Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 25 (2016) 326 – 332

Procedia Technology

Global Colloquium in Recent Advancement and Effectual Researches in Engineering, Science and Technology (RAEREST 2016)

Student academic performance prediction model using decision tree and fuzzy genetic algorithm

Hashmia Hamsa*, Simi Indiradevi, Jubilant J. Kizhakkethottam

Department of Computer Science and Engineering, Musaliar College of Engineering and Technology, Pathanamthitta, 689653, India

Abstract

The research on the educational field that involves Data Mining techniques is rapidly increasing. Applying Data Mining techniques in an educational background are known as Educational Data Mining that aims to discover hidden knowledge and patterns about student's performance. This work aims to develop student's academic performance prediction model, for the Bachelor and Master degree students in Computer Science and Electronics and Communication streams using two selected classification methods; Decision Tree and Fuzzy Genetic Algorithm. Parameters like internal marks, sessional marks and admission score were selected to conduct this work. Internal marks are the combination of attendance marks, average marks obtained from two sessional exams and assignment marks. Admission score for degree students is the weighted score obtained from 10^{th} and 12^{th} examination marks and entrance marks. In the case of master's degree students, it includes degree examination marks and entrance marks. Resultant prediction model can be used to identify student's performance for each subject. Thereby, the lecturers can classify students and take an early action to improve their performance. Systematic approaches can be taken to improve the performance with time. Due to early prediction and solutions are done, better results can be expected in final exams. Students can view their academic information and updates. Reputed companies having a tie-up with the institution can search students according to their requirements.

Peer-review under responsibility of the organizing committee of RAEREST 2016

Keywords: Educational Data Mining; Classification; Prediction; Decision Tree; Genetic Algorithm; Fuzzy Logic

* Corresponding author
E-mail address:hashmia.hamsa@gmail.com

Peer-review under responsibility of the organizing committee of RAEREST 2016
doi:10.1016/j.protcy.2016.08.114

1. Introduction

Education is a very important issue regarding the development of a country [1]. The main objective of higher education institutions is to present quality education to its students. One way to accomplish the higher level of quality in higher education scheme is by predicting student's academic performance and thereby taking early actions to improve student's performance and teaching quality. The relevant knowledge is hidden with the educational data set and it is extractable during data mining techniques. The present paper is planned to validate the capabilities of data mining techniques in the background of higher education by offering a data mining model. In this work, classification task is used to evaluate student's performance [2]. Constant evaluations of student's performance have to be done for each subject. So that the exact area, the student lose his/her marks can be identified. That helps the lecturer to take necessary actions like more attention to the student for that particular subject, teaching in a different way that he/she can grasp quickly, conducting exams, etc. Finally, that improves student's calibre and academic status. The application of Data Mining in the Educational framework is referred as Educational Data Mining (EDM).

The appliance of analytics in the educational background has increased in the last decade. Ferguson presents in [3] three drivers for this to arise: primarily, the volume of data that are composed of educational institutions have seriously improved; secondary driver is the use of e-learning: it has helped to collect data, still brought some learning issues such as possible lack of motivation and difficulties for the educators to collect direct feedback regarding the mood, level of interest or even the understanding of the students; after all, the political concerns: countries are getting a superior understanding of the significance of higher education in their development and government have an attention in improving it, to propose enhanced learning opportunities that direct to better academic results.

In this work, two data mining approaches are proposed to predict student's performance. Prediction is done using two algorithms: Decision Tree (DT) and Fuzzy Genetic Algorithm (FGA). Prediction is carried out with academic records along with initial academic information.

The rest of the text is ordered as follows: section II present background study and related work; section III present the data sets and preparation; the proposed data mining model is presented in section IV; section V describe the experimental setup; results and analysis are presented in section VI; finally, the conclusions.

2. Background Study and Related Work

The study in [4] identifies that Bayes classifier performance result was decreased by adding more academic records. This may be caused by the assumption of independence required by the algorithm. For getting accuracy in prediction, selected attributes should be relevant and noise free. By adding more relevant attributes the accuracy can be increased. Related studies are:

2.1. Classification Methods

Ina classification task, the objective is to assign a predefined label or class to a record based on a set of known attributes. An important feature of a classification model is that it is built by part of the data, also known as the training set, which is used to train the model. All the attributes in the subset are known, even the class. After the model is built, it is used to assign the label to new records where the class attribute is unknown. To build the models, two techniques are used: DT and FGA.

A DT is a representation made of nodes and arcs where an internal node present a decision based on attribute values and the arcs stand for the option made in the node. It ends on a leaf node, which represents the label or the class to be assigned. To categorize a record with DT, it starts from the root node and goes one level down at a time that depends on the results of the condition tested on every node; when it ends on a leaf node, the record is classified according to the label on the leaf node. The knowledge represented by DT be extracted and presented in the form of IF-THEN rules [5].

Fuzzy Logic (FL) addresses applications that resemble human decision making with a skill to generate exact solutions from certain or approximate information. The use of FL based techniques is for either improving GA behaviour or modelling GA components; the results obtained have been called FGA. A fuzzy fitness finding (FFF) mechanism guides the GA through search space by combining various criteria/features that have been identified as

governing factors for the formation of the clusters.

2.2. Prediction of Academic Performance

Academic records and initial academic information's are used for prediction [6]. Initial academic information includes tenth, twelfth, entrance examination marks; and degree and entrance marks for master degree students. Weightage is given for each parameter in three categories like eight for marks 80 and above, five for marks in between 60 and 80, 2 for marks 60 and below. The total of weight is calculated and termed as admission score. Academic records include attendance, two sessional marks, assignment marks and internal marks. Internal mark a combination of average marks obtained from two sessional exams, assignment marks and attendance. From prediction, students at risk and not, as in [7]; to pass a subject are identified. The results will assist the educational institutions to improve the excellence of teaching after evaluating the marks scored by the students in an academic year. Student's expert area can be added by lecturers, helps companies for recruitment.

3. Data Set and Preparation

Data set is the whole data used for mining. Relevant data for prediction is collected through schedules like admission time, daily attendance, assignment submission time, examination conducting by a lecturer at scheduled time. Here it includes 120 and 48 students from bachelor and master degree program respectively. Among 120 students it includes two batches of 30 students from degree computer science department and another two batches of 30 students from degree electronics and communication department. Likewise, two batches of 12 students from master's computer science and two batches of 12 students from master's electronics and communication department are the other dataset. The data set used for learning is called training set and data set used for testing is called the test set. Stratified sampling is used to get the training set and the remaining is used for testing.

For mining purpose, the relevant attributes as in [8] are admission score, average for two sessional marks and internal marks. Prediction is not possible if student details are not entered or not valid. Following tables shows the data set for predicting first-year mathematics result for each 30 students from computer science department:

Table 1: Data set for Mathematics result prediction

Roll No	Internal	Sessional	AdmScore
1	47.5	25	24
2	44	24	24
3	46	23	24
4	48	25	21
5	47	22	24
6	30	12	15
7	40	23	24
8	39	24	9
9	37	18.5	15
10	32	23	15
11	42	24	18
12	48	25	24
13	47.5	24	24
14	45	22.5	24
15	44	20.5	24
16	35	16.5	18

17	34.5	13	15
18	21	12	6
19	22	11	12
20	32	15	16
21	34	12	12
22	35	13	10
23	35	10	9
24	26	12	8
25	29	14	10
26	39	17.5	11
27	43	20.5	15
28	34	18	11
29	32	15	12
30	28	10	10

Following tables shows the data set for predicting second-year big data result for each 12 students in computer science master degree.

Table 2: Data set for Big data result prediction

Roll No	Internal	Sessional	AdmScore
1	48.5	25	32
2	47	24	29
3	46	23	26
4	48	25	26
5	38	15	20
6	43	25	32
7	44	25	32
8	29	14	20
9	26	18	29
10	25.5	14	17
11	28	19	12
12	29	18	12

4. Proposed Data Mining Model

This section introduces the student classification models to predict the student performance due to low academic performance, that uses the student's initial academic information (all gathered in the admission process); and the academic records of current academic periods. The proposed model is seen in Figure. 1.

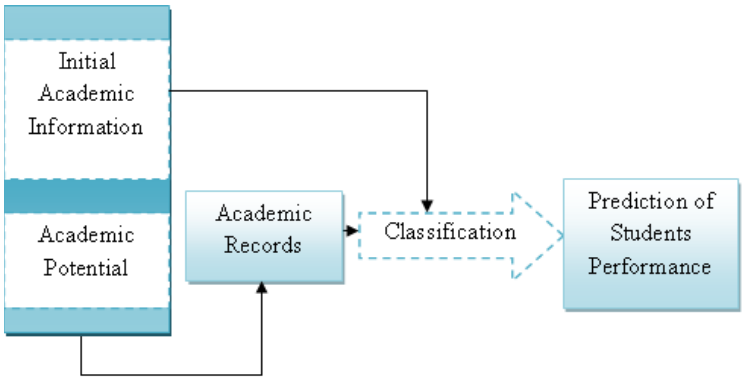

Fig.1.Proposed Data Mining Model

5. Experimental Setup

DT and FGA are used for implementation. The attribute set used for both algorithms are same: internal marks, session marks and admission score. Session marks are considered separately because not to avoid the chance for students who scored better in those fields only. C4.5 algorithm [9] is used to implement decision tree. Working of the algorithm is as follows:

- Decision tree is a classifier resembles a tree structure
  Decision node: denote a test on a single attribute; Leaf node: specify the value of the target attribute; Branch: split of attribute; Path: a disjunction of test to make the final decision.

- DT classifies instances by starting from the root node of the tree and moving towards the leaf node.

For the working of decision tree, D- the data partition; attribute_list- set of selected attributes and attribute_selection_method are required as input. Attribute_selection_method is a procedure to determine the splitting criterion that "best" partitions the data tuples into separate classes. This criterion consists of a splitting_attribute, either a split point or splitting subset.

The tree starts as a single node, N the root node. If the tuples in D are of the same class, C then the path returns a leaf node labeled with class C. If attribute_list is empty, then the path returns a leaf node with the majority class in D. Otherwise, the algorithm calls attribute_selection_method to determine the splitting criterion. Attribute with highest informational gain is selected as splitting attribute. The node N is labeled with the splitting criterion, which works as a test at the node. A branch comes from node N for each of the outcomes of the splitting criterion. Let the splitting attribute be SpA, then if SpA is discrete-valued, then one branch is grown for each known value of SpA; if SpA is continuous-valued, then two branches are grown, corresponding to SpA<= split point and SpA> split point and if SpA is discrete-valued and a binary tree is produced, then the test is of the form SpA an element of Sa, where Sa is the splitting subset for SpA. The same process recursively applied to form a decision tree for the tuples at each resulting partition. If any of the terminating conditions like all tuples in partition D belongs to same class or no remaining attributes to partition the tuples or no tuples for a given branch, is true the recursive partitioning is stopped. Thus the resulting decision tree is returned as output.

The FGA in [10] has two computational elements that work together: The Genetic Algorithm (GA) and Fuzzy Fitness Finder (FFF). The working is shown in Figure.2.

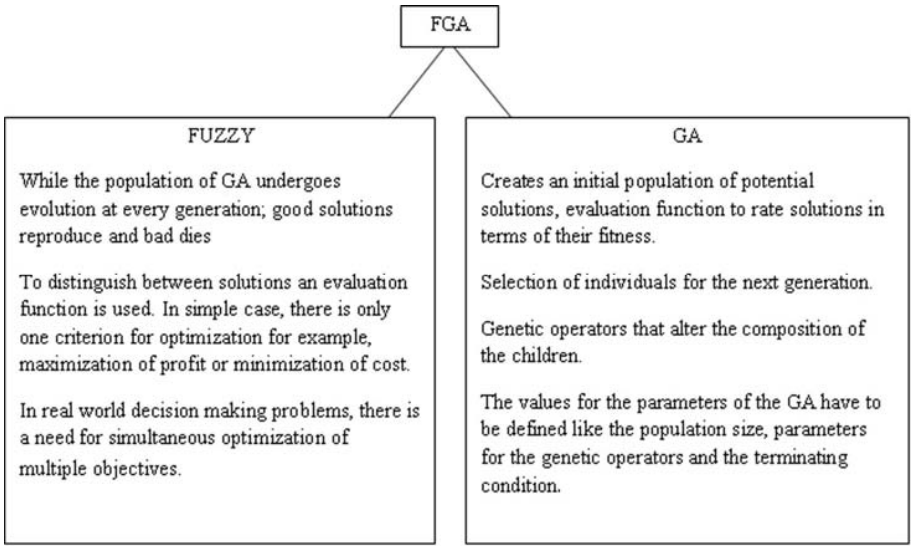

Fig.2.Fuzzy Genetic Algorithm working

6. Results and Analysis

In DT, predefined strict decisions are followed, so students at the barrier of success will be identified as at risk. So in lecturer's point of view, average-level students will also be under the care of experts. So in the final exams, a good result can be expected. In FGA, a student having a low score in some attributes, have a chance to be at safe due the high score attained from other attributes which make students comfortable. Along the graph, the list of students and their status will be displayed. Companies can search students according to their requirement and the list of safe students who are expert in their field will be displayed. Following Figure.3 and Figure.4 are the resulting graphs obtained after applying DT algorithm and FGA for subject's mathematics and big data for bachelor's and master's degree respectively.

Fig.3. Prediction of mathematics result

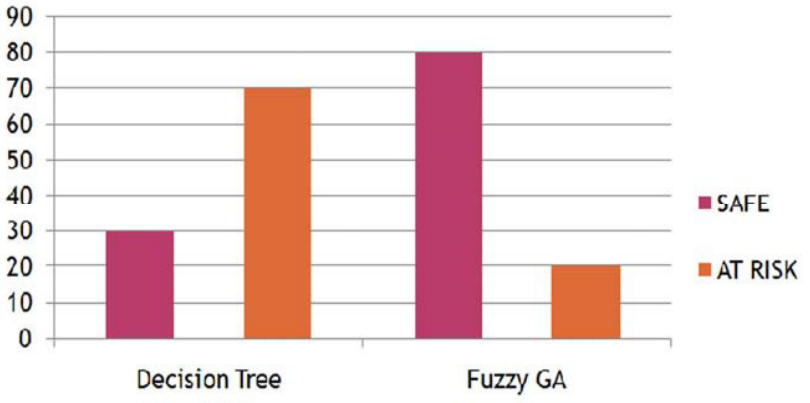

Fig.4.Prediction of big data result

7. Conclusion

Prediction of student's academic performance in bachelor's and master's degree for each subject was done independently using decision tree algorithm and fuzzy genetic algorithm. Results from decision tree algorithm made more students at risk class, which makes lecturers a decision to take more care for those students. That helps to expect a better and almost cent percent results from the final exams. Results from fuzzy genetic algorithm give more passed students because of considering those who are in between risk and safe, to safe state that gives students a mental satisfaction. But the lecturers will take attention on them indirectly. So a friendly environment will be created in between lecturers and students. Also expert students will be recruited by reputed companies that make student's future secure. Student's early recruitment to companies brings both the institution and students to prestige.

References

[1] Gaviria, A. (2002). Los quesuben y los quebajan: educación y movilidad social en Colombia. Fedesarrollo, Alfaomega..

[2] Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417.

[3] Mazzoni, E. (2014). The Cliques Participation Index (CPI) as an indicator of creativity in online collaborative groups.Journal of Cognitive Education and Psychology,13(1), 32-52.

[4] Lopez Guarin, C. E., Guzman, E. L., & Gonzalez, F. A. (2015).A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. Tecnologias del Aprendizaje, IEEE RevistaIberoamericana de,10(3), 119-125.

[5] Ogunde, A. O., &Ajibade, D. A. (2014). A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm.Journal of Computer Science and Information Technology, 2(1), 21-46.

[6]Adelman, C. (1999). Answers in the Tool Box.Academic Intensity, Attendance Patterns, and Bachelor's Degree Attainment.

[7] Swamy, M. N., &Hanumanthappa, M. (2012). Predicting academic success from student enrolment data using decision tree technique. Int. J. Appl. Inf. Syst, 4, 1-6.

[8] Ramesh, V., Parkavi, P., &Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. International Journal of Computer Applications, 63(8), 35-39.

[9] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[10] Lakshmi, T. M., Martin, A., &Venkatesan, V. P. (2013). An Analysis of Students Performance Using Genetic Algorithm. Journal of Computer Sciences and Applications, 1(4), 75-79.