable keyword, then the UI would let it included in the question paper. Or else , the keywordrank table is consulted to find the nearest rank of the required keyword. If the keyword inputted has a rank less than the proposed keyword , then the suggestion for excluding the old and including the new keyword is made .Or else , the question is accepted.

6. CONCLUSIONS AND DISCUSSIONS:

In this paper, ANN algorithm and rpart algorithm are applied on students' marks data to analyze and predict the clusters of students who have scored a certain range of marks and what percentage of question types have been answered by them. This analysis would give us an insight as to what is the aptitude of the students. In the second part, question papers of the entrance examination are being analyzed with Bloom's taxonomy as the frame of reference. The researcher also presents his own types

January-February 2020
ISSN: 0193-4120 Page No. 12529 - 12535

of questions that are there and not there in the question paper. This will lead to recommending types of questions which could test the aptitude of the students correctly.

REFERENCES

1. Data Mining Algorithms to Classify Students:- Cristóbal Romero, Sebastian Ventura, Pedro G. Espejo and César Hervás {cromero, sventura, pgonzalez, chervas}@uco.es Computer Science Department, Córdoba University, Spain --- Educational Data Mining 2008 The 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008 Proceedings.
2. Brijesh Kumar Baradwaj, Saurabh Pal-Mining Educational Data to Analyze Students‟ Performance-(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
3. Nitya Upadhyay :- Educational Data Mining by Using Neural Network-International Journal of Computer Applications Technology and Research Volume 5– Issue 2, 104 - 109, 2016, ISSN:- 2319–8656
4. Priyanka Gaur:- Neural Networks in Data Mining-International Journal of Electronics and Computer Science Engineering 1449 Available Online at www.ijecse.org ISSN- 2277-1956
5. Delgado, M., Gibaja, E., Pegalajar, M.C., Pérez, O. Predicting Students' Marks from Moodle Logs using Neural Network Models. Current Developments in TechnologyAssisted Education, Badajoz, 2006. pp.586-590.
6. Breiman, L. Friedman, J.H., Olshen, R.A., Stone, C.J. Classification and Regression Trees. Chapman & Hall, New York, 1984.
7. Data Mining: A prediction for Student's Performance Using Classification Method-Abeer Badr El Din Ahmed , Ibrahim Sayed Elaraby-World Journal of Computer Application and Technology 2(2): 43-47, 2014 http://www.hrpub.org DOI: 10.13189/wjcat.2014.020203
8. Brijesh Kumar Baradwaj, Saurabh Pal, Data mining: machine learning, statistics, and databases, 1996.
9. Efficiency of decision trees in predicting student's academic performance -s. Anupama kumar and dr. Vijayalakshmi m.n
10. Kolo David Kolo, Solomon A. Adepoju, John Kolo Alhassan,"A Decision Tree Approach for Predicting Students Academic Performance", International Journal of Education and Management Engineering(IJEME), Vol.5, No.5, pp.12-19, 2015.DOI: 10.5815/ijeme.2015.05.02
11. Surjeet Kumar Yadav, Saurabh Pal Head- Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification - World of Computer Science and Information Technology Journal (WCSIT) - ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012
12. Anwar Ali Yahya, Addin Osman, Ahmed Abdu Alattab, Educational Data Mining: A Case Study of Teacher's Classroom Questions
13. Bloom B. S. (1956). Taxonomy of educational objectives, Handbook I: The Cognitive Domain. New York: David McKay Co Inc. [4] Carner, R. L. (1963) Level of Questioning, Education, , 83, 546-55
14. Terry Peckham, Gord McCalla, Mining Student Behavior Patterns in Reading Comprehension Tasks