ISSN 1870-4069

Evaluating Predictive Techniques in Educational Data Mining: An Unbalanced Data Set Case of Study

Lourdes Sánchez-Guerrero, Josué Figueroa-González,
Beatriz González-Beltrán, Silvia González-Brambila

Universidad Autónoma Metropolitana, System Department,
CDMX, Mexico

{lsg,jfgo,bgonzalez,sgb}@azc.uam.mx

Abstract. This work presents an evaluation of the predictive techniques decision trees using CART algorithm, Naïve Bayes Classifier, Gradient Boosting Machine and Support Vector Machine for predicting whether a student will successfully complete a programming course or not. Factors considered for prediction were university-entrance and personal criteria like entrance age, gender, scholarship, high school GPA, mark in admission exam and other related with student's performance in a prerequisite introductory programming course. The predicted variable takes two values, 'Approved' or 'Not Approved', and the data record contains an unbalanced portion of the class 'Approved'. For the analysis were considered two data sets, unbalanced and balanced. Evaluation of algorithms was performed considering the concepts of accuracy and ROC area. Results show that accuracy is bigger for the unbalanced data set, but its ROC area was very poor. Using the balanced data set, results were more reliable because accuracy and ROC area are closer. Best results were obtained with Naïve Bayes and Support Vector Machine algorithms. The most important factor in the prediction was whether a student had a scholarship or not.

Keywords. Educational data mining, predictive techniques evaluation, unbalanced data set, students' performance prediction, ROC curve evaluation.

1 Introduction

The increment in the use of technology has allowed gathering and storing a lot of data in many fields; as time passed, was observed that these data could be analyzed using several techniques for obtaining information. This process of analyzing large sets of data for finding patterns that lead to knowledge is known as Data Mining (DM) [8], and nowadays is applied in a lot of fields. In an educational environment, also is gathered a lot of information from many aspects of students, like their academic performance, personal characteristics or even social and affective ones. All these data can be exploited using DM techniques,

for finding valuable information about them and the aspects that could affect their academic performance. Applying DM techniques for analyzing educative information received the name of Educational Data Mining (EDM) [15].

The main goal of educational institutions is ensuring that their students obtain a good preparation and finish their studies. Use of EDM has grown in the last decades, and right now it is applied for analyzing several aspects related with education [11], being one of the most studied the prediction of students' achievement at different levels, from a single work or exam, to the complete path of their higher education. Students' performance has been analyzed mainly with DM regression and classification techniques [2].

One common problem in EDM is that data sets are not big enough compared with other fields. Small data sets are common with educational data, and it is possible that these data sets are unbalanced, having a big difference between the amount of data of one class and another. This generates a problem when creating predictive models because obtained results could not be totally reliable, this is known as the *unbalanced data set problem.*

For dealing with unbalanced data sets, there exists some methods known as 'Sampling Methods' that balance the distribution of classes modifying the size of the original data. This work presents an analysis and comparison of the accuracy of predictive techniques: decision trees using CART algorithm [4], Naïve Bayes classifier (NB) [12], Support Vector Machines (SVM) [9] and Gradient Boosting Machine (GBM) [13] applied to an small educational data set with an unbalanced distribution of classes.

In particular, the goal of this paper is comparing the accuracy of different predictive algorithms for determining the most suitable for predicting whether a student will successfully complete a programming course or not, taking into account personal characteristics and academic performance in a previous programming course, and considering an unbalanced data set and a balanced one.

The structure of this paper is: Section 2 presents the related work concerning predicting student performance. Section 3 shows the experiments and the obtained results. Section 4 presents the analyzed results, and finally, Section 5 contains conclusions and future work.

2 Related Work

Predicting students' performance is the most studied topic in EDM because there are a lot of different approaches and factors to consider.

In [5], the authors compared the effectiveness to identify early fail of students in an introductory programming course using Neural Networks, Decision Tree, Support Vector Machine and Naïve Bayes classifiers, using data from one distance course (considering 22 variables) and the other from on-campus (considering 16 variables).

They conclude that the SVM technique outperformed the other ones by predicting with 92% and 83% of effectiveness, the failures of students that have performed at least 50% of the courses by distance education and on-campus,

respectively. They conclude that the information is useful for teachers to take decisions, but is necessary make more experiments that permit the generalization of their results.

In [14], CART, J48, C4.5 and C5.0 decision trees algorithms and artificial neural network are applied for predicting study branch enrollment decision, future grades and satisfaction level of students in an Indian University. Processed data varied depending the classification. For enrollment decision (i.e. school and engineering), were considered national and state ranks of student and personal and social factors. For future grades, were considered academic criteria like: midterm grades, participation, and understanding of practice and theory. Finally, for satisfaction level were considered obtained grades, selected branch, social and family factors and student expectations about its studies. Results shows that enrollment decision prediction was the best, with a level of confidence over 98% for every technique. Other decisions obtain 60% and 67% of accuracy. The algorithm with the highest accuracy was C5.0.

In [6] was predicted students' performance using linear regression and matrix factorization approaches. They predicted a) students' next-term course grades, and b) within-class assessment performance. In particular, they investigated four methods: the course-specific regression (CSpR), the personalized linear multi-regression (PLMR) methods, the standard matrix factorization (MF) and the MF method based on factorization machines (FM) to predict the grade that a student will achieve in a specific course. They shown that PLMR and MF can predict next-term grades with lower error rates than traditional methods. PLMR were also useful for predicting grades on assessments within a traditional class or online course.

In [16], was developed a method for predicting student performance based in the cumulative Grade Point Average in a certain area. Main contribution of the paper is focusing on three main problems: differences of students at the moment of taking courses, importance of different courses at the moment of the prediction and incorporation of student progress in the prediction. Proposed method, which consists in a structure of base predictors that later were assembled in a cascade of predictors which incorporated the progress of student performance in the prediction, obtained better performance than traditional methods which gives different courses the same importance and do not include the evolution of students performance in later predictions.

In [7], is presented a predictive analysis for determining students performance in a public school in Brazil considering two aspects: first, personal and before entrance school criteria, and then, other set considering academic factors in a certain moment of the scholar period. Data sets were taken separately from years 2015 and 2016 and then combined. Classification technique was Gradient Boosting Machine and the process was performed with CRISP methodology. Results showed that academic factors like grades and absences are the most important, but also some personal factors, specially the neighborhood and the age of students have an important role over their performance at the end of the scholar year.

3 Experiments and Results

This section presents the analysis performed to the data set and the obtained results. We considered the steps of CRISP-DM methodology [1]: business understanding, data understanding, data preparation and modeling for obtaining the accuracy of considered algorithms. The evaluation stage is presented in Section 4 and the deployment step is out the scope of the present work.

3.1 Business Understanding

The first step involved the understanding and defining the goal of the analysis. The proposed goal for the study case was predicting whether a student would successfully complete a programming course or not, considering personal, entrance and academic factors. Also, it is presented the importance of different criteria at the moment of predicting grades and measured the effect of processing balanced and unbalanced data set.

It was used a study case related with predicting whether a student will approve the programming course "Object Oriented Programming" (OOP) or not, based in its personal characteristics, university entrance factors and performance in the prerequisite introductory course "Structured Programming" (SP) in the Mexican Universidad Autónoma Metropolitana Azcapotzalco (UAM-A). Prerequisite or seriation relationship means that for taking OOP, a student must have approved SP. This prerequisite relationship in the study programs is based on the idea of linking a set of courses so, one or more give the appropriate knowledge to students for having a good performance in the next ones.

The main goal of this work was measuring the accuracy of different algorithms applied to balanced and unbalanced data sets, for determining the algorithms more suitable, according to the characteristics of data.

3.2 Data Understanding

For accomplishing the goals, the student data set is about academic, personal and entrance data. Data for the analysis were obtained from two sources, the General File of Students (GFS) that contains personal and entrance information of the student like: age of entrance, score in the entrance exam, high school GPA, gender and scholar period of entrance.

Another source was the historical record of marks file (called *kardex* at UAM-A), which contains: student identifier, course identifier, obtained mark, scholar period where the grade was obtained and the way a course was approved, at UAM-A exist two: taking the course (called Global evaluation - GLO.) or through an extraordinary exam (called Recuperation Exam - REC.); these data were considered as part of the performance in SP.

It was only considered the grade obtained the first time a student took OOP (not mattering whether it was approved or not).

Grades at UAM-A are assigned with letters: MB (Very Good), B (Good), S (Sufficient) and NA (Not Approved). Some criteria are expressed in scholar periods, which at UAM-A represents three months.

Were only considered Computing Engineering students that are currently studying and for which OOP is mandatory. The total of students processed for analyzing the relationship between SP and OOP was 258.

3.3 Data Preparation

Considered criteria for predicting whether a student will approve OOP or not were:

- *Entrance age*. Age of the student at the moment of being accepted at UAM-A.
- *Gender of the student*. Male or Female.
- *High school GPA*. Average of the student in the former scholar level.
- *Score in entrance exam*. Score in the admission exam, maximum score is 1000 points.
- *Scholarship*. Whether the student has a scholarship or not.
- *Time before SP*. Time elapsed before the student took SP, in the first opportunity. Notice that SP it is supposed to be taken 2 or 3 scholar periods after a student enters the university.
- *Grade in SP*. Grade obtained at the moment of approving SP.
- *Tries for approving SP*. Number of attemps that took the student approving SP.
- *Time invested in approving SP*. Time invested by the student in approving SP, notice that this time and the number of tries it is not necessarily the same.
- *Time elapsed before taking OOP*. Number of scholar periods elapsed since the student approved SP and took OOP, in its first opportunity, approving it or not.

The predicted variable was the *performance of student in OOP in its first opportunity*; that is, whether the student approved OOP or not OOP, in its first opportunity.

Possible values and acronym for each criteria used in predicting if a student approved or not OOP are presented in Table 1.

Kardex did not contain neither the time invested in approving SP nor the number of tries. First value was obtained considering the entrance scholar period and the one in which the student approved SP. Number of tries needed for approving SP was obtained considering the amount of not approved marks obtained taking the SP before approving it. Time elapsed before taking OOP after approving SP was calculated considering the scholar period in which SP was approved and the one in which the OOP was taken. According the study program of Computing Engineering, OOP should be taken the next scholar period after approving SP.

Table 1. Criteria, acronyms and possible values used in predicting grade "Object Oriented Programming".

Criteria	Acronym	Values
Entrance age	AGE	17 to 37
Gender	GEN	Male or Female
High school GPA	MLA	7.7 to 10.0
Score in entrance exam	EXM	548 to 909
Scholarship	SCH	Yes or No
Progress level	PRO	0 to 15
Mark in SP	MSP	S, B or MB
Number of tries needed for approving SP	TRSP	1 to 5
Time invested in approving SP	TMSP	1 to 10
Time elapsed before taking OOP	TIMEP	0 to 14

Initially, grades to be predicted had four possible values: MB, B, S and NA; however after some tests, accuracy of different algorithms was very low, about 30%. For improving this accuracy, the approved marks (MB, B and S) were grouped in a single criteria. This also allows obtaining a binary classification where the predicted variable could have one or two possible values: AP, when the student successfully complete the course, and NAP when the student did not approve the course.

Distribution of grades in OOP was: 181 students approved (AP) in their first try and 77 did not approve (NAP). This represents a 45.5% of NAP. Difference it is not very large, but considering the reduced amount of data could be treated as an unbalanced data set.

For balancing the data, it was applied the Random Over Sampling Examples (ROSE) [10] package of R software.

3.4 Modeling and Algorithm Accuracy

The modeling stage was focused on measuring accuracy for each algorithm.

As already mentioned, the data set were analyzed with the predictive techniques Decision Tree using CART algorithm, NB, GBM and SVM.

Considering that the amount of data was small, we tested different percentages for training and testing models. Best results for predicting OOP grade were obtained with 90% for training and 10% for testing. A cross-validation process with 10-fold was applied for all predictive algorithms.

Accuracy was measured using the concept of Receiver Operating Characteristics (ROC) [3] which represents the rate of Positive values classified as positive, True Positve (TP), against the Negative values classified as positive, False Positive (FP), rate. Accuracy of the model is represented by the area under a curve, as more area, more will be the efficiency of the predictive model.

Accuracy of unbalanced data set. First, it was measured the accuracy of algorithms considering the original data set. The distribution of AP and NAP classes for training and testing sets is shown in Table 2.

The distribution of classes were: 68.9% of AP and 31.1% NAP for the training set and 80.7%for AP and 19.3% for the testing set.

Table 2. Class distribution in unbalanced data set.

	Approved	Not Approved
Training set	160	72
Testing set	21	5

The best accuracy and the ROC area for each algorithm is presented in Table 3.

Table 3. Accuracy and ROC area for unbalanced data set.

Algorithm	Accuracy	ROC area
CART	65.38%	0.557
SVM	80.76%	0.5
NB	76.9%	0.629
GBM	80.76%	0.5

ROC curves for each algorithm are presented in Figure 1.

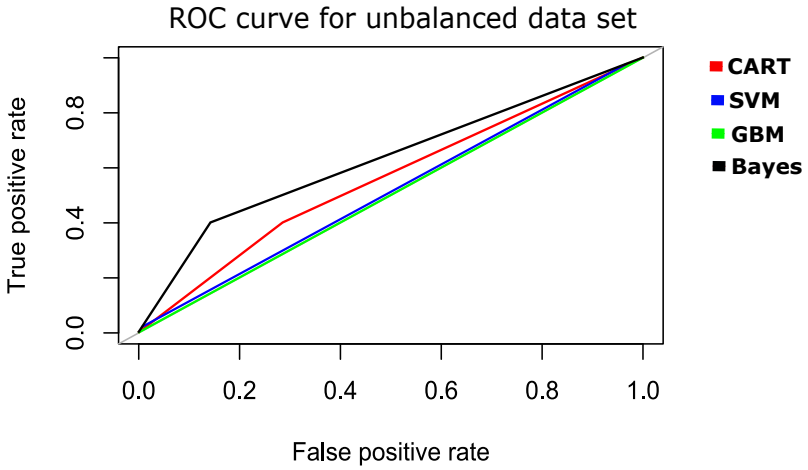

Fig. 1. ROC curves for algorithms using unbalanced data set.

Accuracy of balanced data set. After balancing the data set, the difference between the amount of AP and NAP was reduced for training and testing sets, as shown in Table 4. Distribution of classes were: 53.8% of AP and 46.2% NAP for both training and testing set.

Table 4. Class distribution in balanced data set.

	Approved	Not Approved
Training set	125	107
Testing set	14	12

Best accuracy for each algorithm is presented in Table 5.

ROC curves for each algorithm processing the balanced data set are presented in Figure 2.

Table 5. Accuracy and ROC area for balanced data set.

Algorithm	Accuracy	ROC area
CART	38.46%	0.643
SVM	76.92%	0.774
NB	80.76%	0.815
GBM	57.69%	0.601

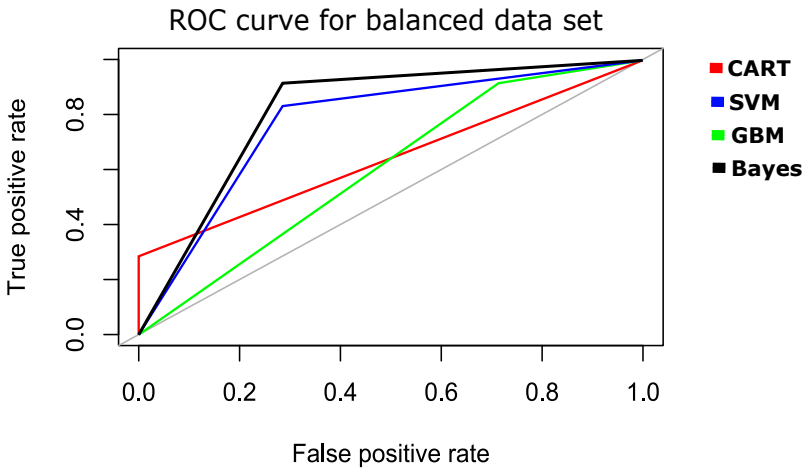

Fig. 2. ROC curves for algorithms using balanced data set.

For unbalanced and balanced data sets, it was created a confusion matrix which is presented in Table 6.

Table 6. Confusion matrix of prediction for unbalanced and balanced data.

		CART		SVM		Bayes		GBM	
Data set		AP	NAP	AP	NAP	AP	NAP	AP	NAP
Unbalanced	AP	15	6	21	0	18	3	21	0
	NAP	3	2	5	0	3	2	5	0
Balanced	AP	10	4	10	2	10	4	4	10
	NAP	12	0	2	10	1	11	1	11

Importance of criteria over approving OOP or not. The importance of each criteria in the prediction was obtained from the model produced by each algorithm. All of them showed the same results for balanced and unbalanced data sets. Figure 3 shows the obtained results.

4 Results Analysis

We evaluated the accuracy and ROC value for each algorithm for both data sets, balanced and unbalanced.

4.1 Unbalanced Data Set

As shown in Table 3, accuracy of SVM and GBM has a good value, more than 80%; however, the ROC area of these algorithms is very low, 50% that is the value of random prediction, this is also presented in Figure 1 where is clear that CART and NB have a bigger area even if their accuracy is smaller.

Analyzing Table 6, it is clear that accuracy of SVM and GBM is because of the classification of AP cases. Both classified correctly total of AP cases, but also classified incorrectly total of NAP cases. This is because there are few cases of NAP class and do not contribute with enough information for prediction. CART and NB classified more NAP cases correctly but less AP ones. The four algorithms are efficiently classifying AP cases correctly, mainly for the mentioned reason about class distribution.

4.2 Balanced Data Set

From Table 5, accuracy obtained analyzing the balanced data set is smaller than the unbalanced one; however, ROC area is bigger. Consider SVM with a 76.9% of accuracy and 0.774 of ROC area and NB with 80.76% of accuracy and 0.815

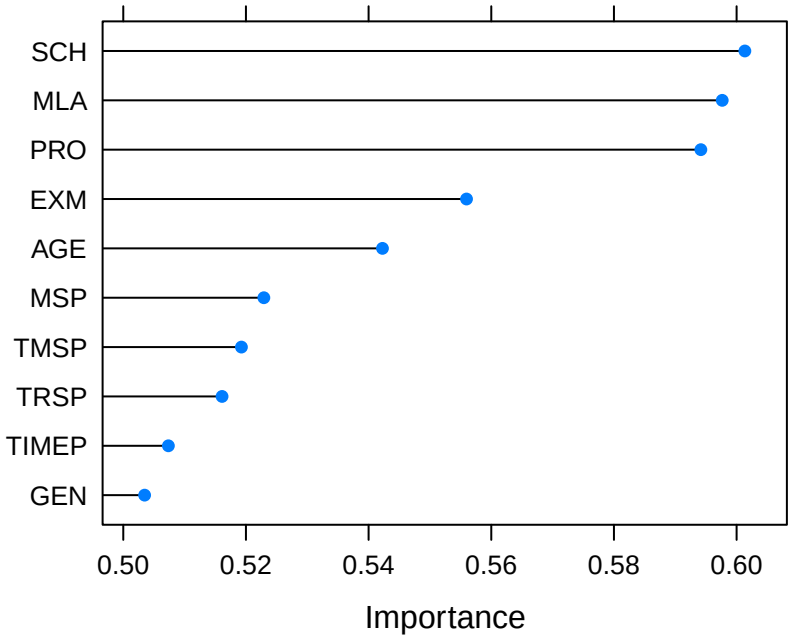

Fig. 3. Importance of criteria for predicting grade in "Object Oriented Programming".

of ROC area. For both algorithms, ROC area and accuracy are closer than using unbalanced data. This results also are presented in Figure 2 where is clear that SVM and NB have a bigger ROC curve area than CART and GBM.

Also, as is shown in the rows corresponding to the balanced results in Table 6, the distribution of correct predictions is bigger for NAP cases which was the main problem in the unbalanced data set. Here, for SVM and NB the distribution of correct predicted classes is more balanced, meanwhile for CART is good for AP cases, but very low for NAP and for GBM is good for NAP and very poor for AP.

4.3 Importance of Criteria in Prediction

Finally, Figure 3 shows that the most important factor in predicting whether a student will approve a programming course or not is related with having a scholarship (SCH), followed by high school GPA (MLA), level of progress (PRO), entrance exam (EXM) and entrance age (AGE).

This represents that most important factors are not related with the performance in a previous course, but with the general performance of the student.

Most important factor related with prerequisite course is the obtained mark (MSP).

5 Conclusions

The goal of this work was evaluating predictive techniques for predicting whether a student will approve a programming course or not, considering both personal and performance criteria over a prerequisite course. Also, analyzing the impact of working with an unbalanced data set has over the predictions and determining if balancing the data set helps for obtaining better results.

Results show that using unbalanced data could generate good results; however these can be misleading because the minority class could not contribute to the prediction and cases of that class could be wrong classified, or if there are not enough cases, could lead to a false high accuracy depending only in the majority class.

Using a balanced data set reduced a little the accuracy of the algorithms; however, results are more reliable because now both classes contribute to the prediction and eliminate the problem produced by the minority class.

About measuring the efficiency of algorithms, using ROC area and its graphic representation (ROC curve) is a better way of evaluating predictions when are analyzed unbalanced data sets. Using only accuracy, comparing right predictions against total of cases is affected by the lack of minority class cases, meanwhile ROC area consider false positive and true positive classifications, so its results are more reliable.

Study case presented the opportunity of working with an unbalanced data set; however the amount of information it is very reduced. This impacts in the efficiency of the algorithms obtaining as best results 80.76% of accuracy and 0.815 for ROC area using Naïve Bayes Classifier and 76.92% of accuracy and 0.774 for ROC area for Support Vector Machine.

From the results, was observed that it is more important when the student has a scholarship and its general performance, than its mark in the prerequisite course.

As future works, a similar analysis for a course that offers more data and that also could have a unbalanced relationship in its classes will help to verify the importance of balancing the data set and considering other evaluating criteria more than accuracy.

References

1. Azevedo, A.I., Santos, M.F.: Kdd, semma and crisp-dm: a parallel overview. IADS-DM (2008)
2. Bakhshinategh, B., Zaiane, O.R., ElAtia, S., Ipperciel, D.: Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies 23(1), 537–553 (2018)

3. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition 30(7), 1145–1159 (1997)
4. Breiman, L.: Classification and regression trees. Routledge (2017)
5. Costa, E.B., Fonseca, B., Santana, M., de Araújo, F., Rego, J.: Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior 73, 247–256 (2017)
6. Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., Rangwala, H.: Predicting student performance using personalized analytics. Computer 49(4), 61–69 (Apr 2016)
7. Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., Van Erven, G.: Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research (2018)
8. Hand, D.J.: Principles of data mining. Drug safety 30(7), 621–622 (2007)
9. Hsu, C., Chang, C., Lin, C.: A practical guide to support vector classification (2003)
10. Lunardon, N., Menardi, G., Torelli, N.: Rose: A package for binary imbalanced learning. R Journal 6(1) (2014)
11. Manjarres, A.V., Sandoval, L.G., Suárez, M.: Data mining techniques applied in educational environments: Literature review. Digital Education Review (33), 235–266 (2018)
12. Murphy, K.P.e.a.: Naive bayes classifiers. University of British Columbia 18 (2006)
13. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Frontiers in neurorobotics 7, 21 (2013)
14. Rajni, J., Malaya, D.B.: Predictive analytics in a higher education context. IT Professional (4), 24–33 (2015)
15. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40(6), 601–618 (2010)
16. Xu, J., Moon, K.H., Van Der Schaar, M.: A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing 11(5), 742–753 (2017)