International Journal of Ethics in Engineering & Management Education

Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 10, October 2014)

Data Mining Techniques for School Failure and Dropout System

[1]K.Swathi
M.Tech (CSE)
Aurora Technical Research and Institute, Hyderabad

[2]D.Sujatha
Associate Professor (Ph.d)
Aurora Technical Research and Institute, Hyderabad

Abstract: Data mining techniques are applied to predict college failure and bum of the student. This is method uses real data on middle-school students for prediction of failure and drop out. It implements white-box classification strategies, like induction rules and decision trees or call trees. Call tree could be a call support tool that uses tree-like graph or a model of call and their possible consequences. A call tree is a flowchart-like structure in which internal node represents a "test" on an attribute. Attribute is the real information of students that is collected from college in middle or pedagogy, each branch represents the outcome of the test and each leaf node represents a class label. The paths from root to leaf represent classification rules and it consists of three kinds of nodes which incorporates call node, likelihood node and finish node. It is specifically used in call analysis. Using this technique to boost their correctness for predicting which students might fail or dropout (idler) by first, using all the accessible attributes next, choosing the most effective attributes. Attribute choice is done by using WEKA tool.

Keywords: dataset, classification, clustering.

1. INTRODUCTION

Past years have shown a growing interest and concern in several countries about the problem of college failure and the determination of its main contributing factors [1]. The great deal of analysis [2] has been done on characteristic the factors that affect the low performance of students (school failure and dropout) at totally different instructional levels (primary, secondary and higher) using the large quantity of data that current computers can store in databases. To identify and find useful information hidden in massive databases is a trouble task [3]. A very promising solution to reach this goal is that the use of information discovery in databases techniques or data mining in education, referred to as instructional data processing, EDM [4]. This new area analysis focuses on the event of methods to better understand students and therefore the settings in which they learn [5]. In fact, there are good samples of the way to apply EDM techniques to make models that predict dropping out and student failure specifically [6]. These works have shown promising results with respect to those social science, economic, or instructional characteristics that may be additional relevant in the prediction of low educational performance [7]. It is important to note that almost of the analysis on the application of EDM is to resolve the issues of student failure and drop-outs has been applied primarily to the particular case of higher education [8] and specifically to on-line or distance education[9]. However very little information concerning specific analysis on elementary and secondary education has been found, and what has been found uses only statistical methods, not DM techniques [10].

EXISTING SYSTEM:

Starting from the previous models (rules and decision trees) generated by the DM algorithms, a system is used to alert the teacher and their parents about students who are potentially at risk of failing or drop out can be implemented.

LIMITATIONS:

- 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.
- It does not include complete academics of any student as it covers only till higher secondary school.

PROPOSED SYSTEM:

We propose that Data mining techniques are applied to Engineering colleges and once students were found at risk, they would be assigned to a tutor in order to provide them with both academic support and guidance for motivating and trying to prevent student failure.

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

LIMITATIONS:

- Where the whole control comes under students rather than parents or teaching faculty.
- If student does not maintain attendance and credits properly it will lead to serious problem and there are high chances to student failure and drop outs.

ADVANTAGES:

- Data mining is a broad process that consists of several stages and includes many techniques.
- This knowledge discovery process comprises the steps of pre-processing, the application of DM techniques and the evaluation and reading of the results.
- DM is aimed at working with very large amounts of data (millions and billions).
- The statistics does not usually work well in large databases with high dimensionality.

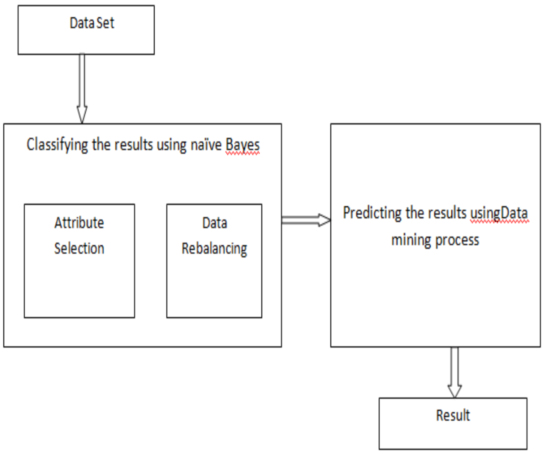

Data Set
Classifying the results using naïve Bayes
Attribute Selection
Data Rebalancing
Predicting the results usingData mining process
Result

Fig:1 System Architecture

RELATED WORK:

Prediction of Higher Education Admissibility using Classification Algorithms

This paper presents the results from data mining research, performed at one of the famous and prestigious Government Arts and Science Colleges in Tamil Nadu, with the main goal to predict the higher education admissibility of women students. In this research, real data about 690 under-graduate students from Government arts college (W) was taken. The research is focused on the development of data mining models for predicting the students likely to go for higher studies, based on their personal, precollege and graduate performance characteristics.

Predicting School Failure Using Data Mining

This paper proposes to apply data mining techniques to predict school failure. They have used real data about 670 middle-school students from Zacatecas, México. Several experiments have been carried out in an attempt to improve accuracy in the prediction of final student performance and, specifically, of which students might fail. In the first experiment the best 15 attributes has been selected.

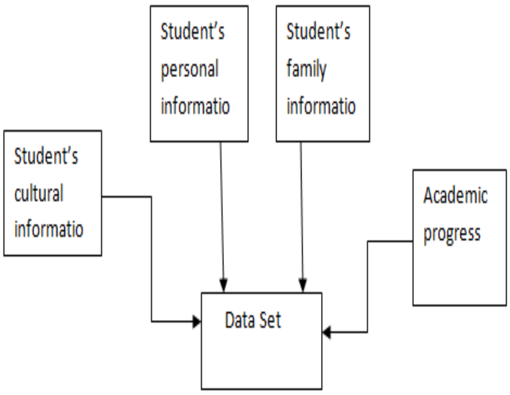
Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out

In this paper, they consider three issues of factors affecting students drop out rate. These factors are conditions related to the students before admission, factors related to the students during the study periods in the university, and all factors including the target value to be predict for factors analysis. They use tree-based classification algorithm, J48 or C4.5, and Naïve Bayes to analyze the data.

SYSTEM IMPLEMENTATION:

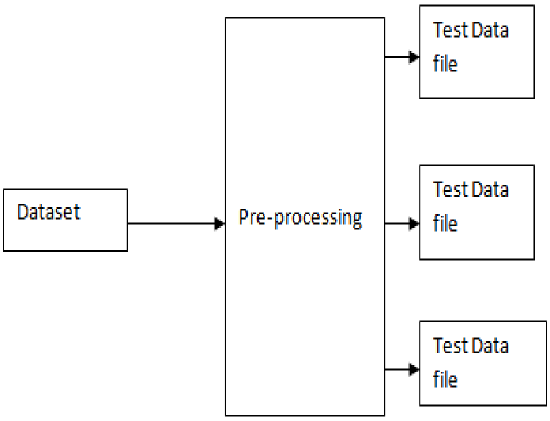
1. DATA GATHERING

The process of data gathering is that involves in collecting all available information about students. The set of factor should be identified that can affect student's performance and collected from different available data sources. The collected characteristics or risk factors that can influence to students failure or dropped out. Risk factors contain the information about student's cultural, social, educational background, socioeconomic status, psychological profile and academic progress. In which most of the students are aged between 15 and 16 and this is the years with the highest rate of failure. Finally the survey is to obtain personal and family information to identify important risk factors of all students and school services provides the score obtained by the students in all subjects of course. All those information are integrated into single dataset.


Student's personal informatio
Student's family informatio
Student's cultural informatio
Academic progress
Data Set

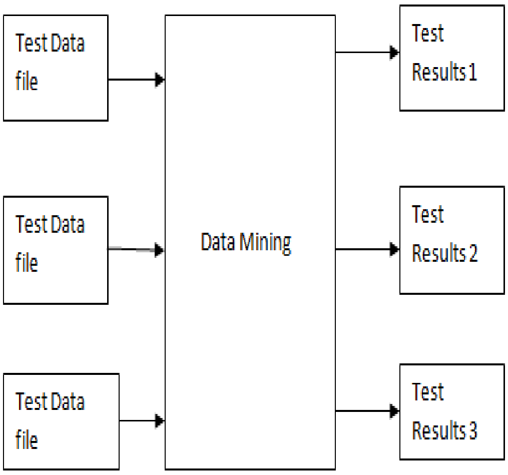
2. PRE-PROCESSING

In this stage dataset is prepared for applying data mining technique. Before applying data mining technique, pre-processing methods like cleaning, variable transformation and data partitioning and other techniques for attribute selection must be applied. Here new attribute of age is created using date of birth of each students. The continuous variables are transformed into discrete variable that is scores obtained by each student is changed into categorical values (i.e) Excellent score between 9.5 and 10,Very good the score between 8.5 and 9.4.all information's are integrated in single dataset that is stored in .arff format of Weka tool. Finally entire dataset is divided randomly into 10 pairs of training and test data files. After pre-processing we have attributes or variables for each student. Each test file will contain best attributes and rebalanced.



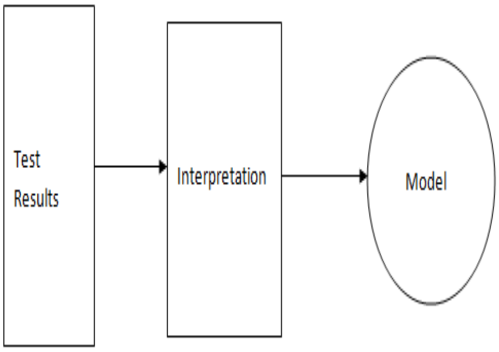
3. DATA MINING

In this stage Data mining techniques are applied. Here the data mining technique used for classification. The classification is based on best attribute selection from data set. In which the naive bays algorithm is implemented for classification of data. Traditionally the Weka Software tool is used for data mining. It contains vareity of data mining algorithms. Weka implements decision tree, it is a set of condition organized in hierarchical structure. Decision tree algorithms are like J48, C4.5, Random Tree etc. Here the classification algorithms were executed using cross- validation and all available information. Finally the result with the test file of classification is shown.



4. INTERPRETATION

In which, the obtained results are analyzed to predict student failure or drop out. To achieve this, previous test results are taken for comparison. At this stage classification rules are applied for predicting relevant factors and relationships that lead to student pass or fail. There are attributes that indicate that student who failed are older than 15 year and some of the attributes are show marks of poor, not presented and regular students. Finally the risk factors are analyzed from previous results of classification algorithms.



5. CONCLUSION

As we've seen, predicting student failure in a class is a troublesome task, not solely as a result of it's a multifactor drawback (in that there are plenty of private, family, social, and economic factors which will be influential). To resolve these issues, we've shown the utilization of various DM algorithms and approaches for predicting student failure. We've applied totally different classification approaches for predicting the educational status or final student performance at the end of the course. Moreover we've shown that some approaches like choosing the most effective attributes, cost-sensitive classification, and information equalization can even be very helpful for improving accuracy

6. FUTUREWORK

As future work, we like to extend the work as
1) To develop our own algorithmic program for classification/prediction based on grammar, using descriptive genetic programming which will be compared against classic algorithms.
2) To predict the student failure as soon as possible. The sooner the higher, so as to find students in danger and in time before it's too late.
3) To propose actions for serving to students known among the danger cluster. Then, to ascertain the speed of the days it's potential to stop the fail or dropout of that student antecedent detected.

REFERENCES

[1]. L. A. Alvares Aldaco, “Comportamiento de la deserción y reprobación en el colegio de bachilleres del estado de baja california: Caso plantel ensenada,” in Proc. 10th Congr. Nat. Invest. Educ., 2009, pp. 1–12.

[2]. F. Araque, C. Roldán, and A. Salguero, “Factors influencing university drop out rates,” Comput. Educ., vol. 53, no. 3, pp. 563–574, 2009.

[3]. M. N. Quadril and N. V. Kalyankar, “Drop out feature of student data for academic performance using decision tree techniques,” Global J. Comput. Sci. Technol., vol. 10, pp. 2–5, Feb. 2010.

[4]. C. Romero and S. Ventura, “Educational data mining: A survey from 1995 to 2005,” Expert Syst. Appl., vol. 33, no. 1, pp. 135–146, 2007.

[5]. C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 6, pp. 601–618, Nov. 2010.

[6]. S. Kotsiantis, K. Patriarcheas, and M. Xenos, “A combinational incremental ensemble of classifiers as a technique for predicting students’ performance in distance education,” Knowl. Based Syst., vol. 23, no. 6, pp. 529–535, Aug. 2010.

[7]. J. Más-Estellés, R. Alcover-Arándiga, A. Dapena-Janeiro, A. Valderruten-Vidal, R. Satorre-Cuerda, F. Llopis-Pascual, T. Rojo-Guillén, R. Mayo-Gual, M. Bermejo-Llopis, J. Gutiérrez-Serrano, J. García-Almiñana, E. Tovar-Caro, and E. Menasalvas-Ruiz, “Rendimiento académico de los estudios de informática en algunos centros españoles,” in Proc. 15th Jornadas Enseñanza Univ. Inf., Barcelona, Rep. Conf., 2009, pp. 5–12.

[8]. S. Kotsiantis, “Educational data mining: A case study for predicting dropout—prone students,” Int. J. Know. Eng. Soft Data Paradigms, vol. 1, no. 2, pp. 101–111, 2009.

[9]. I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, “Dropout prediction in e-learning courses through the combination of machine learning techniques,” Comput. Educ., vol. 53, no. 3, pp. 950–965, 2009.

[10]. A. Parker, “A study of variables that predict dropout from distance education,” Int. J. Educ. Technol., vol. 1, no. 2, pp. 1–11, 1999.

[11]. T. Aluja, “La minería de datos, entre la estadística y la inteligencia artificial,” Quaderns d’Estadística Invest. Operat., vol. 25, no. 3, pp. 479–498, 2001.

[12]. M. M. Hernández, “Causas del fracaso escolar,” in Proc. 13th Congr. Soc. Española Med. Adolescente, 2002, pp. 1–5.

[13]. E. Espíndola and A. León, “La deserción escolar en américa latina: Un Tema prioritario para la agenda regional,” Revista Iberoamer. Educ., vol. 1, no. 30, pp. 39–62, 2002.

[14]. I. H.Witten and F. Eibe, Data Mining, Practical Machine Learning Tools and Techniques, 2nd ed. San Mateo, CA, USA: Morgan Kaufman, 2005.

[15]. 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.