Analysis of Classification Algorithms for Prediction of Student Failure Using EDM

Miss. Trupti Diwan
Department of Information Technology
MIT College of Engineering
Pune, India
Truptid969@gmail.com

Prof. Bharati Dixit
Department of Information Technology
MIT College of Engineering
Pune, India
dixit.bharati@gmail.com

Abstract*—Data mining has a unlimited scope of uses running from business to medication to engineering. Educational data mining (EDM) is a rising interdisciplinary research zones that arrangement with the advancement of routines to investigate data starting in an educational context. In this paper, we gather data of different students from MIT College of Engineering in Pune University. This data is preprocessed and prediction of students failure/succeed will be done in order to improve an accuracy. This paper proposes to apply data mining procedures to foresee college disappointment and dropout. In this paper, a genetic programming calculation and LADTree algorithm are proposed for prediction of students. In the end accuracy of algorithms is analyzed. ADTree algorithm showed correctly classified instances 89.3939% & relative absolute error 76.893.

Keywords- College failure, Classification, Educational data mining (EDM), Genetic algorithm, LADTree algorithm, Prediction,.

I. INTRODUCTION

The data in any given educational organization is developing quickly. There is a need to change this data into useful data and knowledge; hence we make use of data mining. Educational data mining is the area of science where different methods are being developed for making discoveries inside data. This data is obtained from an educational foundation. These methods provide an understanding into a student's behavioral patterns and the environment in which they learn [10].

The abundant deal of research [12] has made on recognizing the elements that affect the less performance of students (college failure and dropout) at different educational levels (essential, secondary furthermore higher) utilizing the large measure of the data that large amount of data is store in databases. All these data are a "gold mine" of valuable data about students. Identify and find useful data hidden in large databases is a troublesome errand [13]. A very making a guarantee to solution to achieve this objective is the use of knowledge discovery in databases methods or data mining in education, called Educational Data Mining, EDM [14].

EDM is used for developing approaches to discover the unique types of data in educational surroundings and, utilizing these approaches, for better understanding of students as well as the settings in which they learn [1]. EDM has occurred as a examination area in modern years for researchers universally throughout the world from various as well as associated research areas.

The EDM process converts crude data originating from educational systems into useful data that could potentially have a great effect on educational study and training. This process does not change much from other use areas of DM, like business, genetics, medicine, etc, because it takes after the same steps as the general DM process[9] preprocessing, DM, and post-processing. However, it is imperative to note that in this paper, the term DM is utilized in a broad sense than the unique/customary DM definition, i.e., we are going to describe not just EDM studies that use commonplace DM procedures, such as classification, sequential mining, association-rule mining, clustering, etc., moreover describe other approaches, for example, regression, correlation, visualization, etc., which are not measured to be DM in a strict sense.

Detecting student failure at college is a real social problem what's more it has become very imperative for educational professionals to better understand why such a variety of young people neglect to complete their college studies. In any case, this is a troublesome problem to resolve due to the

large measure of danger elements or characteristics of students that can impact college failure, for example, demographics, social, social, family, or educational foundation, socioeconomic status, mental profile, and academic progress [1]. Truth be told, this problem is otherwise called the "one thousand elements problem" [13]. In the earlier decades, a excessive deal of research has done on identifying the fundamental components that affect the low performance of students, for example, failure furthermore dropping out at different educational levels, for example, essential, secondary, and higher [2]. One of the most influential theoretical explanations of this problem and its causes also cures is the way examination model of Tinto [9]. Model suggest that the student's social and academic integration into the educational organization is the significant determinant of accomplishment and classifies some key influences on integration for example, the student's family foundation, personal characteristics, previous educating, former academic performance, also interactions between student and the staff. Currently, there has been consent that recognition and prevention of student failure at college and early intervention make significantly more sense than remediation [3].

This study proposes to predict student failure at college in secondary education by utilizing DM. Actually; we need to detect the variables that most influence student failure in adolescent students by utilizing classification techniques. Different DM approaches have likewise been used to attempt to increase the exactness of the classification model and to resolve the problems of high dimensionality and imbalanced data.

II. RELATED WORK

This paper [1] proposed to apply data mining procedures to calculate school disappointment and dropout. We utilize genuine information on 670 center school students from Zacatecas, México, and apply white-box classification strategies, for example, decision trees and induction rules [2].

This report [2] is overview of their state of the craftsmanship with respect to EDM and surveys the absolute most relevant perform of this type to date. Each study has been categorized, not just by the sort of data and DM methods applied, additionally and more vitally, by the type of instructional occupation that they resolve [2].

In this paper, [3] author studied about the utilization of data mining in training for student profiling and collecting. The author make utilization of Apriori algorithm for student profiling which is the famous approaches for mining associations i.e. finding co-relations amongst set of items. Another calculation utilized, for gathering students is K-means clustering which allots a set of observations into subsets [3].

In this paper [9], the author presents the utilization of data mining method, especially classification method, to help secondary school students in choice of UG programs. The paper additionally exhibits the study on educational structure in Thailand, and foundation of data mining ideas and procedures [9].

III. IMPLEMENTATION DETAILS

A. System Overview

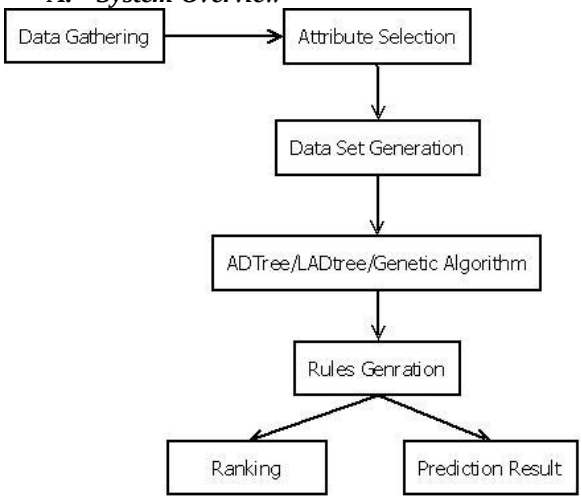

Figure1: System Architecture

System architecture is shown in figure1.In this paper we are gathering data of students from MIT College of Engineering in Pune University. For removing unwanted data we need to preprocess gathered data. Based on the rule student dropout and failure is being predicted. We used AdTree Algorithm to predict student failure. In proposed work we will use LADTree and Genetic algorithm (ICRM) to predict student failure. Accuracy of these classification algorithms is compared in order to check best performance. Ranking algorithm is used based on the marks obtained by the particular student. The student ranking will be based on average percentage calculated and by sorting average percentage in descending order.

B. Mathematical Model for Proposed Work

Let, System S is represented as: S = {A, B, C, D}

Data Gathering:

Let, A is a set of students dataset A= {a1, a2,…}

Where,

a1, a2… are the students data gathered.

Attribute Selection:

Let B is a set attribute selection B = {f1, f2, f2, f3, f4, f5}

Where,

f1, f2, f3, f4, f5 are the selected attributes.

Classifier and Genetic Algorithm:

Let, C is a classifier C = {e1; e2}

Where,

e1, e2 are the classification process.

Ranking:

Let, D is a ranking process,

D = {d1, d2, d3,….dn}

Where,

d1, d2,… are number marks obtained by students.

C. Algorithm

ADTree Algorithm:

Input: Set of instances from a dataset.

Step 1: calculate weight of each instance

$w_i=\frac{1}{m}$ for all i

Step 2: Calculate value of a

$a=\frac{1}{2\ln\frac{w+(true)}{w-(true)}}$

Where,

W+ (true) returns sum of wt of +ve

Step 3: Declare precondition p {true} and condition c={true}

Step 4: For i=1 to T

Where, T is number of iterations

P belongs to P and c belongs to C

Calculate values that min 1=2 $\sqrt{(w+(p\hat{}c)w-(p\hat{}c)}+\sqrt{(w+(p^{\sim}c)w-(p\hat{}\sim c)}$ + w(~p)

Step 5: p+= p^c + p^~c

Step 6: $a1=\ln\frac{1}{2}\ \frac{W+(P\hat{}C)+1}{W_{-P\hat{}C}+1}$

$a2=\ln\frac{1}{2}\ \frac{W+(P\hat{}\sim C)+1}{W_{-P\hat{}\sim C}+1}$

Where,

a1 and a2 are weights.

Step 7: Calculate new rules R with precondition p and c and weight a1 and a2.

R_j=rules

Step 8: Calculation of weight

$w_i=w_i e^{-yi}Rj(xi)$

End for

Step 9: Return set of R_j

LADTree Algorithm:

To obtain a tree we require error criterion and hence LAD is used, LAD stands for Least Absolute Deviation. LAD is built on the operations of logical expression. LAD is basically a binary classifier and gives distinction between –ive and +ive examples. The basic assumption of LAD model is that a binary purpose coated by some positive patterns, however not coated by any negative pattern is positive, and equally, a binary purpose coated by some negative patterns, however not coated by positive pattern is negative. Logical Analysis of Data (LAD) tree is that the classifier for binary target variable supported learning logical expression which will distinguish between positive and negative samples in an exceedingly data set. The construction of LAD model for a given knowledge set generally involves the generation of huge set patterns and therefore the choice of a set of them that satisfies the on top of assumption specified every pattern within the model satisfies sure needs in terms of prevalence and homogeneity. LADTree produces a multi-category LADTree. It has the capability to have more than two class inputs. It performs additive logistic regression using the Logistics Strategy.

Grammar rules generated by ICRM Algorithm:

Evolutionary calculations are an ideal model based on the Darwin evolution process, where each individual codifies an answer what's more evolves into a improved individual by means of genetic operators (transformation and crossover). Genetic Programming (GP) is an evolutionary calculation based methodology used to discover computer programs that perform a user-defined and. It is a specialization of genetic calculations where every single is a computer program. Therefore, it's a machine learning process utilized to optimize a populace of computer projects as indicated by a fitness landscape determined by a program's capacity to perform a given computational undertaking.

(*S*) → (*cmp*) | (*cmp*) *AND* (*S*)

(*cmp*) → (*op_ca) (variable)*(*value*)

(*op_cat*)→ = | ≠

(*variable*)→*Any valid attribute in dataset*

(*value*)→*Any valid value*

The fitness capacity evaluates the nature of the represented arrangements. We have used a mix of two measures that are having common position in classification, called the sensitivity (Se) and the specificity (Sp). After that the fitness value is considered as the result of the sensitivity and the specificity to maximize the precision. This capacity achieves fine regardless of may be the data is imbalanced.

Fitness = Se · Sp

At last, there are several approaches to assemble a rule base. In this paper we proposed three various versions of the ICRM (Interpretable Classification Rule Mining) calculation to get rule bases that are both accurate and useful to the user in search for

student failure data. The initial method (ICRM v1) states that there is only one rule per class. The second approach (ICRM v2) permits multiple rules for each class. The third techniques (ICRM v3) spreads the second approach yet focuses on the student's failure problem.

D. Experimental Setup

The system is built using Java framework (version jdk 6) on Windows platform. The Netbeans (version 6.9) is used as a development tool.

IV. RESULTS AND DISCUSSION

A. Results

The following Table I. shows the performance measure parameters and there values in percentage for ADTree algorithm.

TABLE I PERFORMANCE MEASURES TABLE

Performance Measures	ADTree
Correctly Classified Instances	89.3939%
Incorrectly Classified Instances	10.6061%
Relative absolute error	76.8938
Root relative squared error	73.9185

The following Figure 2: shows performance measure parameters and there values for ADTree Algorithm.

1. **Correctly Classified Instances**: It classifies the correct instances.
2. **Incorrectly Classified Instances**: It classifies the incorrect instances.
3. **Relative Absolute Error:** The relative absolute error is very similar to the relative squared error in the sense that it is also relative to a simple predictor, which is just the average of the actual values.

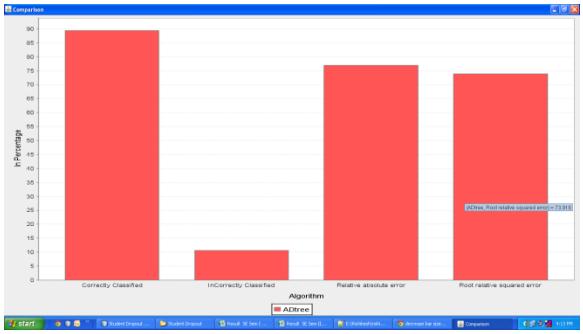

Figure.2: Graph of Performance Measure

$$Relative\ Absolute\ Error = \frac{\sum_{j=1}^{n}|P_{(ij)} - T_j|}{\sum_{j=1}^{n}|T_j - \bar{T}|}$$

where $P_{(ij)}$ is the value predicted by the individual program i for sample case j (out of n sample cases); T_j is the target value for sample case j; and $\bar{T}$ is given by the formula:

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$$

4. **Root relative squared error**: The root relative squared error is relative to what it would have been if a simple predictor had been used.

$$Root\ Relative\ Squared\ Error = \sqrt{\frac{\sum_{j=1}^{n}(P_{(ij)} - T_j)^2}{\sum_{j=1}^{n}(T_j - \bar{T})^2}}$$

where $P_{(ij)}$ is the value predicted by the individual program i for sample case j (out of n sample cases); T_j is the target value for sample case j; and $\bar{T}$ is given by the formula:

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$$

V. CONCLUSION

To predict student failure at college can be a troublesome assignment not just because it is a multi-factor problem (in which there are a ton of personal, family, social, and economic variables that can be influential) additionally because the available data are typically imbalanced (the greater part of the students go to the next course). DM calculations & approaches for predicting student failure are used to solve these problems. Several experiments are performed on ADTree algorithm. Real data from MIT College of Engineering in Pune University is used. Performance of ADTree is measured in terms of correctly classified instances, incorrectly classified instances, and relative absolute error or in TP Rate, TN Rate, GM, and Accuracy. Different classification approaches will be applies to predict academic status or last student performance toward the end of the course. The classification algorithms are proposed for acquiring accurate and comprehensible classification rules.

VI. REFERENCES

[1] Cristobal Romero, Sebastian 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, NOVEMBER 2010.

[2] Carlos Márquez-Vera, Cristabal Romero Morales, and Sebastian Ventura Soto-Predicting School Failure and Dropout by Using Data Mining Techniques" IEEE JOURNAL OF LATIN-AMERICAN LEARNING TECHNOLOGIES, VOL. 8, NO. 1, FEBRUARY 2013.

[3] Suhem Parack, Zain Zahid, Fatima Merchant, "Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns".

[4] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan, "Mining Social Media Data for Understanding Students Learning Experiences" IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 7, NO. 3, JULY-SEPTEMBER 2014.

[5] Oktariani Nurul Pratiwi, "Predicting Student Placement Class using Data Mining" 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE).

[6] Tripti Mishra, Dr. Dharminder Kumar, Dr. Sangeeta Gupta, "Mining Student's Data for Performance Prediction" 2014 Fourth International Conference on Advanced Computing & Communication Technologies.

[7] Yohannes Kurniawan, Erwin Halim., "Use Data Warehouse and Data Mining to Predict Student Academic Performance in Schools: A Case Study (Perspective Application and Benefits)" 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE).

[8] Brijesh Kumar Baradwaj, Saurabh Pal, "Mining Educational Data to Analyze Students" Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[9] Waraporn Jirapanthong, "Classification Model for Selecting Undergraduate Programs" 2009 Eighth International Symposium on Natural Language Processing.

[10] Aloise-Young PA, Chavez EL (2002) Not all school dropout are the same: Ethnic differences in the relation between reason for leaving school and adolescent substance use. Psychol Sch 39(5):539–547

[11] K. Wisaeng , "A Comparison of Different Classification Techniques for Bank Direct Marketing", International Journal of Soft Computing and Engineering (IJSCE), Volume-3, Issue-4, September 2013, pp-116-119 .

[12] Araque F, Roldan C, Salguero A (2009) Factors influencing university dropout rates. Comput Educ 53:563–574.

[13] Hernández MM (2002) Causas del Fracaso Escolar. In: XIII Congreso de la Sociedad Espanola de Medicina

[14] Klösgen W, Zytkow JM (2002) Handbook of data mining and knowledge discovery. Oxford University Press, London