

# Predicting School Failure Using Data Mining

C. MÁRQUEZ-VERA

Autonomous University of Zacatecas, Mexico

C. ROMERO AND S. VENTURA

University of Cordoba, Spain

---

This paper proposes to apply data mining techniques to predict school failure. We 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. Then two different approaches have been applied in order to resolve the problem of classifying unbalanced data by rebalancing data and using cost sensitive classification. The outcomes of each one of these approaches using the 10 classification algorithms and 10 fold-cross validation are shown and compared in order to select the best approach to our problem.

Key Words and Phrases: School failure, Educational Data Mining, Prediction, Classification.

---

## 1. INTRODUCTION

Recent years have shown a growing interest and concern in many countries about the problem of school failure and the determination of its main contributing factors. This problem is known as the “the one hundred factors problem” and a great deal of research has been done on identifying the factors that affect the low performance of students (school failure and dropout) at different educational levels (primary, secondary and higher) (Araque et al., 2009). A very promising solution to resolve this problem is the use of Data Mining (DM) that is called Educational Data Mining (EDM) when applied to an educational context (Romero and Ventura, 2010). There are examples about how to apply EDM techniques for predicting drop out and school failure (Kotsiantis, 2009). These works have shown promising results with respect to those sociological, economic or educational characteristics that may be more relevant in the prediction of low academic performance. It is also important to notice that most of this research on EDM applied to resolve this problems have been applied primarily to the specific case of higher education (Kotsiantis, 2009) and more specifically to online or distance education (Lykourentzou et al., 2009). However, very little information has been found in specific research on elementary and secondary education, and what has been found only uses statistical methods, not DM techniques (Parker, 1999).

## 2. DESCRIPTION OF THE DATA USED

This paper uses data from students (of about 15 years of age) admitted to the Academic Program 2 of UAPUAZ. It must be pointed out that a very important task in this work was information gathering and data pre-processing due to the quality and reliability of available information which directly affects the results obtained. All the information used in this study has been gathered from three different sources during the period from August to December 2010:

---

Authors' addresses: C. Márquez-Vera, Autonomous University of Zacatecas, Mexico. E-mail: [cmarquezvera@hotmail.com](mailto:cmarquezvera@hotmail.com); S. Ventura, C. Romero, Department of Computer Science, University of Cordoba, Spain; E-mail: [sventura@uco.es](mailto:sventura@uco.es), [cromero@uco.es](mailto:cromero@uco.es);

- a) A specific survey was designed and administered to all students in the middle of the course. Its purpose was to obtain personal and family information to identify some important factors that could affect school performance.
- b) A general survey from the National Evaluation Center (CENEVAL) for admission to many institutions of secondary and higher education. When students register for the admission exam (EXANI I), this Center also carries out a socioeconomic study to obtain this information.
- c) The final scores obtained by students in different subjects in the course provided by the Department School Services of the UAPUAZ at the end of the semester.

Finally, the output variable/attribute or class to be predicted in our problem is the academic status or final student performance that has two possible values: PASS (student who pass the course) or FAIL (student who has to repeat the course). This attribute has been provided by the Academic Program of UAPUAZ at the end of the course.

Starting from all this information we have created a dataset (split in 10 folds) with 77 attributes/variables about 670 students of whom 610 passed and 60 failed the course.

### 3. EXPERIMENTATION AND RESULTS

To do all the experiments, ten classification algorithms have been used that are available in the well-known Weka DM software (Witten et al., 2011): five rule induction algorithms such as JRip, NNge, OneR, Prism and Ridor; and five decision tree algorithms such as J48, SimpleCart, ADTree, RandomTree and REPTree. These algorithms have been selected because they are considered as “white box” classification model, that is, they provide an explanation for the classification result and can be used directly for decision making.

In the first experiment, the 10 classification algorithms have been executed using 10 fold-cross validation and all the available information, that is, the original data file with 77 attributes of 670 students. The results with the test files (an average of 10 executions) of classification algorithms are shown in Table I (A). This table shows the overall accuracy rate (Accuracy), the rates or percentages of correct classification for each of the two classes: Pass (TP rate) and Fail (TN rate) and the Geometric Mean (GM) that is a measure of the central tendency used with unbalanced datasets. It can be seen in Table II (A) that the values generally obtained are high in accuracy (greater than 91.5%) and in the TP rate (greater than 95.7%), but not so high in the TN rate (greater than 25%) and the Geometric mean (greater than 49.9%). The best algorithm in the TP rate and Accuracy was ADTree (99.7% and 97.6% respectively), in the TN rate and Geometric mean was Jrip (78.3% and 87,5% respectively).

After this first experiment using all available attributes, we have notice that in all the obtained model only few of the large number of attributes used (77) appear. So, we decided to do carry out a study of feature selection in order to try to identify which of them has the greatest effect on our output variable (academic status). The objective is to reduce the number of attributes without losing reliability in classification. Weka provides several feature selection algorithms from which we have selected the following ten (Witten et al., 2011): CfsSubsetEval, ChiSquaredAttributeEval, ConsistencySubsetEval, FilteredAttributeEval, OneRAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, ReliefFAttributeEval, SymmetricalUncertAttributeEval. The results obtained have been ranked by these ten algorithms to select the best attributes using our 77 available attributes. In order to find the ranking of the attributes, we have counted the number of times each attribute was selected by one of the algorithms (see

Table I) and only those with a frequency greater than 2 have been selected as the best attributes.

TABLE I. MOST INFLUENTIAL ATTRIBUTES RANKED BY FREQUENCY OF APPEARANCE

Attribute	Frequency
Scores in Humanities 1, and in English 1	10
Scores in Social Science 1, Math 1, Reading and Writing 1, Physics 1, and Computer 1	9
Level of motivation	5
Grade Point Average in secondary	3
Age, number of brothers/sisters, classroom/group, smoking habits, and average score in EXANI I	2
Studying in group, marital status, time spent doing exercises, and score in History	1

The selection of the attributes with a frequency greater than 2 has reduced the dimensionality of our dataset from the original 77 attributes to only the best 15 attributes. Starting from these 15 attributes a second experiment has been carried. Table I (B) shows the results with the test files (the average of 10 executions) using only the best 15 attributes. When comparing the results obtained versus the previous one using all the attributes, that is, Table II (A) versus (B), we can see in general that all the algorithms have improved in some measures (TN Rate and Geometric mean). And about the other measures (TP rate and Accuracy) there are some algorithms that obtain a bit worse or a bit better values, but very similar in general to the previous ones. In fact, the maximum values obtained now are better than the previous ones obtained using all attributes in two evaluation measures (TN rate and Geometric mean). The algorithm that obtains these maximum values is Jrip (81.7% TN rate and 89% Geometric Mean). However, although these results are better than the previous one; they are still very lower than the obtained by TP rate (greater than 95.6% and a maximum value of 99.2%) and Accuracy (greater than 93.1% and a maximum value of 97.3%). This is because our data are imbalanced.

TABLE I. COMPARISON OF RESULTS: (A) USING ALL ATTRIBUTES, (B) USING THE BEST ATTRIBUTES.

Algorithm	A				B			
	TP rate	TN rate	Accuracy	Geometric Mean	TP rate	TN rate	Accuracy	Geometric Mean
Jrip	97,7	<b>78,3</b>	96	<b>87,5</b>	97	<b>81,7</b>	95,7	<b>89</b>
NNge	98,5	73,3	96,3	85	98	76,7	96,1	86,7
OneR	98,9	41,7	93,7	64,2	98,9	41,7	93,7	64,2
Prism	99,5	25	93,1	49,9	<b>99,2</b>	44,2	94,7	66,2
Ridor	96,6	65	93,7	79,2	95,6	68,3	93,1	80,8
ADTree	<b>99,7</b>	76,7	<b>97,6</b>	87,4	<b>99,2</b>	78,3	<b>97,3</b>	88,1
J48	97,4	53,3	93,4	72,1	97,7	55,5	93,9	73,6
RandomTree	95,7	48,3	91,5	68	98	63,3	94,9	78,8
REPTree	98	56,7	94,3	74,5	97,9	60	94,5	76,6
SimpleCart	97,7	65	94,8	79,7	98	65	95,1	79,8

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 (Gu et al., 2008). Traditional classification algorithms have been developed to maximize the overall accuracy rate, which is independent of class distribution; this causes majority class classifiers in the training stage, which leads to low sensitivity classification of minority class elements at the test stage. One way to solve this problem is to act during the pre-processing of data by making a sampling of or a balance of class distribution. There are several data balancing or rebalancing algorithms and one that is widely used and available in Weka as a supervised data filter is SMOTE (Synthetic Minority Over-sampling Technique). In general, SMOTE (Nitesh et al., 2002) introduces minority class elements synthetically, considering the nearest neighbor elements of the same class. In our case, only the 10 training files (with the best 15 attributes) have been

rebalanced using the SMOTE algorithm, obtaining 50% Pass students and 50% Failed students and not rebalancing the test files. The results obtained after re-executing the 10 classification algorithms using 10 fold-cross validation are summarized in Table III (A). If we analyze and compare this Table versus the previous Table II, we can observe that slightly over half of the algorithms have increased the values obtained in all the evaluation measures, and some of them also obtain the new best maximum values in almost all measures except accuracy: Prism (99.8% TR rate), OneR (88.3% TN rate) and ADTree (97.2% Accuracy and 92.1% Geometric Mean).

TABLE III. COMPARISON OF RESULTS: (A) USING DATA-BALANCING, (B) USING COST-SENSITIVE

Algorithm	A				B			
	TP rate	TN rate	Accuracy	Geometric Mean	TP rate	TN rate	Accuracy	Geometric Mean
Jrip	97.7	65	94.8	78.8	96.2	<b>93.3</b>	96	<b>94.6</b>
Nnge	98.7	78.3	96.9	87.1	98.2	71.7	95.8	83
OneR	88.8	<b>88.3</b>	88.8	88.3	96.1	70	93.7	80.5
Prism	<b>99.8</b>	37.1	94.7	59	<b>99.5</b>	39.7	94.4	54
Ridor	97.9	70	95.4	81.4	96.9	58.3	93.4	74
ADTree	98.2	86.7	<b>97.2</b>	<b>92.1</b>	98.1	81.7	<b>96.6</b>	89
J48	96.7	75	94.8	84.8	95.7	80	94.3	87.1
RandomTree	96.1	68.3	93.6	79.6	96.6	68.3	94	80.4
REPTree	96.5	75	94.6	84.6	95.4	65	92.7	78.1
SimpleCart	96.4	76.7	94.6	85.5	97.2	90.5	<b>96.6</b>	93.6

A different approach to solving the problem of imbalanced data classification is to apply cost-sensitive classification (Elkan, 2001). Optimizing the classification rate without taking into consideration the cost of errors can often lead to suboptimal results because high costs can result from the misclassification of a minority instance. In fact, in our particular problem, we are much more interested in the classification of Fail students (the minority class) than Pass students (the majority class). These costs can be incorporated into the algorithm and considered during classification. In the case of 2 classes, costs can be put into a 2x2 matrix in which diagonal elements represent the two types of correct classification and the off-diagonal elements represent the two types of errors. Weka allows any classification algorithm to be made cost sensitive by using the meta-classification algorithm CostSensitiveClassifier and setting its base classifier as the desired algorithm. In fact, the CostSensitiveClassifier and our 10 classification algorithms have been applied as base classifiers using the original test and training files with the best 15 attributes. We have also selected (0, 1; 4, 0) as the cost matrix because it obtained the best results. This matrix indicates that performing the classification takes into consideration that it is 4 times more important to correctly classify Fail students than Pass students. Table III (B) shows the results with test files obtained after applying 10 fold-cross validation. On analyzing and comparing Table III (B) versus Table III (A), some algorithms can be seen to obtain better values in some evaluation measures while other algorithms obtain worse values. So, there is no clear general improvement. However, one algorithm (Jrip) does obtain the current best maximum values on the TN rate (93.3%) and Geometric mean (94.6), which is very important in our problem.

#### 4. DISCUSSION AND FUTURE WORK

Regarding the different used approaches and the classification results obtained, the main conclusions are:

- We have shown the utility of feature selection techniques when we have a great number of attributes. In our case, we have reduced the number of attributes used from the 77 to 15 attributes, without losing classification performance.
- We have shown two ways to address the problem of imbalanced data classification by rebalancing the data and considering different classification costs. These approaches have been able to improve the classification results obtained in one or several evaluation measures.
- We can select cost-sensitive classification as the best approach because it obtains not only very good classification results in the minority class (Fail students), but also in the majority class (Pass students).

Regarding the specific knowledge extracted from the classification models obtained, the main conclusions are:

- White box classification algorithms obtain models that can explain their predictions at a higher level of abstraction by IF-THEN rules. These types of rules are easily understood and interpreted by non-expert DM users. In this way a non-expert user of DM such as a teacher or instructor can directly use the output obtained by these algorithms to detect students with problems (classified as fail) and to make decisions about how to help them and prevent their possible school failure.
- There are some factors/attributes and specific values that appear more in the models when predicting the students who will fail in the classification models obtained. For example, the scores/grades that most appear in the obtained classification rules are the values of “Deficient” or “Not Presented” in the subjects of Physics, Humanities, Math and English. Other factors frequently associated with failing are: to be over 15 years of age, to have more than one brother/sister, to be attending an evening classroom/group, and to have a low level of motivation to study.

Finally, as the next step in our research, we want to develop our own classification algorithm using grammar-based genetic programming and cost sensitive classification for comparison versus other classification algorithms.

## ACKNOWLEDGMENTS

This research is subsidized by P08-TIC-3720 and TIN2008-06681-C06-03 projects, and FEDER funds.

## REFERENCES

- ARAQUE F., ROLDAN C., SALGUERO A. 2009. Factors Influencing University Drop Out Rates. *Computers & Education*, 53, 563–574.
- ELKAN, C. 2001. The Foundations of Cost-Sensitive Learning. *International Joint Conference on Artificial Intelligence*, 1-6.
- Gu, Q., Cai, Z, Zhu, L., Huang, B. 2008. Data Mining on Imbalanced Data Sets. *Proceedings of International Conference on Advanced Computer Theory and Engineering*, Phuket, 1020-1024.
- Kotsiantis S. 2009. Educational Data Mining: A Case Study for Predicting Dropout – Prone Students. *Int. J. Knowledge Engineering and Soft Data Paradigms*, 1(2), 101–111.
- LYKORENTZOU I., GIANNOUKOS I., NIKOPOULOS V., MPARDIS G. AND LOUMOS V. 2009. Dropout Prediction in e-learning Courses through the Combination of Machine Learning Techniques. *Computers & Education*, 53, 950–965.
- NITESH V. CHAWLA ET. AL. 2002. Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
- PARKER A. 1999. A Study of Variables That Predict Dropout From Distance Education. *International Journal of Educational Technology*, 1(2) 1-11.
- ROMERO C. AND VENTURA S. 2010. Educational Data mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics*. 40(6), 601-618.
- WITTEN I. H., EIBE F. AND HALL, M.A. 2011. *Data Mining, practical Machine Learning Tools and Techniques*. Third Edition. Morgan Kaufman Publishers.