ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Scholar Performance Prediction using Boosted Regression Trees Techniques

Bernardo Stearns[1], Fabio Rangel[1], Flavio Rangel[1],
Fabrício Firmino de Faria[1] and Jonice Oliveira[1]

1- Federal University of Rio de Janeiro (UFRJ)
Graduate Program in Informatics (PPGI)
Av. Athos da S. Ramos, 149. Rio de Janeiro, RJ. - Brazil

Abstract.
The possibility of predicting a student performance based only on their socioeconomic status may help to infer what cultural features are important in education. This work was based on scores and socioeconomic data from the most popular exam to enter universities in Brazil: the National High School Exam. Statistical and computational methods used in data mining were applied on a data set of 8 millions data points from Brazil's National High School Exam to examine the predictability of the performance in Mathematics based on socioeconomic status. The results showed that it is possible to predict a students' scores using two ensemble techniques: AdaBoost and Gradient Boosting. The latter presented better results.

1 Introduction

Educational Data Mining is a emerging field which consists in analyzing a large volume of educational data in order to extract non-trivial patterns [1]. By the application of machine learning techniques in the educational field, it is possible to construct complex cognitive modules based on users behavior while playing games [2], aiming to understand the human learning process, or even to create a recommendation system to help students during their learning process, by informing them about topics which they need to strengthen [3].

ENEM (National High School Exam) is a policy created in 1998 by the Brazilian federal government in order to evaluate the quality of high school education and provide a national standardization for knowledge. Although it has always been non-mandatory, the number of subscribers grew over the years, as well as the amount of universities using ENEM scores as admission criteria. The 2014 edition had more than 9.5 million subscribers, whereas the first edition in 1998 had 115 thousand. As for the Brazilian universities, 25 would select students based on ENEM scores in 1999 against more than 5 hundreds nowadays. When subscribing to the exam, students must answer to socioeconomic questions which are freely available to download [1].

The present work studies the feasibility of predicting the student performance using the socioeconomic information as features. Considering the data set volume used in this work, it is important to chose a model capable of learning

[1] http://portal.inep.gov.br/web/enem

with large amount of data. Classification and Regression Trees (CART) [4] can perform well when dealing with large data sets, moreover, CART's based trees are white boxes models and can provide rules in order to understand the regression process, differing from other algorithms, such as Neural Networks and Support Vector Machine, which are black boxes models [5] [6]. Furthermore, ensemble tree methods based on CART algorithms, which use multiple trees to obtain better predictive performance, can be used to enhance the regression task, improving the results [7].

Gradient Boosting [8] and AdaBoost [9] were used to construct decision tree ensembles, and their prediction error were obtained through the application of a 10-Fold Cross Validation. In order to obtain a good set of parameters for the learning algorithms, which is a problem known as model selection, Particle Swarm Optimization was used. The study presents that it is possible to predict the student performance with a Mean Absolute Error of 65.9 points, using Gradient Boosting.

2 Methodology

The present work proposes the use of machine learning methods in order to predict student's scores in the mathematics exam using socioeconomic status as explanatory variables. The mathematics exam presented the higher variance, and due this characteristic was chosen as target. It is considered feasible if the regressing algorithms can predict better than the mean value, i.e. if these algorithms can learn from the features and estimate a better target than the mean value, using the Mean Absolute Error and R-Square metrics.

2.1 Data Set Description

The data set is divided in 3 parts: (i) The first part of the data set contains: data fulfilled by the student on the registration for the exam, which consists in the following features: age, gender, marital status, ethnicity, schooling status, type of school attended (if public or private), state of residence, city of residence; (ii) The second part consisted of answers for the following questions, which were answered also during registration, such as "How long your parents studied?" or "How many people live with you?"; (iii) The last part is the performance on each subjected, which consists in the scores for the following exams: Mathematics, Human Science, Languages, Nature Science and Writing.

2.2 Pre-Processing

The pre-processing applied to the data set in the present work can be summarized in: (i) filtering students who did not attend to the exam since they would not have either scores or socioeconomic information; (ii) and Min-Max Normalization which uses the maximum and minimum value of the feature to change its values. In the Equation 1 the new value v' in the feature A is obtained using the actual value v, the maximum value at this same feature, as well as its minimum value.

The upper value u and lower value l are two variables that indicate the maximum and minimum values desired for the output.

$$v' = \frac{v - min_A}{(max_A - min_A)}(u - l) + l \qquad (1)$$

2.3 Learning Algorithms

Two methods based on CART algorithm that use Boosting technique were used in this work: AdaBoost and Gradient Boosting. Boosting refers to a general and provably effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rules [10]. The following subsections describe both of these methods.

2.3.1 AdaBoost

The detailed theory behind AdaBoost has been reported in [9]. An AdaBoost regressor is a meta-estimator that uses multiples trees combined ("weak learners") to produce a final result. Each weak learn is fitted using part of the data set. The weights of each weak learn instances are adjusted based on the error of the prediction. AdaBoost is sensitive to noisy data and outliers. In some problems, it can be less susceptible to the overfitting problem than other learning algorithms. Each individual learner can be weak, but as long as the performance of each one is slightly better than random guessing, it is guaranteed that the combined result will have an exponential loss [11].

2.3.2 Gradient Boosting

The Gradient Boosting implementation used in this work was the XGBoost. XGBoost is an improvement on Gradient Boost algorithm [12]. Like other boosting methods, Gradient Boosting combines weak learns to produce a final result. To build a XGBoost regressor the trees that minimize a loss function are chosen. A loss function is composed of two factors: an error rate that is calculated over a validation data set and a regularization factor to avoid overfitting the model. XGBoost is a scalable end-to-end tree boosting system [8].

2.4 Evaluation Metrics

The following metrics were used to evaluate the performance of the regressors predictions: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and R-Square (R^2). In the following equations, y_i is the correct target for the example i, and $\hat{y_i}$ is the prediction for the same example.

MAE calculates the absolute distance between the target and the prediction, summing this value for each example. The Equation 2 presents the MAE formula.

MAPE measures the mean absolute percentage deviation. The main reason MAPE was used in this work was the composition of an object function for the hyper-parameter optimization which needed to include an error metric and the R^2. MAPE equation is presented in the Equation 3.

R^2 is a measure that calculates the ratio between the estimator variance and the variance from the observation. Its value ranges from 0 to 1, where 1 is the better performance. In the Equation 4, the R^2 measure is presented with $\bar{y}$ being the mean value for the target y, and $\hat{y}$ is the estimated value for the target y.

$$\sum_{i=1}^{n} |y_i - \hat{y_i}| \qquad (2) \qquad \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{\hat{y_i}}\right| \qquad (3) \qquad 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2} \qquad (4)$$

2.5 Hyper-parameters Optmization

Hyper-parameter optimization is the problem of choosing a set of hyper-parameters for a learning algorithm [13]. The set of hyper-parameters of an algorithm defines a model, and this task is also called model selection. Since our set of hyper-parameters were all numbers, mostly in a continuous space, Particle Swarm Optimization (PSO) was a suitable option for the task. PSO was applied using 100 particles during 100 iterations, using same weights for velocity, *pbest* and *gbest*.

The objective function for the optimization was constructed using MAPE (Mean Absolute Percentage Error) and R^2 (R-Square). For each particle position, the mean and standard deviation was calculated repeating the function presented in the Equation 5 30 times, and for every repetition a sample is taken from the 30% of the data set. Half of this sample is used for training and the remaining is used for evaluation. With a confidence interval of 95%, and error of 5% of the data set size, the sample size used contained 1052 examples. In the end, the result of the objective function is the mean value plus the standard deviation, and it is a minimization problem. In the Equation 5, X is a sample taken from the data set. This function is one of the contribution of this work, which output lies between the range $[0, +\infty)$, and is composed using two different metrics.

$$function(X) = 1 + MAPE(X) - R^2(X) \qquad (5)$$

For XGBoost (Gradient Boosting), the optimized hyper-parameters were: maximum depth, learning rate, minimum child weight, gamma, sub-sample size, column sample by tree, and number of rounds. Where gamma is the weight of the regularization term. For the AdaBoost, the optimized hyper-parameters were: maximum depth, number of estimators, and learning rate.

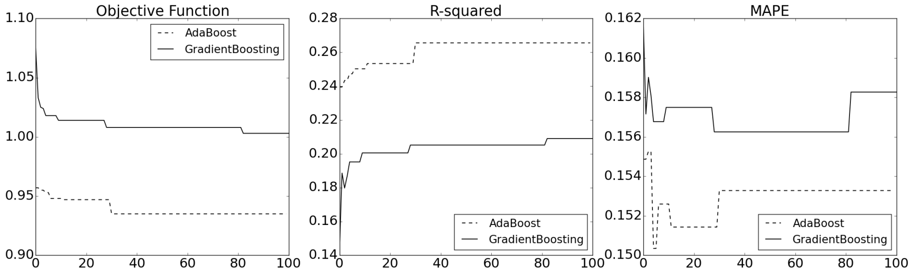
2.6 Validation

The validation step was applied to evaluate the models using the last 70% of the data set. This work used k-Fold Cross Validation [14] with $k = 10$. This validation is used to calculate the final prediction result, after the model selection.

3 Experiment Results

The first step in the experiment aimed to find a good set of hyper-parameters for both ensemble techniques. PSO was applied in this optimization problem using as objective function the one described in Section 2.5. The image of the objective function is presented in the Figure 1. Although the objective function image had a short decrement, it is important to notice that its value is composed of percentage functions, and this short decrement may represent a huge difference in the prediction.

Fig. 1: Search for good set of hyper-parameters using PSO.



After the optimization step, the last 70% of the data set was used for the application of a 10-Fold Cross Validation. A paired t-test was applied in order to statistically validate the results. The null hypothesis states that both models have same average values for the metrics. Each metric average value for the folds is presented in the Table 1. It is possible to notice that Gradient Boosting was superior than AdaBoost in both metrics. Furthermore, both techniques predicts better the target than the mean value (90.27), representing that the predictability is feasible. And predicting for the mean value also gives a zeroed R^2.

Table 1: 10-Fold Cross Validation results.

	Gradient Boosting	AdaBoost	p-value
MAE	65.90 ± 0.11	72.66 ± 1.37	1.27×10^{-7}
R^2	0.35 ± 0.0014	0.18 ± 0.0354	1.34×10^{-7}

4 Conclusion and Future Works

By the application of learning techniques, it was possible to predict the math grade on ENEM using socioeconomic data. The feasibility of this task was showed by comparing the prediction results with the mean value. Furthermore, this work presented a comparison between two ensemble techniques: Gradient Boosting and AdaBoost. PSO was used to perform a model selection for both ensemble techniques.

For future works, we expect to study the feature importance for the models. One way to ranking features is proposed in [15]. Besides that, other model selection techniques could be applied to enhance the exploration in the search.

References

[1] Mehrnoosh Vahdat, Alessandro Ghio, Luca Oneto, Davide Anguita, Mathias Funk, and Matthias Rauterberg. Advances in learning analytics and educational data mining. *Proc. of ESANN2015*, pages 297–306, 2015.

[2] Seong Jae Lee, Yun-En Liu, and Zoran Popovic. Learning individual behavior in an educational game: A data-driven approach. In *Educational Data Mining 2014*, 2014.

[3] Avi Segal, Ziv Katzir, Kobi Gal, Guy Shani, and Bracha Shapira. Edurank: A collaborative filtering approach to personalization in e-learning. In *Educational Data Mining 2014*, 2014.

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[5] N Satyanarayana, CH Ramalingaswamy, and Y Ramadevi. Survey of classification techniques in data mining. *International Journal of Innovative Science, Engineering & Technology*, 1, 2014.

[6] V Krishnaiah, G Narsimha, and N Subhash Chandra. Survey of classification techniques in data mining. *International Journal of Computer Science and Engineering*, 2, 2014.

[7] Michał Woźniak, Manuel Graña, and Emilio Corchado. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16:3–17, 2014.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[9] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997.

[10] Robert E. Schapire. *Nonlinear Estimation and Classification*, chapter The Boosting Approach to Machine Learning: An Overview, pages 149–171. Springer New York, New York, NY, 2003.

[11] Guy Lebanon John Lafferty. Boosting and maximum likelihood for exponential models. *Advances in neural information processing systems*, 14:447, 2002.

[12] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[13] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[14] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. In *Encyclopedia of database systems*, pages 532–538. Springer, 2009.

[15] J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008.