Predicting Dropout Students Using Data-Mining Techniques

Omkar S. Patil[1] & Parag M. Dhere[2]
Vidya Pratishthan's College of Engineering, Baramti, Pune, India
**E-mail:*omkarp21@gmail.com & paragchintoo007@gmail.com

Abstract:

Now-a-days the amount of data stored in educational database increasing rapidly. These databases contain hidden information of students' performance. Educational data mining is used to study the data available in the educational field and bring out the hidden knowledge from it. Classification methods like decision trees can be applied on the educational data for predicting the students' performance. This prediction will help to identify the dropout students and in order to provide them with both academic support and guidance for motivating and trying to prevent student failure. Weka is an open source data mining software is used to explore the influence of factors on predicting students' academic performance. Dataset is applied to different classifiers of Weka : J48,RandomForest, RepTree and BFTree of Decision Tree and JRip rule.

Keywords:

Data mining; Decision Tree and Educational data mining

Introduction

Data Mining is process to discover or extract useful information from a large dataset or database, in a way that those useful information can't be extracted using a simple query. Data mining concepts and methods can be applied in various fields like marketing, medicine, real estate, customer relationship management, engineering, web mining etc. Educational data mining is a new emerging technique of data mining that can be applied on the data related to the field of education. Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments.

Examination plays a vital role in any student's life. The marks obtained by the student in the examination decide his future. Therefore it becomes essential to predict whether the student will pass or fail in the examination. If the prediction says that a student tends to fail in the examination prior to the examination then extra efforts can be taken to improve his studies and help him to pass the examination.

The stages to predicting the academic failure of students belongs to the process of Knowledge Discovery and Data Mining and they are :

a) Gathering all available information of students with predictive variables,

b) Identification of different factors, which affects a student's learning behavior and performance during academic career,

c) Construction of a prediction model using classification data mining techniques on the basis of identified predictive variables and

d) The obtained models are analyzed to detect student failure.

Predicting successful and unsuccessful students at an early stage of the academic year it help not only to concentrate more on the bright students but also to apply more efforts in developing programs for the failure ones in order to improve their progress while attempting to avoid student dropouts.

Data Mining and Knowledge Discovery

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. The sequences of steps identified in extracting knowledge from data are: shown in Figure 1.

The more common Techniques in current data mining practice include the following.

1) Classification: classifies a data item into some of several predefined categorical classes.

2) Regression: maps a data item to a real valued prediction variable.

3)Clustering: Clustering is maximization of similarity and minimization of dissimilarity between categorical classes.

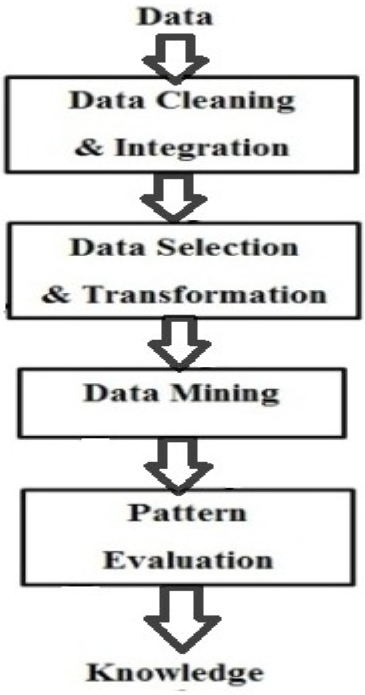

Fig 1. The Steps of Extracting Knowledge from Data

4) Rule generation: extracts different classification rules from the data.

5) Discovering association rules: describes association relationship among different attributes.

6) Summarization: provides a compact description for a subset of data.

7) Dependency modeling: describes relating dependencies among variables.4) Rule generation: extracts different classification rules from the data.

5) Discovering association rules: describes association relationship among different attributes.

6) Summarization: provides a compact description for a subset of data.

7) Dependency modeling: describes relating dependencies among variables.

Method

The method proposed for predicting the academic failure of students belongs to the process of Knowledge Discovery and Data Mining (see Fig. 1).

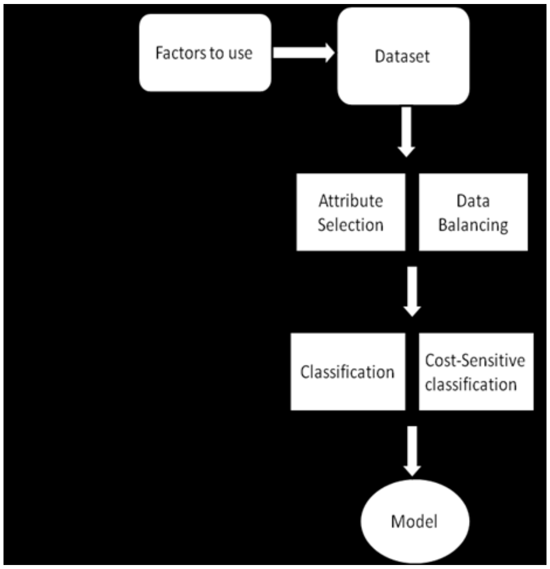

Figure 1. Method proposed for the prediction of student failure.

The main stages of the method are

1) Data gathering
2) Pre-processing
3) Data mining
4) Interpretation

Data Gathering

This stage collecting information on students and find which factors can most affect the students' performance and collect these information from the different sources of data available. Finally, the information of students is being used to prepare dataset.

All the information of students is collecting from three different sources :

Table 1. Variables Used And Information Sources

Data Pre –Processing

Pre-processing stage which included cleaning, integration, discretization and variable transformation tasks. This tasks can be improve the quality of dataset and reliability of available information, which directly affects the results obtained. This stage also include techniques which are the selection of attributes and the re-balancing of data.These techniques are applied on student database in order to solve the problems of high dimensionality and imbalanced data.

In this stage those students who are not having complete information in dataset were eliminated. This also include fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies. The Preprocessing can create new attribute such as age of student with help of its DOB. Furthermore, the continuous variables were transformed into discrete variables. For example, the numerical values of the scores obtained by students in each subject were changed to categorical values.

The Attribute Selection techniques can be used for selecting best attribute. In this identify most affecting factors which are directly effect on output. The problem of imbalanced data by students in each subject were changed to categorical values.

Source	Type of information	Attributes	Number of attributes
Specific survey	Personal, social, family and school factors	Classroom/group, number of students in group, attendance during morning /evening sessions, number of friends, number of hours spent studying daily, methods of study used, place normally used for studying, having one's own space for studying, resources for study, study habits, studying in group, parental encouragement for study, marital status, having any children, religion, having administrative sanctions, type of degree selected, influence on the degree selected, type of personality, having a physical disability, suffering a critical illness, regular consumption of alcohol, smoking habits, family income level, having a scholarship, having a job, living with one' s parents, mother' s level of education, father' s level of education, number of brothers/ sisters, position as the oldest/ middle/ youngest child, living in a large city, number of years living in the city, transport method used to go school, distance to the school, level of attendance during classes, level or boredom during classes, interest in the subjects, level of difficulty of the subjects, level of motivation, taking notes in class, methods of teaching, too heavy a demand of homework, quality of school infrastructure, having a personal tutor, level of teacher' s concern for the welfare of each student.	45
General survey	Socioeconomic factors and previous marks	Age, sex, previous school, type of school , type of secondary school, Grade Point Average (GPA) in secondary school, mother' s occupation, father' s occupation, number of family members, limitations for doing exercises, frequency of exercises, time spent doing exercises, score obtained in Logic, score in Math, score in Verbal Reasoning, score in Spanish, score in Biology, score in Physics, score in Chemistry, score in History, score in Geography, score in Civics, score in Ethics, score in English, and average score in the exam.	25
Scores	Current marks	Score in Math 1, score in Physics 1, score in Social Science 1, score in Humanities 1, score in Writing and Reading 1, score in English 1 and score in Computer 1.	7

The Attribute Selection techniques can be used for selecting best attribute. In this identify most affecting factors which are directly effect on output. The problem of imbalanced data classification occurs when the number of instances in one class is much smaller than the number of instances in another class or other classes. This imbalanced problem can solve in preprocessing stage.

Data Mining

Data mining stage is used for obtaining the prediction models of students' academic status. Data mining algorithms are applied on student dataset to predict student failure based on rules and decision trees. A decision tree is a set of conditions organized in a hierarchical structure. The decision tree is make decision at each node and following the path of satisfied conditions from the root of the tree until a leaf is reached. In Rule induction algorithms, in which obtain rules from a description of each class. A decision tree can be directly transformed into a set of IF-THEN rules which are obtaining from rule induction algorithm.

This stage also include cost sensitive classification approach is also used in order to solve the imbalanced data problem. At last, different algorithms have been executed, evaluated and compared in order to determine which one obtains the best accuracy results and best result model can used for analyzing failure student.

Table 2. Confusion matrix

Pred./Act.	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

In order to evaluate the performance of classification algorithms, normally the confusion matrix is used. This matrix contains information about actual and predicted classifications(see Table 3).

– Accuracy (Acc) is the overall accuracy rate or classification accuracy and is calculated as

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

– True Positive rate (TP rate), also called sensitivity (Se) or recall, is the proportion of actual positives which are predicted to be positive and is calculated as

$$TP_{rate} = \frac{TP}{TP + FN}$$

– True Negative rate (TN rate), or specificity (Sp), is the proportion of actual negatives which are predicted to be negative and is calculated as

$$TN_{rate} = \frac{TN}{TN + FP}$$

– Geometric Mean (GM) indicates the balance between classification performances in the majority and minority classes; that is, GM is a measure of the central tendency used with imbalanced datasets and is calculated as

$$GM = \sqrt{TP_{rate} \cdot TN_{rate}}$$

Interpretation Of Results

The obtain model from data mining stage which is used in interpretation stage to analyze failure student. In Interpretation stage algorithms can discovers rules. These rules show factors and relationships that student to pass or fail. The attributes which are associated to fail they are mostly concerning marks in subject.

Weka Implementation

Weka is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. From the student dataset with 75 attribute of 450 students, student.arff file was created. This file was loaded into WEKA explorer. The classify panel enables the user to apply classification algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself. There are 16 decision tree algorithms like ID3, J48, ADT etc. implemented in WEKA. The algorithm used for classification is ID3, C4.5 and CART. Under the "Test options", the 10-fold cross-validation is selected as our evaluation approach. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. The model is generated in the form of decision tree. These predictive models provide ways to predict whether a student will fail or not.

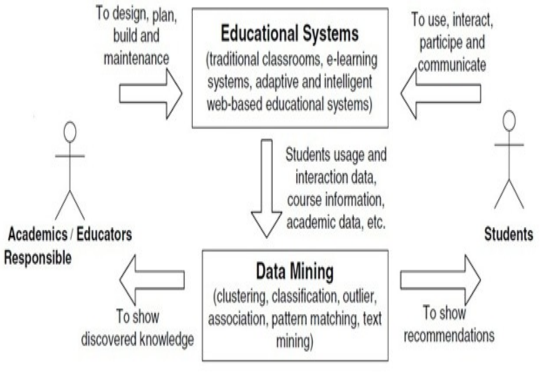
To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Students usage and interaction data, course information, academic data, etc.
Academics / Educators Responsible
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Fig 3. Data Mining in Education systems.

Here as shown in Fig. 3, educators and academics responsible are in charge of designing, planning, building and maintaining the educational systems. Students use and interact with them. Starting from all the available information about courses, students, usage and interaction, different data mining techniques can be applied in order to discover useful knowledge that helps to improve the e-learning process. The discovered knowledge can be used not only by providers (educators) but also by own users (students). So, the application of data mining in educational systems can be oriented to different actors with each particular point of view.

Conclusion

We have shown that classification algorithms cab be used successfully in order to predict a student's academic performance and, in particular, to model the difference between Fail and Pass students.

We have shown two different ways to address the problem of imbalanced data classification by rebalancing the data and considering different classification costs. In fact, rebalancing of the data has been able to improve the classification results obtained in TN rate, Accuracy, and Geometric Mean.

Concerning the specific factor or attributes related with student failure, there are some specific values that appear most frequently in the classification models obtained. For example, the values of scores/grades that appear most frequently in the obtained classification rules are the values “Poor”, “Very Poor”, and “Not Presented” in the subjects of Physics 1, Humanities 1, Math 1 and English 1. Other factors frequently associated with failing are being over 15 years of age, having more than one sibling, attending evening classroom/group, having a low level of motivation to study, to live in a big city (with more than 20 thousand inhabitants), and students which consider Math as a difficult subject. It is also striking that the failing grades for a subject like Humanities, that a majority of students usually pass, appear in the obtained models.

REFERENCES

1)Carlos Mrquez-Vera, Cristbal Romero Morales, and Sebastin Ventura Soto. "Predicting School Failure and Dropout by Using Data Mining Techniques,"IEEE ,Latin-American Learning Tech.,vol.8, no. 1, Feb 2013.

2) A. Parker, "A study of variables that predict dropout from distance education," Int. J. Educ. Technol., vol. 1, no. 2, pp. 1–11, 1999.

3) Y. Freund and L. Mason, "The alternating decision tree algorithm," in Proc. 16th Int. Conf. Mach. Learn., 1999, pp. 124–133.

4) M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for data mining," Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, Tech. Rep. 00/10, Jul. 2002.