ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2016.0161

ENHANCED PREDICTION OF STUDENT DROPOUTS USING FUZZY INFERENCE SYSTEM AND LOGISTIC REGRESSION

A. Saranya[1] and J. Rajeswari[2]

[1,2]Department of Information Technology, Adhiparasakthi Engineering College, India

E-mail: [1]saranyamit12@gmail.com, [2]rajee.apec@gmail.com

Abstract

Predicting college and school dropouts is a major problem in educational system and has complicated challenge due to data imbalance and multi dimensionality, which can affect the low performance of students. In this paper, we have collected different database from various colleges, among these 500 best real attributes are identified in order to identify the factor that affecting dropout students using neural based classification algorithm and different mining technique are implemented for data processing. We also propose a Dropout Prediction Algorithm (DPA) using fuzzy logic and Logistic Regression based inference system because the weighted average will improve the performance of whole system. We are experimented our proposed work with all other classification systems and documented as the best outcomes. The aggregated data is given to the decision trees for better dropout prediction. The accuracy of overall system 98.6% it shows the proposed work depicts efficient prediction.

Keywords:

Data Mining, Fuzzy Inference System, Logistic Regression, Decision Trees, Student Dropout

1. INTRODUCTION

In recent days student dropout ratio is increasing because of many issues like social hazards, financial problems. Discovering student dropouts and failures is a main social crisis and it has become very significant for educational professionals to solve and satisfies their problem to fulfill their studies. But it is not an easy task, while taking the student datasets there are many factors or characteristics involved to review about dropout and failures, such as cultural, demographics, social, family, or educational background, socioeconomic status, psychological profile, and academic progress [1].

Student endeavors (i.e., the level of school attachment, involvement, and commitment) is related with optimistic academic outcomes. Examiner committees have an important effect on academic outcomes. Peer relationships will produce a set of norms and standards that either promote or destabilize academic achievement. Poor marks, lack of attendance, and detachment [4] from school become particularly aggressive to the completion of high school at this stage and four major high school dropout categories begin to emerge. Using data mining techniques predictions can be made at simple and efficient. There are many techniques involved in dropout prediction such as data preprocessing, attribute selection, classification and rule association. Association rules are useful to find the association between two elements and shows relationship between them.

There are seven different parameters [3] are used to find the relationship between two different factors affecting the school dropout. The parameters discussed in this paper are support, confidence, cosine, added value, lift, correlation, conviction. Educational Data Mining is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. Educational Data Mining (EDM) uses computational approaches to analyze educational data in order to study educational questions [5]. Mining the knowledge from the database or dataset contain [15], [16] several steps: (1) process the data in correct format, (2) apply the data mining techniques to the appropriate field, (3) collect the output predicted data from the field, (4) visualize the data using visualization techniques, etc. The main motivation of this work is to the models (rules and decision trees) generated by the DM algorithms, a system to alert the teacher and their parents about students who are potentially at risk of failing or drop out can be implemented.

The paper is organized as follows: section 2 presents the related work carried out in this field using data mining. Section 3 describes our proposed method for predicting college dropouts. Section 4 describes different experiments carried out (interpretation of our results) and the results obtained. Finally in section 5 summarizes the main conclusions and future research.

2. RELATED WORK

Our country education level is not attaining the growth, due to school/college dropouts. According to the R&D connection, the dropout ratio is 15.9% of Indian people dropout their school. The reason beyond dropouts is low income, less attendance and not interested in subjects. Pattern matching and rule association [2], easily resolve the problem of frequent item identification. Based on the frequently used motives are consider to rectify the problem of dropouts. Student behavior is predicted by using psychological activities defined by “theory of planned behavior”. Educational data mining system throws a report for dropout, it includes attitude of the students, subject norms and perceived control are the main problem for dropout [6]. Decision tree is one of technique, which used to predict the future information. The most important parameters were used in ID3 [8, 9] techniques to filtering the dropout students with their most possible way to dropout. Data mining techniques are particularly used for abnormal behavior and irrelevant pattern identification. Group the students based on their own behavior [10, 11] and interest using the clustering techniques. Most of the reputed institution are organizing “student counseling centre” for reducing the stress of student. Train the students to get back them on their problems, to increase concentration in studies and curriculum.

3. PROPOSED WORK

Student behaviors and related data are collected from the various sources. The dataset is investigated by different kind of students they differ from their nature of behavior, family type, interest, social impacts. Some (student) of the real time data are multidimensional, it can be rebalanced and converted into normal and processed data with the help of data cleansing and sampling techniques. This process referred as data preprocessing. In data preprocessing Data Mining (DM) techniques are applied to prepare the formatted data. We are extracted the features from the student dataset that affects the student performance and education system. Data mining techniques and soft computing techniques are much more effective in the field of learning and prediction. The extracted features are given to the training phase, in this phase the data samples are learned their operation and actions to perform effectively. Once the training method is completed, the models itself learns the state and achieve the accurate result. Based on the trained samples the new data samples are given to the model to test the student performance. The algorithm proposed in this paper is to predict the student Dropouts and failures can be achieved by means of this diagram is shown in Fig.1.

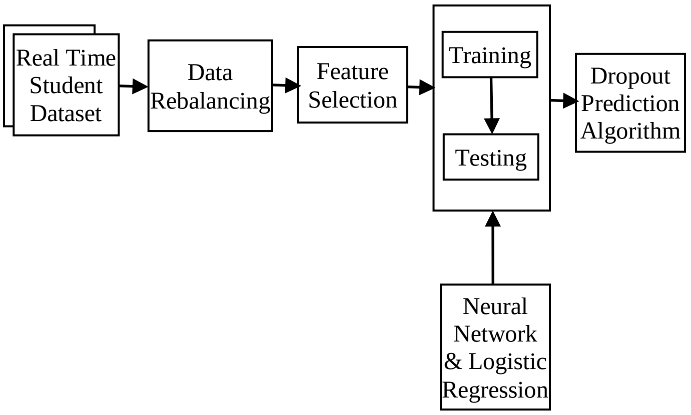
Real Time Student Dataset
Data Rebalancing
Feature Selection
Training
Testing
Dropout Prediction Algorithm
Neural Network & Logistic Regression

Fig.1. Architecture for Dropout Prediction Algorithm

The task of prediction is carrying out the by the following steps:

- Data Collection
- Data Cleansing
- Feature Selection
- Classification of Samples
- Dropout Prediction

3.1 DATA COLLECTION

The student data samples are gathered from various college students. The motive for Students dropout is enormous so predicting their imminent process is not a simple and easy task. The every collected samples are integrated together to frame a dataset. We have gathered 300 data samples from students to lead dropout success scenarios. Most of the real time data sets are multi-dimensional therefore there is a need to rebalance the data. For this process data mining techniques are utilized to evaluate and filter the samples.

3.2 DATA CLEANSING

Cleansing and preprocessing is begins with a collected dataset and to accumulate the significant metrics to accomplish a filtering process. In this work, we have enabled attribute based filtering and instance based filtering. Selecting the excellent attributes lead to achieve the efficient results in filtering. In instance based filtering SMOTE produce an optimal result. There are several preprocessing techniques are available, that can be broadly classified in to two approaches called supervised and unsupervised, attained by Weka data mining tool.

This tool is one of the recognizable in data mining techniques, it includes classification, feature selection, clustering, rule association and prediction algorithms. Some of the neural networking techniques are also available in Weka tool. Weka accepts the data samples in the form of .arff and .csv file formats. The input data samples are collectively varies with their attribute because each and every student has different philosophy to discontinue the course. Groping and balancing the data set is more important to obtain the better result.

3.3 FEATURE SELECTION

All the dropout information is lies into reduced 23 features, in that we have to select the most appropriate attributes to predict the dropouts. Best attribute selection algorithm lead to achieve great effect in dropout scenario. This can be broadly categorized into filter and wrapper; filter includes two process of selection and evaluation of attribute independently. Wrapper uses the learning algorithm to determine the desirability of a sample. Weka provides several attribute selection algorithms such as, CfsSubsetEval, ChiSquaredAttributeEval, Consistency-SubsetEval, FilteredAttributeEval, OneRAttributeEval, FilteredSubsetEval, GiniIndexBased, PrincipalComponents, GainRatioAttributeEval and InfoGain-AttributeEval. We have deeply analyses the student log to predict the dropouts. There are lots of factors affecting student performance in college and schools. In that we finalize the best 23 attributes to conclude the dropout problem shows in Table.1.

Table.1. List of Attributes

Source	Attributes
General Survey	Daily attendance, friends list, Study method, parental survey, study habits, marital status, religion, type of personality, physical disability, Participation in extra-curricular activities, residential, satisfaction level, bad habits, family income level, campus environment, family type, enrolled in other institution, history of arrear transport method, score in mathematics, score in English, parent occupation, sports, social impacts
Reduced Features (proposed data collection)	Family type, parent occupation, participation in extra-curricular activities, residential, satisfaction level, enrolled in other institution, change of goal, rules, placement, motivation of study, campus environment, HSC marks, Score in mathematics, score in English, history of arrear, social impacts, bad habits, stress, health problem, nature of character, FIRs, medical Surgery

The best attributes are selected based on the attribute selection algorithms. Each algorithm of attribute selection operates on unique search methods. Significantly, 10 attribute selection algorithms have been applied and results are shown in Table.2.

Table.2. Best Attribute Selected

Sl. No.	Algorithm Name	Selected Attributes
1	CfsubsetEval	History of arrear, enrolled in other institution, change of goal, rules, stress
2	ChiSquaredAttributeEval	Family type, parent occupation, participation in extra-curricular activities, residential, satisfaction level, enrolled in other institution, change of goal, rules, placement, motivation of study, campus environment, history of arrear, social impacts, bad habits, stress, health problem, nature of character, FIRs, medical surgery
3	FilteredAtributeEval	Family type, parent occupation, participation in extra-curricular activities, enrolled in other institution, change of goal, rules, placement, motivation of study, campus environment, history of arrear, social impacts, bad habits, stress, health problem, nature of character, FIRs, medical surgery
4	FilteredSubsetEval	Enrolled in other institution change of goal, history of arrear, social impacts
5	GainRatioAttributeEval	Enrolled in other institution, change of goal, rules, placement, motivation of study, campus environment, history of arrear, social impacts, bad habits, stress, health problem, nature of character, FIRs, medical surgery
6	InfoGainAttributeEval	Enrolled in other institution, change of goal, rules placement, motivation of study, campus environment, history of arrear, social impact, bad habits, stress, health problem, nature of character, FIRs, medical surgery
7	OneRAttributeEval	Family type, parent occupation, participation in extra-curricular activities, residential, satisfaction level, enrolled in other institution, change of goal, rules, placement, motivation of study, campus environment, history of arrear, social impacts, bad habits, stress, health problem, nature of character, FIRs, medical surgery
8	PrincipalComponent	Family type, parent occupation, participation in extra-curricular activities, residential, satisfaction level, enrolled in other institution, change of goal, rules, placement, motivation of study, campus environment, history of arrear, social impacts, bad habits, stress, health problem, nature of character, FIRs, medical surgery
9	SymmetricalUncertAttributeEval	Enrolled in other institution, change of goal, rules, placement, motivation of study, campus environment, history of arrear, social impacts, bad habits, stress, health problem, nature of character, FIRs, medical surgery
10	WrapperSubsetEval	Change of goal, rules, placement, motivation of study history of arrear, social impacts, stress, health problem, nature of character, FIRs, medical surgery

.

The best attribute is selected based on the frequent occurrence in all attribute selection algorithm. In reduced 23 attributes, we have filtered 12 best attribute to predict the student dropout in college to attain high education growth in institutions and society. The best selected attributes are Parent occupation, Participation in non-collegiate activities, Satisfaction level; Enrolled in other institution, change of goal, rules, placement, campus environment, history of arrear, Bad habits, stress, and nature of character is depicted in Table.3.

Table.3. Best Attribute Selection

Attribute Name	Frequency of Occurrence
Parent occupation	10
History of arrear	10
Satisfaction level	9

Change of goal	8
Participation in extra-curricular activities	8
Stress	8
Enrolled in other institution	7
Health problem	7
Placement	7
Social impacts	7
Bad habits	6
Nature of character	6

3.4 CLASSIFICATION OF SAMPLES

The selected attributes are forwarded to the next phase of classification. Here train the data samples based on the selected feature and test the remaining data samples. Weka includes more than 76 classification and regression algorithm. We have used decision tree and logistic regression to predict the dropouts, because logistic regression is best for prediction function and formulate the accurate decision also help to choose a best path. Based on the extracted features decision tree classify the data items.

3.5 DROPOUT PREDICTION

We generate the rules for predicting dropouts using the fuzzy inference system. Each and every attributes are classified based on the circumstances. There are 12 important attributes covered in this phase to predict the student's performance, in that parent occupation is classified based on the income (high, average, low), history of arrear is classified based on the number of arrears (below 5, 5 to 10 and above 10), placement is categorized into three levels (like good, average, poor) and other attributes has classified in Boolean conditions (like yes or no). All fuzzy rules are combined to frame the Dropout Prediction Algorithm (DPA) helps to efficiently predict the student dropouts in college. This system will improve the college academic performance and students pass percentage. The prediction algorithm is described in below Table.4.

Table.4. Rules for Predicting Dropouts

If(parent occupation = average) and (change of goal = yes) and (history of arrear = above 5) then dropout = yes

If(parent occupation = high)) and (history of arrear = below5) and (stress = yes) and (health problem = yes) then dropout = yes

if(enrolled in other = no) and (placement facility = yes) and (bad habits = no) and (satisfaction = poor) and (social impacts = low) then dropouts = no

if(social impacts = high) and (satisfaction = poor) and (change of goal = yes) and (participation in other activities = yes) then dropout = yes

if(participation in other activities = yes) and (history of arrear = below 5) and (stress = no) and (placement = yes) then dropouts = no

4. EXPERIMENT RESULTS AND ANALYSIS

Our results improve the classification accuracy to determine the dropout predictions. This section describes the student's performance to improve the academic status of individual. We executed 10 classification algorithms to obtain the better results about the student performance. The experiments results differentiated from both the selected attributes and overall attributes. We consider the decision tree and neural network techniques to classify the samples based on the student behavior. In progress of classification data balancing and preprocessing techniques are applied ahead. The accuracy parameters of data classification True positive, false positive precision and recall are measured by Weka data mining tool is displayed in Table.5.

Table.5. Classification Results based on All Attributes

Algorithm	TP Rate	FP Rate	Precision	Recall
NaïveBayes	0.892	0.545	0.846	0.892
MultiLayer Perceptron	0.838	0.636	0.816	0.838
RBF Network	0.892	0.455	0.868	0.892
Logistic Regression	0.973	0.545	0.857	0.973
SMO	0.865	0.909	0.762	0.865
AdaBoost	0.973	0.545	0.857	0.973
Decision Stump	0.973	0.455	0.878	0.973
NBtree	0.946	1	0.761	0.946
JRip	0.946	0.545	0.854	0.946
Ridor	0.892	0.818	0.786	0.892

Based on the correctly classified records, the True Positive (TP) and False Positive (FP) rates are calculated by following Eq.(1) and Eq.(2).

$$\text{True Positive Rate (TPR)} = TP/(TP + FN) \tag{1}$$

$$\text{False Positive Rate (FPR)} = FP/(FP + TN) \tag{2}$$

In experimented data, how many samples are correctly classified based on the attributes and constraints are taken as true positive. The Remaining and imperfectly classified records are considers as false records are taken as false positive is shown in Table.6.

Table.6. Classification Results based on Selected Attribute

Algorithm	TP Rate	FP Rate	Precision	Recall
NaïveBayes	0.892	0.455	0.868	0.892
Multilayer Perceptron	0.784	0.727	0.784	0.784
RBFNetwork	0.919	0.455	0.872	0.919
LogisticRegression	0.973	0.455	0.878	0.973
SMO	0.919	0.727	0.81	0.919
AdaBoost	0.973	0.636	0.837	0.973
DecisionStump	0.973	0.455	0.878	0.973

NBtree	0.865	0.909	0.762	0.865
JRip	0.919	0.545	0.85	0.919
Ridor	0.892	0.364	0.892	0.892

Weka allows any classification algorithm to be made cost sensitive by using the Meta classification algorithm CostSensitive Classifier and setting its base classifier as the desired. We executed all the classification algorithms using tenfold cross-validation, considering different costs of classification and introducing a coast matrix algorithm. The rule set is associated with group of attributes; any changes in the attribute selection majorly affect the overall performance. Classification accuracy depends on the true positive rate, false positive rate, F- score, recall and precision is shown in Fig.2 and Fig.3.

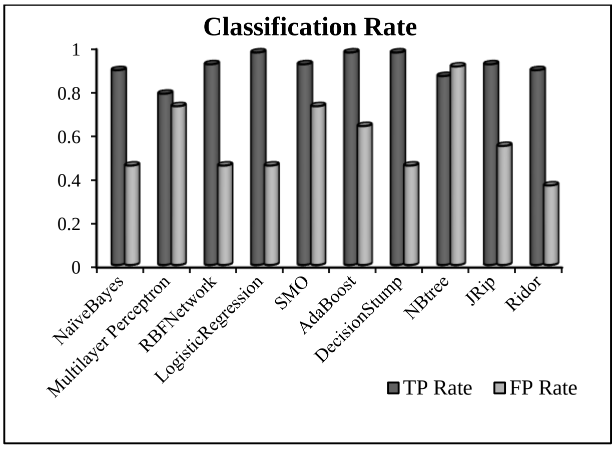

Fig.2. Results of Classification Accuracy

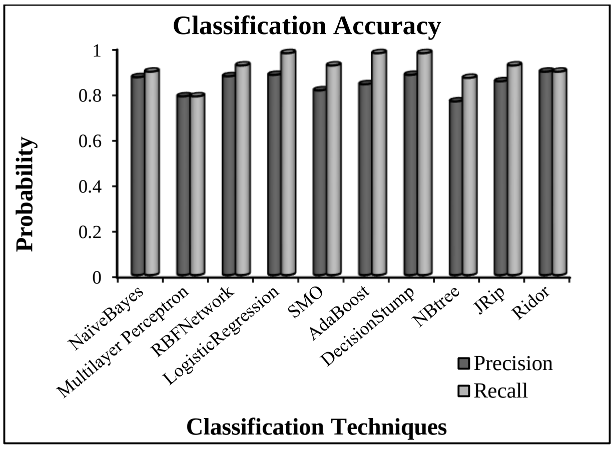

Fig.3. Classification Accuracy Measures

The data samples are tested in all classification techniques and their results has been visualized and compared with other algorithms. In Fig.4, shows the result of logistic regression and neural network techniques achieve the best outcome. Decision tree always prone to produce higher level of prediction (ID3, Decision Stump etc.).

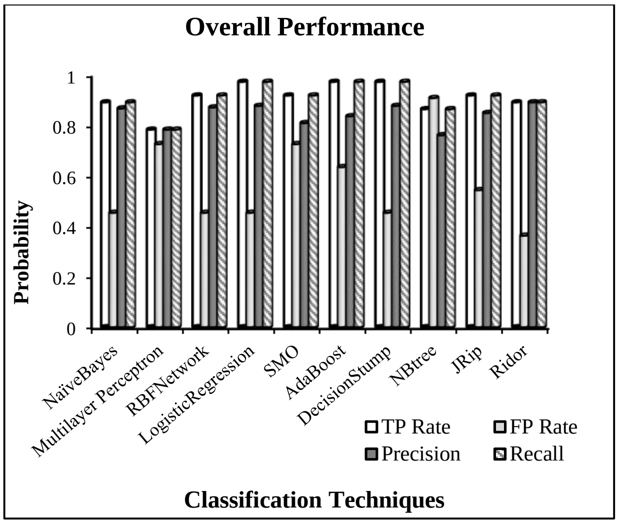

Fig.4. Comparison with other Techniques

5. CONCLUSION AND FUTURE WORK

Predicting the student dropouts in college is not a simple task. It consists of various direct and indirect processes to produce the solution to this problem. The major risk involved in this work is data balancing and efficient prediction. We achieve both in our proposed work with help of efficient attribute selection algorithm and data preprocessing techniques. Gathering the data is difficult task because student behavior and nature depends on the situation and environment. Generating the IF-THEN rules is used to predict the dropouts; the rules can be structured based on the dropout historical information. The proposed system learns the rules from the trained samples and applied into testing samples. We propose that once students were found at risk, mentor will be assigned and they counsel with both academic support and guidance for motivating and trying to prevent student failure. Our experimented result shows the outcome by achieving of highly predictive rate using the logistic regression and inference system. Our future enhancement is to design the automatic student performance evaluation system to minimize the dropouts and improves the academic performance.

REFERENCES

[1] Patricia A. Aloise-Young and Ernest L. Chavez, “Not All School Dropouts are the same: Ethnic Differences in the Relation between Reason for Leaving School and Adolescent Substance Use”, *Psychology in the Schools*, Vol. 39, No. 5, pp. 539-547, 2002.

[2] Carlos Márquez-Vera, Cristóbal Romero Morales and Sebastián Ventura Soto, “Predicting School Failure and Dropout by using Data Mining Techniques”, *IEEE Journal of Latin-American Learning Technologies*, Vol. 8, No. 1, pp. 7-14, 2013.

[3] Francisco Araque, Concepción Roldán and Alberto Salguero, “Factors Influencing University Dropout Rates”, *Computers and Education*, Vol. 53, No. 3, pp. 563-574, 2009.

[4] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art”, *IEEE Transactions on*

Systems, Man, and Cybernetics, Part C, Applications and Reviews, Vol. 40, No. 6, pp. 601-618, 2010.

[5] Rajni Jindal and Malaya Dutta Borah, "A Survey on Educational Data Mining and Research Trends", *International Journal of Database Management Systems*, Vol. 5, No. 3, pp. 53-73, 2013.

[6] Bharat Inder Fozdar, Lalita S. Kumar and S. Kannan, "A Survey of a Study on the Reasons Responsible for Student Dropout from the Bachelor of Science Programme at Indira Gandhi National Open University", *International Review of Research in Open and Distance Learning*, Vol. 7, No. 3. pp. 1-15, 2006.

[7] Shreenath Acharya and N. Madhu, "Discovery of Students' Academic Patterns using Data Mining Techniques", *International Journal on Computer Science and Engineering*, Vol. 4 No. 6 pp. 1054-1062, 2012.

[8] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, "Predicting Students' Performance Using ID3 and C4.5 Classification Algorithms", *International Journal of Data Mining & Knowledge Management Process*, Vol. 3, No. 5, pp. 39-52, 2013.

[9] Mohammed M. Abu Tair and Alaa M. El-Halees, "Mining Educational Data to Improve Student's Performance", *International Journal of Information and Communication Technology Research*, Vol. 2, No. 2, pp. 140-146, 2012.

[10] Sotiris B. Kotsiantis, "Educational Data Mining: A Case Study for Predicting Dropout-Prone Students", *International Journal of Knowledge Engineering Soft Data Paradigms*, Vol. 1, No. 2, pp. 101-111, 2009.

[11] L. Fourtin, D. Marcotte, P. Potvin, E. Roger and J. Joly, "Typology of Students at Risk of Dropping Out of School: Description by Personal, Family and School Factors", *European Journal of Psychology of Education*, Vol. XXI, No. 4, pp. 363-383, 2006.

[12] L.G. Moseley and D.M. Mead, "Predicting Who Will Drop Out of Nursing Courses: A Machine Learning Exercise", *Nurse Education Today*, Vol. 28, No. 4, pp. 469-475, 2008.

[13] M.A. Hall and G. Holmes, "Benchmarking Attribute Selection Techniques for Data Mining," Working paper No. 00/10, Department of Computer Science, University of Waikato, 2002.

[14] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, "*Classification and Regression Trees (Wadsworth Statistics/Probability)*", New York: Chapman & Hall, 1984.

[15] Yoav Freund and Llew Mason, "The Alternating Decision Tree Algorithm," *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 124-133, 1999.