International Journal of Data Mining and Emerging Technologies
DOI: 10.5958/j.2249-3220.2.2.010

IndianJournals.com

Research Article

Predicting Students' Performance using the Machine Learning Approach in Educational Data Mining

K. Shanmuga Priya[1*] and A.V. Senthil Kumar[2]

[1]M.Phil Scholar, Department of Computer Science, [2]Director, Department of MCA, Hindustan College of Arts and Science, Bharathiar University, Coimbatore – 28, Tamil Nadu, India

[*]Email: priyasiva.0229@yahoo.com

[2]Email: avsenthilkumar@yahoo.com

ABSTRACT

The objective of this paper is to provide assistance to students, which when given at the appropriate level is invaluable in the learning process. Not only does it aid the student's learning process but it also prevents problems such as student frustration and floundering. Students' key demographic characteristics and their marks determined using a small number of written assignments, internal exams, attendance, etc., can constitute the training set for a regression method in order to predict their performance and marks. The scope of this work compares some of the state-of-the-art regression methods in this thesis of predicting students'marks. A number of experiments have been conducted using six methods, which were trained using datasets provided by the Hindustan College of Arts and Science.

Keywords: Educational Data Mining, Decision Tree, Machine Learning, Rule Learning, CN3 Algorithm

1. INTRODUCTION

This paper examines the usage of machine learning (ML) techniques in order to predict a student's performance in the ML system. Even though ML techniques have been successfully applied in numerous domains such as pattern recognition, image recognition, medical diagnoses, commodity trading, computer games and various control applications, to the best of our knowledge, no previous attempt was made in the presented domain. Thus, we use a ML approach for each one of the most common ML techniques, namely

- Decision Trees,
- Instance-Based Learning and
- Rule Learning

Indeed, it is proved that learning algorithms can predict student performance with satisfying accuracy long before the final examination. In this paper, we also try to find the characteristics of students that mostly influence the induction of the methods. This will reduce the information needed to be stored and will also speed up the induction. For our study, the "MCA" course of the Hindustan College of Arts and Science provided the dataset.

As mentioned above, the ML approach is widely used for pattern recognition, image recognition, medical diagnoses, commodity trading, computer games and various control applications. However, this proposed method is used for students' mark prediction.

The inputs of this paper will present details of the students, which will predict their future details. The input details are:

- First semester marks, subject-wise.
- Attendance report of the corresponding semester.
- Internal marks of the corresponding semester.
- Health report of the corresponding semester.

- Behaviours and activities.
- Areas of interest.
- Marital status.
- Gender.

These are the key inputs for the prediction process. Using the ML approach and decision tree methods, a student's final semester marks can be predicted.

This prediction can be done through the primary input as the first semester marks of the corresponding semester. In case a student is interested in a particular semester, naturally he or she will score marks in a particular subject. However, he or she will not score the same marks in all subjects. Hence, here the prediction will give more priority to the area of interest. The supporting inputs are the internal marks, health report, activities, etc. These provide less prediction details, but play an important role. In case a student scored more marks in the first semester, but unfortunately his health was not good during the semester exams, then the prediction will focus on his first semester marks, internal marks of the second semester, area of interest and present health report.

Initially, all input details will be separated for the database and the data set. These datasets work under the decision code written on paper using the ML approach. Hence, the prediction can be carried out.

1.1 DECISION TREE

Basically, decision trees area flowchart-like structure in which the internal node represents test on an attribute in that, each branch represents the outcome of test and each leaf node represents a class label. A path from root to leaf represents classification rules.

Decision trees are widely used for operations research and more specifically for decision analysis, to help identify a strategy most suitable for prediction. If in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or an online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

Decision trees influence diagrams, utility functions and other decision analysis tools, and methods are taught to undergraduate students in schools of business, health economics and public health; they are also examples of operations research or management science methods.

2. RELATED WORK

Prediction plays a vital role in all departments. In case of biology for prediction, a biological response of molecules from their chemical properties is used. Using various kinds of prediction methods in various departments has become a trend. Our proposed work is based on predicting marks for college students.

Fora university academic performance in Colombia [1], a survey of 850 students with their high school marks and college marks was conducted. High school marks highly affect the result of college marks. Students were categorised into two different cases, namely best case and worst case.

Busato *et al*. [2] say that motivation given to students during higher studies will reflect their achievements throughout the educational period.

In their research, Busato *et al*. [3] justified that psychology students are self-motivated and highly skilled in intellectual ability, learning style, personality, achievement motivation and academic success.

Minnaert and Janssen [4] investigated the additive, beneficial effect of regulatory activities on top of verbal, numerical and diagrammatic intelligence in the prediction of academic performance. About 500 freshmen of different study domains participated in this research.

The findings supported both the mixed and the independency model of the relationship between intelligence and cognitive skills. Analyses of variance revealed significant main effects of verbal and numerical (crystallised) intelligence, and also of cognitive regulatory activities on academic performance. The effect of diagrammatic (fluid) intelligence on academic performance was just short of being significant. Implications for further research and for educational practice are being discussed.

3. PROPOSED WORK

The CN2 induction algorithm is a kind of learning algorithm for rule induction. It is designed to work the training data is imperfect and also designed for prediction. The basic ideas are generated from the aq algorithm and the ID3 algorithm. As a consequence, it creates a rule set like that created by aq but is also able to handle noisy data like ID3. Most of the similarities lie between CN2 and ID3 algorithms. Here, we used CN2 for prediction along with the improvement of ID3 methods.

Basically, college students will be involved in highly matured activities; they are above the level of school students. However, their behaviour and attitude are decided from the school age, including their memory skills. According to this concept, for the prediction of marks for college students, their previous study marks are much important. Hence, here, all possible input methods are taken such as their tenth and twelfth marks, attendance percentage, current semester marks, internal marks, behaviour, health conditions, creativity and marital status.

4. DATASET

The dataset was collected from the Department of Computer Science, Hindustan College of Arts and Science. The data size was about 90 to 100. Most of the information mentioned above was collected as per academic data. All data used were obtained originally and confidentially was well-maintained. From the dataset, very minimum data were omitted due to irrelevant order of data.

Table 1: Attributes Used

Attributes	Values	Description
10m	Tenth marks	In percentage <100
12m	Twelfth marks	In percentage<100
Semester 1	Subject 1	ca+ca+s1 = SM1
Semester 2	Subject 2	ca+ca+s2 = SM2
Semester 3	Subject 3	ca+ca+s3 = SM3
Semester 4	Subject 4	ca+ca+s1 = SM4
Semester 5	Subject 5	ca+ca+s1 = SM5
Semester 6	Subject 6	ca+ca+s1 = SM6
LM1	Lab Mark 1	ca+ca+l1 = LM1
LM2	Lab Mark 2	ca+ca+l2 = LM2
ATT	Attendance	% < 100
HL	Health	Poor/Fair/Average/Healthy
BH	Behaviour	Poor/Fail/Average/Excellent
CR	Creativity	Poor/Fail/Average/Excellent
MS	Marital status	Yes/No

4.1 DATASET ERROR RATES

When working with real-world datasets, it is not possible to know exactly when irrelevant data start to occur, which type of data is present or even if something irrelevant has occurred. Hence, it is not possible to perform a detailed analysis of the behaviour of algorithms in the presence of the concept of data mining using only pure real-world datasets. In order to analyse the effect of low/high-diversity ensembles in the presence of data mining and to assist the analysis of prediction, we first used artificial datasets. Then, in order to reaffirm the analysis of datasets, we performed experiments using artificial datasets.

4.2 Data Analysis

This section presents an analysis of the accuracy of data analysis of the ensembles/strategies. Accuracy is the average accuracy obtained by the prediction of each example to be learned, before its learning, calculated online. The rule used to obtain the frequential accuracy on time step is presented in the study by Baena-Garcýá *et al*. [5]:

$$acc(t) = \begin{cases} acc_{ex}(t), & \text{if } t = f, \\ acc(t-1) + \dfrac{acc_{ex}(t) - acc(t-1)}{t-f+1}, & \text{otherwise,} \end{cases}$$

where access is 0 if the prediction of the current training example ex before its learning is wrong and 1 if it is correct; and f is the first time step used in the calculation.

In order to analyse the behaviour of the ensembles before and after the beginning of data, the accuracy shown in the graphs is reset whenever a drift starts (f2f1;Nþ1g). The learning of each ensemble is repeated thirty times for each dataset. The online ensemble learning algorithm used in the experiments is the modified version of online bagging proposed in the study by Minku *et al*. [6]. As commented in that section, Minku *et al*. [6] showed that higher/lower s produces ensembles with higher/lower average *Q*-statistic (lower/higher diversity).

4.3 CN3 ALGORITHM

CN3 (Input Data, Values, Attributes, TMP SM1=FM1 repeats for SM2, SM3, SM4, SM5, SM6, LM1, LM2)

Step 1: Create index value for the input data.

Step 2: Create root node for the index values for 10m and 12m for best and worst case, respectively.

Step 3: Return +1 for best case and -1 for worst case.

Step 4: Value return ca+ca+s1 = SM1 the root node will contain the total of SM1. It performs a null return value

Step 5: Execute Step 4 +SM1. If int s SM1=+3 else SM1=0, SM1 returns.

Step 6: SM1 which contains value also evaluated for the data contains in the root node and repeat for SM2, SM3, SM4, SM5, SM6, LM1, LM2 along with values.

Step 7: Node contains SM1 return to AT %<100. Returns to FM1 + AT %. Now data contain value in FM1, which will provides the final result.

Step 8: FM1+or - HL for P=-1,F=0,A=0,H=0, returns FM1.

Step 9: FM1+ or - BH for P=-1,F=0,A=0,E=0, returns FM1.

Step 10: FM1+ or – CR for P=-2,F=-1,A=0,E=0 returns FM1.

Step 11:FM1 + or – MS for Y=-1 or N= 0.

Step 12: Repeat for all SM2, SM3, SM4, SM5, SM6, LM1, LM2.

Step 13: End the process.

Step 14: Return the value.

4.4 ROOT EXECUTION

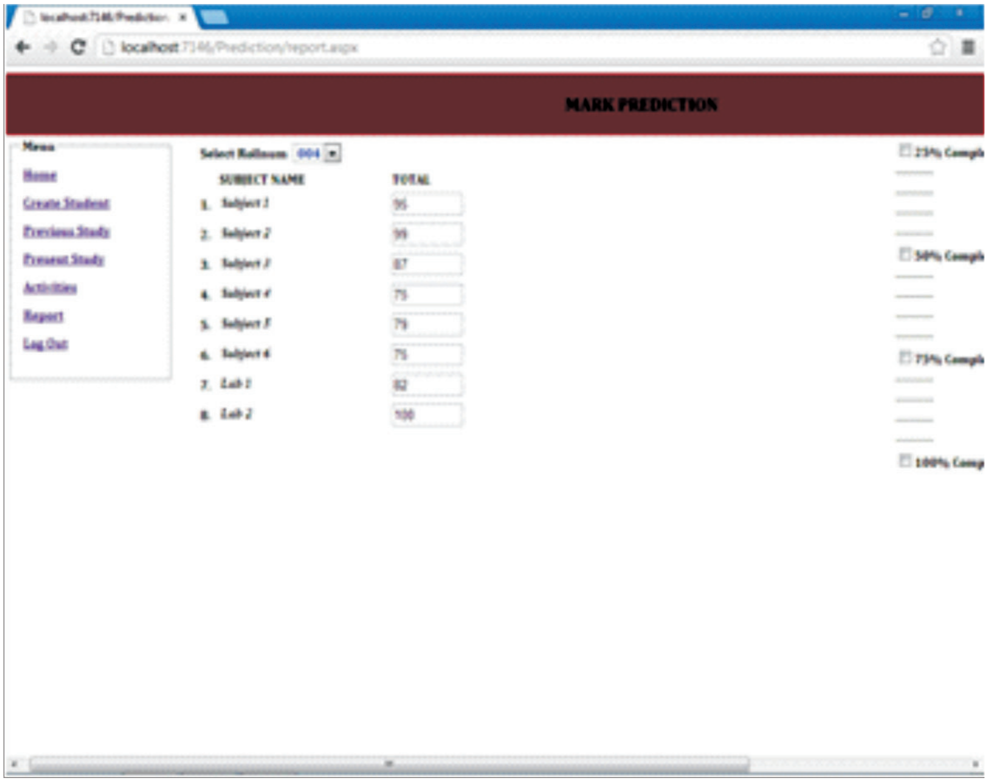
10m= 85,12m=80=Bestcase= ca+ca+s1 = SM1+inst sub checked +3 else 0

AT=SM1>90=0, 75 to 90= -1,<75=-2.HL= P=-1, F=0, A=0, H=0, BH = P=-1,F=0,A=0,E=0

CR = P=-2,F=-1,A=0,E=0, MS = Y=-1 or N= 0. Value return in FM1 and repeat for all.

4.5 RETURN VALUE

85-80=5=<5=best case (SM1=+1)SM1=82=83 lk sub y=+3=86(SM1+AT=<75=-1)=85.repeat HL=P=-1,85-1=84(BH=F=-1,SM1-BH=84-1=83), (CR=A=0, SM1+CR =83+0=83), (MS=Y=0, CM1+MS=83) = 83



MARK PREDICTION

Select Rollnum: 004

	SUBJECT NAME	TOTAL
1.	Subject 1	95
2.	Subject 2	99
3.	Subject 3	87
4.	Subject 4	75
5.	Subject 5	79
6.	Subject 6	75
7.	Lab 1	82
8.	Lab 2	100

Figure 1: Report for Mark Prediction

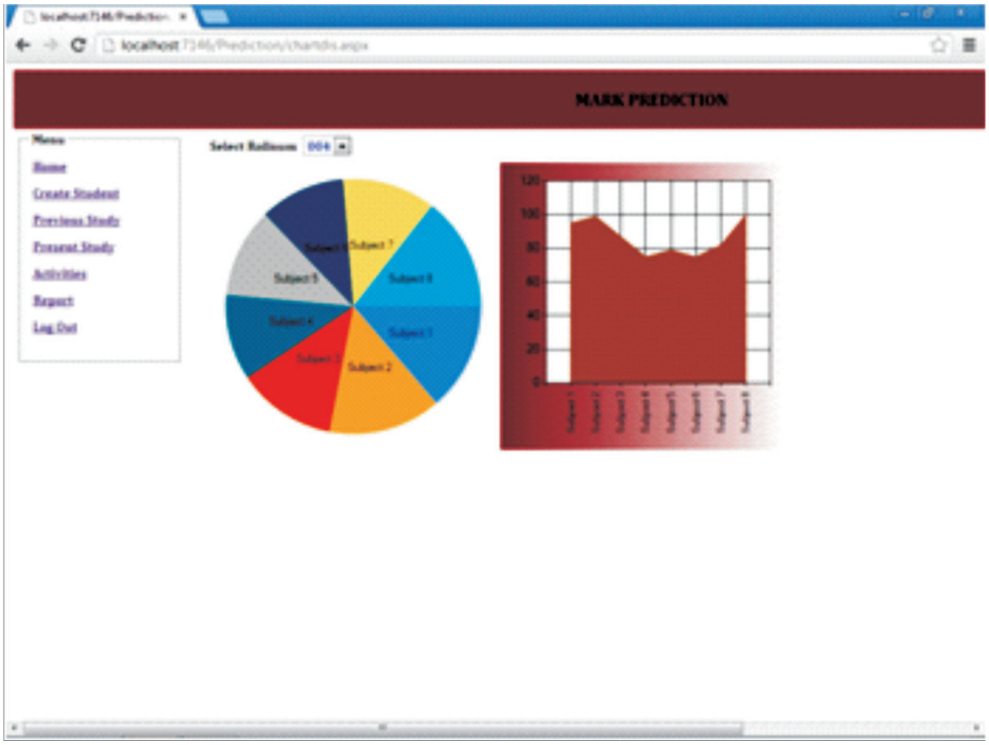


Figure 2: Chart View of Mark Prediction

CONCLUSION

The main objective of this paper to predict the students marks using data mining technique. So here using CN3 algorithm and methods result obtained successfully. The result data was verified with the original data set obtained from Hindustan College of arts and science. Thus the goal was achieved and predicted marks are up to the range from the result obtained. This research contribution will help professors to predict the marks of the student and they can improve the educational quality of the student.

REFERENCES

[1] Ardila, A. 2001. Predictors of University Academic Performance in Colombia. *International Journal of Educational Research,* 35: 411-417.

[2] Busato, V.V., Prins, F.J., Elshout, J.J. and Hamaker, C.1999. The Relation between Learning Styles, the Big Five Personality Traits and Achievement Motivation in Higher Education. *Personality and Individual Differences*, 26: 129-140.

[3] Busato, V.V., Prins, F.J., Elshout, J.J. and Hamaker, C. 2000. Intellectual Ability, Learning Style, Personality, Achievement Motivation and Academic Success of Psychology Students in Higher Education. *Personality and Individual Differences*, 29: 1057-1068.

[4] Minnaert, A. and Janssen, P.J. 1999. The Additive Effect of Regulatory Activities on Top of Intelligence in Relation to Academic Performance in Higher Education. *Learning and Instruction,* 9: 77-91.

[5] Baena-Garcýá, M., Del Campo Avila, J., Fidalgo, R. and Bifet, A. 2006. Early Detection Method, Proc. Fourth ECML PKDD Int Workshop Knowledge Discovery from Data Streams (IWKDDS '06), 77-86.

[6] Minku, F.L., White, A. and Yao, X. 2010. The Impact of Diversity on On-Line Ensemble Learning in the Presence of Data Mining, IEEE Trans. Knowledge and Data Eng., 22(5): 730-742, http://dx.doi.org/10.1109/TKDE.2009.156, May 2010