Indian Journal of Science and Technology, Vol 9(4), DOI: 10.17485/ijst/2016/v9i4/87032, January 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree

Subitha Sivakumar[1*], Sivakumar Venkataraman[1] and Rajalakshmi Selvaraj[2]

[1]Faculty of Computing, Botho University, Botswana and Department of Information Systems, BIUST, Gaborone, Botswana; subitha.sivakumar@bothouniversity.ac.bw, Sivakumar.venkataraman@bothouniversity.ac.bw

[2]Department of Information Systems, BIUST, Gaborone, Botswana; selvarajr@biust.ac.bw

Abstract

Background/Objectives: Educational Data mining is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. The objective of this work is to identify relevant attributes from socio-demographic, academic and institutional data from undergraduate students at the university located in India and develop an improved decision tree algorithm based on ID3 which can able to predict whether the students continue or drop their studies. **Methods/Statistical Analysis:** The traditional ID3 algorithm is improved by using Renyi entropy, Information gain and Association Function and the model generated by improved decision tree algorithm may be beneficial for university administrators to create guidelines and policies related to raise the enrollment rate in university and to take precautionary and advisory measures and thereby reduce student dropout. It can also used to find the reasons and relevant factors that affect the dropout students. **Findings:** Experimental results proved that improved decision tree algorithm provides better prediction accuracy in educational data than that of traditional classification algorithms in the literature. **Improvements/Applications:** Improved decision algorithm was proposed that enhances the ability to form decision trees and thereby to prove that the classification accuracy of improved decision algorithm on educational dataset is greater.

Keywords: Association Function, Decision Tree, Educational Data Mining, Information Gain, ID3, Renyi Entropy

1. Introduction

Data mining helps to extract the original and the valuable data from the large amount of dataset. Data mining can be implemented in different areas such as Fraud detection, Medical, Education, Banking, Marketing and Telecommunications. Feature selection is a process to pick a group of features as subset that are identically suitable for investigation and for future predication by removing the unrelated or redundant features. The ultimate objective of feature selection process is to increase the predictive accuracy and reducing complexity of learner results In the universities or in academic institutions, it's very difficult to predict the frailer or dropout students in early stage[1,2]. Data assimilations is the main process used to reduce student dropout percentage and to increase the student enrolment percentages in the university. Dropout in residential university is caused by academic, family and personal reasons, campus environment and infrastructure of university and varies depending on the educational structure agreed by the university. Thus, this work aims to effectively formulate education program and institutional infrastructure through which the student's enrollment rate at the university will get increased significantly. The main aim of this paper is to develop an improved decision tree model and to derive a classification rules to predict

* Author for correspondence

whether student will graduate or not using the historic dataset. In this paper, improved decision tree model is used to generate the model. Information such as age, parent's qualification, parent's occupation, academic record, attitude towards university was collected from the students to forecast those students requiring periodical monitoring.

In the Era of data mining, Educational Data Mining (EDM) is considered as a potential study topic. Data mining researchers have well explored and discussed the applicability of data mining in higher education. In[3] performed comprehensive study of educational data mining from 1995 to 2005. In[4] applied k–means clustering to analyze learning behavior of students which will help the tutor to improve the performance of students and reduce the dropout ratio to a significant level. In[5] studied on bored and frustrated student. In[6] studied on the factors that predict failure and non-retention in college courses. Many studies included a wide range of potential predictors, including personality factors, intelligence and aptitude tests, academic achievement, previous college achievements, and demographic data and some of these factors seemed to be stronger than others, however there is no consistent agreement among different studies[7-10].

2. Proposed Work

J. Ross Quinlan proposed the Iterative Dichotomized 3 (ID3 algorithm) in the year 1979 which is used to build the decision tree using information theory. Top down approach with no backtracking is used to build the model in the decision tree algorithm. Information gain is used to determine which attribute will best decide the target data classification. The traditional ID3 algorithm is improved by using Renyi entropy, Information gain and Association Function in this work. This combination is used as a new criterion to construct the decision tree and to predict the dropout of the university students. Initially Renyi entropy is determined using which the Information gain is calculated. This value is kept as the old gain for every attribute. Then using Association Factor, normalized information gain is to be calculated. This is the new information gain. This gain value will be used to construct the decision tree. The framework of proposed decision tree model is shown in Figure 1.

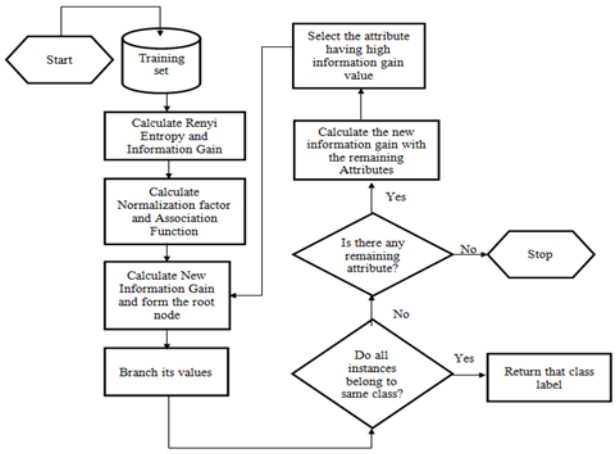

Figure 1. Design of improved decision tree algorithm for educational data mining.

Step 1: The Renyi entropy is used for characterizing the amount of information in a probability distribution. It's generalization of Shannon entropy. Calculate Renyientropy using the formula

$$Entropy = H_{\alpha(X)} = \frac{1}{1-\alpha} * \log \sum \left(p_i \right)^{\alpha}$$

, $\alpha \geq 0$ and $\alpha \neq 1$.

Here X is a discrete random variable with possible outcomes 1, 2…n. α is the order and when it equals to 1 it is Shannon entropy. A completely homogeneous sample has the entropy of 0. Equally divided sample has the entropy of 1.

Step 2: Calculate the Information gain IG of each attribute using the formula:

$$IG = Gain(S,A) = Entropy(S) - \sum_{v \in value(A)} \left(\frac{|s_v|}{|s|} \right) * Entropy(S_v)$$

Step 3: Calculate the Association Function (AF): Suppose A is an attribute of data set D and C is the category attribute of D. the relation degree function between A and C can be expressed as:

$$AF(A) = \frac{1}{n} * \sum_{i=1}^{n} |X_{i1} - X_{i2}|$$

Where Xij indicates that attribute A of D takes the ith value and category attribute C takes the sample number of the jth value, n is the number of values attributes A takes.

Step 4: Calculate the normalization factor: Suppose that there are m attributes and each attribute relation degree function value are AF(1), AF (2) ….AF (m), respectively.

$$V(k)=\frac{AF(k)}{AF(1)+AF(2).....AF(m)}$$

for which 0< k ≤ m.

Step 5: Calculate the new information gain:

New gain(S, A) = Gain(S, A)*V (k).

Now this new Gain can be used as a new splitting criterion for attribute selection to construct decision tree. Decision Trees generated with new information gain are very small and should provide good generalization for small datasets and also avoiding over-fitting.

Step 6: Construct the decision tree with the root node as the attribute which has the maximum information gain value.The splitting criterion tells us which attribute to test at node N by determining the "best" way to separate or partition the tuples in dataset into individual classes.

Step 7: For the root attribute, if all class labels (values) belong to the same class, then it is the terminating condition.

Step 8: If there are different class label, then the tree is further branched with the next node as the attribute which has the next higher information gain value.

Step 9: The above step is repeated until the terminating condition holds.

3. Results and Discussion

A dataset of 240 samples collected randomly through survey at university located at India was used for this empirical study consisting of 32 variables. The dimensionality of dataset is significantly reduced by the use of CFS Feature selector by considering only the correlated features alone with respect to the target classification. CFS Feature selector was applied on original dataset and the selected features are recorded. ID3 and Improved decision tree model was applied on the selected features set and overall accuracy by 10 fold cross validation is recorded. The performance matrix such as accuarcy, precision, recall and F-measure was used to evaluate the performance of both ID3 and Improved Decision Tree algorithms.

True Positive= TP/P;

False Positive=FP/N

Recall = TP/TP+FN;

Precision=TP/TP+FP

F- Measures=2* Recall* Precision/ Recall+Precision

It is found that only 12 features are most relevant to the task of student dropout prediction out of original number of 31 features collected through questionnaire as seen in Table 1. Then the ID3 and improved decision tree algorithm is employed on selected subset of features and record using 10 fold cross validation. Attribute with highest information gain is used as a root node. The

Table 1. Initial Set of features used for the experimentation

11. Residence	27. Like this University
12. Family Type	28. Educational system of University
13. Family Annual Income	29. Infrastructure of university
14. Father's Education	30. Extra-curriculum activities in university
15. Mother's Education	
16. Father's occupation	31. Entertainment in university
17. Mother's occupation	32. Time for self study
18. College Location of student	33. Placement Status
19. Student grade/percentage in High School (10^{th})	34. Participate in extra curriculum activity
20. Student grade/percentage in Senior Secondary (10^{th})	35. Teacher Student relationship
21. Course Admitted	36. Family Problem
22. Admission type	37. Home Sickness
23. Satisfaction with Course	38. Campus Environment
24. Syllabus of Course	39. Change of Goal
25. Parents meet the university expenses	40. Adjustment Problem
26. Family experiences Stress	41. Enrolled in other universities

dropout dataset is classified into two groups Yes and No based on the confusion matrix for Improved Decision Tree was constructed shows accuracy percent 92.50 for ID3 and 97.50 for improved Decision Tree. It indicates that improved decision tree is the best classifier for predicting the student who will dropout or not at the university. The initial set of features used for the experimentation is shown in Table 1. Generated Improved Decision Tree model for student drop out prediction is shown in Figure 2.

Number of features selected by correlation based feature selector, ranking attributes with respect to new information gain, set of rules generated by improved decision tree and confusion matrix of improved decision tree that are shown in Table 2, Table 3, Table 4 and Table 5, respectively.

Table 2. Number of features selected by Correlation Feature Selection

1. Residence
2. Family Type
3. Stream in Senior Secondary
4. Satisfaction with course
5. Family Experience stress
6. Infrastructure of university
7. Participate in extra curriculum activity
8. Family Problem
9. Campus Environment
10. Change of Goal
11. Adjustment Problem
12. Enrolled in other universities.

Table 3. Ranked attribute with respected new information gain

Attribute	Information Gain
Stress	**0.378**
Participate in extra curriculum activity	0.2155
Stream in Senior Secondary	0.1701
Satisfaction with course	0.1672
Enrolled in other institute	0.1359
Change of Goal	0.1288
Campus Environment	0.1245
Family Type	0.1183
Infrastructure of university	0.1064
Adjustment Problem	0.0743
Family Problem	0.0619
Residence	0.0559

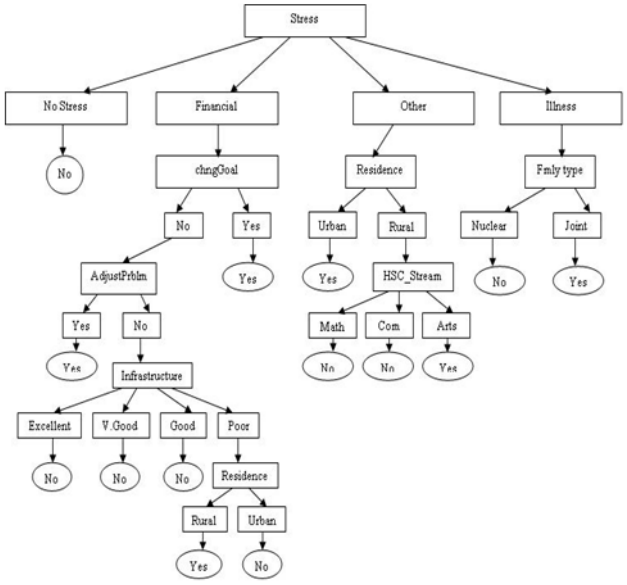

Figure 2. Generated model for prediction of dropout students using Improved Decision Tree.

Table 4. Sample set of rules Generated by Improved Decision Tree

1	IF Attribute_Stress=No THEN Dropout=No
2	IF Attribute_Stress=Financial AND Attribute_Chng-Goal=No AND Attribute_AdjustProblm=NO AND Attribute_Infrastructure=good THEN Dropout=No
3	IF Attribute_Stress=Financial AND Attribute_Chng-Goal=No AND Attribute_AdjustProblm=Yes THEN Dropout=Yes
4	IF Attribute_ Stress=Financial AND Attribute_Chng-Goal=Yes THEN Dropout=Yes
5	IF Attribute_Stress=Other AND Attribute_Resi-dence=Urban THEN Dropout=Yes
6	IF Attribute_Stress=Other AND Attribute_Resi-dence=Rural AND Attribute_HSC_Stream=Math THEN Dropout=No
7	IF Attribute_Stress=illness AND Attribute_Family Type=Nuclear AND THEN Dropout=No
8	IF Attribute_Stress=illness AND Attribute_Family Type=Joint AND THEN Dropout=Yes

The highest dropout reasons were family reasons (10.25%), in institutional factors the highest dropout reason were campus environment (7.58%), low placement rate(4.62%) and in personal problem the highest dropout reason were change of goal (4.92%), adjustment problem

in hostel (2.79%) and home sickness (4.86%). Whereas few students likely to dropout due to illness, home sickness, peer problem, high course fee, adjustment problems and low placement rate etc. The accuracy of conventional and improved decision tree and cause of student dropout are shown in Table 6 and Table 7 respectively.

Table 5. Confusion matrix of Improved Decision Tree

		Predicted class		
		No	Yes	Total
Actual Class	No	195(TP)	2(FN)	197
	Yes	4(FP)	39(TN)	43
	Total	199	41	240

Table 6. Results for the improved decision tree algorithm using 10-fold cross validation as the test option

Performance Metrics	Decision Tree	Improved Decision Tree
Accuracy	92.50 %	97.50 %
TP Rate	0.960	0.990
FP Rate	0.256	0.093
TN Rate	0.744	0.907
FN Rate	0.035	0.010
Precision	0.789	0.914
Recall	0.964	0.990
F-Measure	0.868	0.950
ROC Curve	0.704	0.90

Table 7. Cause of student dropout

Reasons	Percentage
Family Problem	10.25
Home Sickness	4.86
Campus Environment	7.58
Low Placement Rate	4.62
Change of Personal Goal	4.92
Adjustment Problem	2.79

5. Conclusion and Future Work

This paper proposed an improved decision tree algorithm for prediction of dropout student. The objective of this work is to develop an improved decision algorithm that enhances the ability to form decision trees and thereby to prove that the classification accuracy of improved decision algorithm on educational dataset is greater. A new decision tree model is to be constructed by using Renyi entropy for calculating the information gain and the association function will be used which determines the relative degree between the given attribute and class C. Experimental results will prove that improved decision tree algorithm will provide better prediction accuracy on student dropout data than that of traditional classification algorithms.

6. References

1. Ramaswami M, Bhaskaran R. A study on feature selection techniques in educational data mining. Journal of Computing. 2009; 1(1):7-11.
2. Han J, Kamber M. Data mining: Concepts and techniques. San Francisco (CA, USA): Morgan Kaufmann Publishers Academic Press; 2001. p. 550.
3. Romera C, Ventura S. Educational data mining: A Survey from 1995 to 2005. Expert Systems with Applications. 2007; 33(1):125-46.
4. Ayesha S, Musta Fa T, Sattar AR, Khan MI. Data mining model for higher education system. European Journal of Scientific Research. 2010; 43(1):24-9.
5. D'mello SK, Craig SD, Witherspoon A, McDaniel BT, Graesser AC. Automatic detection of learner's affect from conversational cues. User Modeling and User Adapted Interaction. 2008; 18(1):45-80.
6. Romero C, Ventura S, Eapejo PG, Hervas C. Data mining algorithms to classify students. Proceedings of the 1st International Conference on Educational Data Mining; 2008. p. 8-17.
7. Herzog S. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first to second year analysis of new freshmen. Research in Higher Education. 2005; 46(8):883-28.
8. Lassibilille G, Gomez LN. Why do higher education students dropout? Evidence from Spain, Educational Economics. 2008; 16(1):89-105.
9. Touron J. The determination of factors related to academic achievement in the university: implications for the selection and counseling of students. Higher Education. 1983; 12(4):399-10.
10. Malvandi S, Farahi A. Provide a method for increasing the efficiency of learning management systems using educational data mining. Indian Journal of Science and Technology. 2015; 8(28):1-10.