

Mining Student data by Ensemble Classification and Clustering for Profiling and Prediction of Student Academic Performance

Ashwin Satyanarayana

*N-913, Dept. of Computer Systems Technology
New York City College of Technology (CUNY)
300 Jay St, Brooklyn, NY 11229.
{asatyanarayana@citytech.cuny.edu}*

Gayathri Ravichandran

*Dept. of Computer Science
M S Ramaiah Institute of Technology
MSR College Road, MSR Nagar,
Bengaluru, Karnataka 560054, India
{gayathrix7@gmail.com}*

Abstract

Applying Data Mining (DM) in education is an emerging interdisciplinary research field also known as Educational Data Mining (EDM). Ensemble techniques have been successfully applied in the context of supervised learning to increase the accuracy and stability of prediction. In this paper, we present a hybrid procedure based on ensemble classification and clustering that enables academicians -defined cluster

for further advising.

capabilities during team forming and in-class participation. For ensemble classification, we use multiple classifiers (Decision Trees-J48, Naïve Bayes and Random Forest) to improve the quality of student data by eliminating noisy instances, and hence improving predictive accuracy. We then use the approach of bootstrap (sampling with replacement) averaging, which consists of running k-means clustering algorithm to convergence of the training data and averaging similar cluster centroids to obtain a single model. We empirically compare our technique with other ensemble techniques on real world education datasets.

Keywords: *Educational Data Mining, Ensemble Classification, k-means Clustering, Bootstrap averaging, Student academic prediction.*

1. Introduction

The field of Data Mining (DM) is concerned with finding new patterns in large amounts of data. Data Mining (DM) techniques, allow a high level extraction of knowledge from raw data and offer interesting possibilities for the education domain. In particular, several studies have used DM methods to improve the quality of education and enhance school resource management by increasing student retention^{1,2,3,14}.

Educational Data Mining is de

for exploring the unique types of data that come from educational settings, and using those methods to better understand students¹³. The process of tracking and mining student data in order to enhance teaching and learning is one of the goals of Educational Data Mining. Hence, nal environments.

Predicting academic performance o

depends on diverse factors such as personal, socio-economic, psychological and other environmental

variables. Another way to enhance teaching is to identify groups of students with similar learning style and behavioral learning patterns.

The objective of this paper is three-fold: to improve the quality of student data, to predict student academic performance and cluster groups of students with similar learning styles, using data mining techniques such as ensemble classification, anomaly detection and clustering. Ensemble methods have been called the most influential development in data mining and machine learning in the past decade. They combine multiple models usually producing an accurate model than the best of its individual components.

The paper is organized as follows: section 2 surveys data mining techniques for clustering and evaluating student performance, section 3 mentions our contributions, section 4 describes our ensemble (filtering, ensemble classification and clustering

3. To use bootstrap averaged k-means clustering to identify groups of students with similar learning styles.

4. Methodology

4.1 Ensemble Noise Filtering

We propose an ensemble classifier framework for noise filtering and predicting student performance. We show that by having more than one classifier (or model) to evaluate the instances, we *extend* the model space as compared to a single classifier. Thus, by using multiple (in this paper, we use three) classifiers we perform *an approximation*

Algorithm: Bootstrap Averaging

Input: D: Training Data, T: Number of bags, K: Number of clusters

Output: A: The averaged centroids.

```

// Generate and cluster each bag
(1) For i = 1 to T
(2)    $X_i = \text{BootStrap}(D)$ 
(3)    $C_i = \text{k-means-Cluster}(X_i, K)$  // Note  $C_i$  is the set of  $k$  cluster centroids and  $C_i = \{c_{i1}, c_{i2}, \dots, c_{ik}\}$ 
(4) EndFor
// Group similar clusters into bins with the bin averages stored in  $B_1, \dots, B_k$  their sizes are  $S_1, \dots, S_k$ 
(5) For i = 1 to T
(6)   For j = 1 to K
(7)     Index = AssignToBin( $c_{ij}$ ) //See section on signature based comparison
(8)      $B_{\text{Index}} += c_{ij}$ 
(9)   EndFor
(10) EndFor
(11) For i = 1 to K
(12)    $B_i /= S_i$ 
(13)    $A_i = B_i$ 
(14) EndFor

```


Fig 2.

5.1 Student Performance Dataset (UCI):

This dataset is based on a study of data collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal¹¹. The database was built from two sources: school reports, based on paper sheets and including few attributes (i.e. the three period grades and number of school absences); and questionnaires, used to complement the previous information. The final version contained 37 questions in a single A4 sheet and it was answered in class by 788 students. Latter, 111 answers were discarded due to lack of identification details (necessary for merging with the school reports). Finally, the data was integrated into two datasets related to Mathematics (with 395 examples) and the Portuguese language (649 records) classes¹¹.

Table 1. Attributes of the UCI Student performance dataset.

In this work, the Mathematics and Portuguese grades (i.e. G3 of Table 1) will be modeled using 5-Level classification (Table 2) based on the Erasmus (European exchange program) grade conversion system as used by Cortez¹¹. The results are shown in Table 3.

16-20	14-15	12-13	10-11	
A	B	C	D	

2016 ASEE Mid-Atlantic Section Conference

Dataset	Predictive accuracy of student academic performance		
	<i>Decision Tree (J48)</i>	<i>Online Bagging</i>	<i>Ensemble Filtering</i>
Mathematics	0.78	0.82	0.95
Portugese	0.71	0.79	0.94

Table 3. Predictive accuracies after using the different classification techniques

As we can see in Table 3, ensemble filtering which uses multiple classifiers to vote and eliminate noisy instances in the training data produces higher statistically significant predictive accuracies on the test data

2016 ASEE Mid-Atlantic Section Conference

As was done in the previous section, we used ensemble classifiers to firstly eliminate noisy instances and then to predict the final grade of the students on the test set. We use a *majority vote* (which requires at least two out of the three classifiers to mislabel the class value) amongst the classifiers in eliminating the noisy instances. The predictive accuracy numbers are as shown in Table 5.

Dataset	Predictive accuracy of student academic performance	
	<i>Decision Tree (J48)</i>	<i>Online Bagging</i>

2016 ASEE Mid-Atlantic Section Conference

18. Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F., & Seong, H. Y. (2009, December). Predicting NDUM student's academic performance using Data mining techniques. In *Computer and Electrical Engineering, 2009. ICCEE'09. Second International Conference on* (Vol. 2, pp. 357-361). IEEE.
19. Etchells, T. A., Nebot, À., Vellido, A., Lisboa, P. J., & Mugica, F. (2006). Learning what is important: feature selection and rule extraction in a virtual course. In *ESANN* (pp. 401-406).
20. Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.

Ashwin Satyanarayana

Dr. Ashwin Satyanarayana is currently an Assistant Professor with the Department of Computer Systems Technology, New York City College of Technology (CUNY). Prior to this, Dr. Satyanarayana was a Research Scientist at Microsoft in Seattle from 2006 to 2012, where he worked on several Big Data problems including Query Reform Science (Data Mining) from SUNY, with particular emphasis on Data Mining, Machine Learning and Applied Probability with applications in Real World Learning Problems. He is an author or co-author of over 20 peer reviewed journal and conference publications and co-authored a textbook

Engine research. He is also a recipient of the Indian National Math Olympiad Award, and is currently serving as Secretary/Treasurer of the ASEE (American Society of Engineering Education) Mid-Atlantic Conference.

Gayathri Ravichandran

Gayathri Ravichandran is currently pursuing her final year of undergraduate study in M S Ramaiah Institute of Technology, Bangalore, India. Her field of study is Computer Science, and she finds subjects like Data Mining and Artificial Intelligence intellectually stimulating and satisfying. She has authored a paper titled

under the guidance of her professor. She is also a recipient of the Grace Hopper scholarship (GHCI) 2016.