Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance

Ashwin Satyanarayana, Mariusz Nuckowski

N-913, Dept. of Computer Systems Technology,
New York City College of Technology (CUNY),
300 Jay St, Brooklyn NY – 11201.

{ *asatyanarayana@citytech.cuny.edu, mariusz.nuckowski@mail.citytech.cuny.edu* }

Abstract:

In the last decade Data mining (DM) has been applied in the field of education, and is an emerging interdisciplinary research field also known as Educational Data Mining (EDM). One of the goals of EDM is to better understand how to predict student academic performance given personal, socio-economic, psychological and other environmental attributes. Another goal is to identify factors and rules that influence educational academic outcomes. In this paper, 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 also identify association rules that influence student outcomes using a combination of rule based techniques (Apriori, Filtered Associator and Tertius). We empirically compare our technique with single model based techniques and show that using ensemble models not only gives better predictive accuracies on student performance, but also provides better rules for understanding the factors that influence better student outcomes.

1. Introduction

Education is a crucial element in our society. Business Intelligence (BI)/Data Mining (DM) techniques, which allow a high level extraction of knowledge from raw data, offer interesting possibilities for the education domain. In particular, several studies have used BI/DM methods to improve the quality of education and enhance school resource management. Hence, the ability to predict students' academic performance is very important in educational environments. Predicting academic performance of students is challenging since the students' academic performance depends on diverse factors such as personal, socio-economic, psychological and other environmental variables. The scope of this paper is to predict student performance and to determine the factors that influence the academic performance of students, using data mining techniques such as Classification, Filtering and Association Rule Mining. In this paper we use two DM tasks: classification and association rules.

Ensemble methods have been called the most influential development in data mining and machine learning in the past decade. They combine multiple models into one usually more accurate than the best of its components. In this paper, we propose an ensemble classifier framework for analyzing student performance. In the area of classification, we focus on improving the quality of student academic training data by identifying and eliminating mislabeled instances by using multiple classifiers. In the area of generating association rules, we use multiple rule based models to vote on all the individual rules generated by the individual association rule generating algorithms.

The paper is organized as follows: section 2 surveys data mining techniques for evaluating student performance, section 3 mentions our contributions, section 4 describes our ensemble (filtering, association rules) techniques in detail, and section 5 shows our experimental results using datasets from the UCI repository. Finally, in section 6 the conclusions and further research are outlined.

2. Prior Work in this area

Alaa el-Halees [2] show that data mining can be used in educational settings to understand the learning process of identifying, extracting and evaluating variables related to the learning process of students.

Han and Kamber [1] provide a good description of the different data mining tools and software on multidimensional data and their analysis.

Bayes classification was used by Pandey and Pal [3] for student performance prediction based on 600 students from different colleges of Awadh University, Faizabad, India. They use attributes such as category, language and background qualification of students.

Linear regression used by Hijazi and Naqvi [4] on the student performance prediction based on a sample of 300 students (225 males, 75 females) from different colleges affiliated to Punjab university of Pakistan. They consider attributes such as attendance, hours spent studying, family income, mothers age, mothers education. They found that the factors like mother's education and student's family income were highly correlated with the student academic performance.

Several other lines of research have explored data mining methods to predict student academic performance such as: Neural networks for giftedness identification [5], Predicting student performance using data mining with educational web-based system [6], Determination of factors influencing the achievement of the first year university students using data mining [7], Application of GMDH algorithm for modeling of student's quality [8], Predicting persistence of students using data mining methods [9] and Application of data mining methods to the student's dropout problem [10].

3. Our Contribution

Our contributions in this area are as follows:

1. To use data mining filtering techniques on student data to improve the quality of the data.

2. To use ensemble filtering technique to create a more accurate prediction of student performance

2. To use ensemble association rules to create more accurate association mining rules.

4. Methodology

4.1 Ensemble Noise Filtering

An ensemble classifier detects noisy instances by constructing a set of classifiers (base level detectors). A majority vote filter tags an instance as mislabeled if more than half of the *m* classifiers classify it incorrectly. A consensus filter requires that all classifiers must fail to classify an instance as the class given by its training label.

Our filtering approach begins by performing *k*-fold cross validation. *k*-fold cross validation is a commonly used technique which takes a set of *n* examples and partitions them into *k* sets of size *n/k*. For each fold, multiple classifiers are trained on all the other folds and tested on the current fold. Thus *k* hypotheses θ_1, θ_2, …….θ_k are generated. This prediction is equivalent to outputting the average of *k*-hypotheses as shown in equation (1) below:

$$\Pr(y = t \mid x, \theta) = \frac{1}{k}\sum_{i=1}^{k} \delta(t, \theta_i) \qquad (1)$$

where δ is a 0-1 loss function, which returns 1 if θ_i predicts the correct label *t*, else returns 0.

Our Ensemble Filtering algorithm is as shown in Fig 1. It begins with *k* almost equal sized subsets of our dataset *E* (step 1) and an empty output set *A* of detected noisy examples (step 2). The main loop (steps 3-12) is repeated for each fold E_i. In step 4, we form a set E_y which includes all the examples from *E* except E_i. E_y is used as an input for the *k* inductive learning algorithms to generate models *k* models $\theta_{y,1}$, $\theta_{y,2}$……. $\theta_{y,j}$. The set E_i is evaluated by our *j* models in steps 8-11. If more than half of the models misclassify an instance, then it is treated as noise and eliminated.

```
Algorithm: EnsembleFiltering ($E$)
Input: $E$ (training set)
Parameter: $k$ (number of subsets of $E$, typically 10)
           $j$ (number of inductive learning algorithms, typically 3)
Output: A (a detected noisy subset of $E$)
(1) Form $k$ almost equal sized subsets of $E_i$, where $\cup_i E_i = E$
(2) A ← Ø
(3) for i = 1,……,$k$ do
(4)     $E_y \leftarrow E \setminus E_i$
(5)     for $m$ = 1……$j$ do
(6)         $\theta_{y,m}$ ← model built from bootstrap sample $E_y$ and inductive algorithm $m$
(7)     end for
(8)     for every $e \in E_i$ do
(9)         If $e$ is misclassified by more than half the $\theta_{y,m}$ models built, then it is noisy and needs to be eliminated.
(10)         A ← A U {$e$}
(11)    end for
(12) end for
```

Fig 1. Ensemble Filtering Algorithm

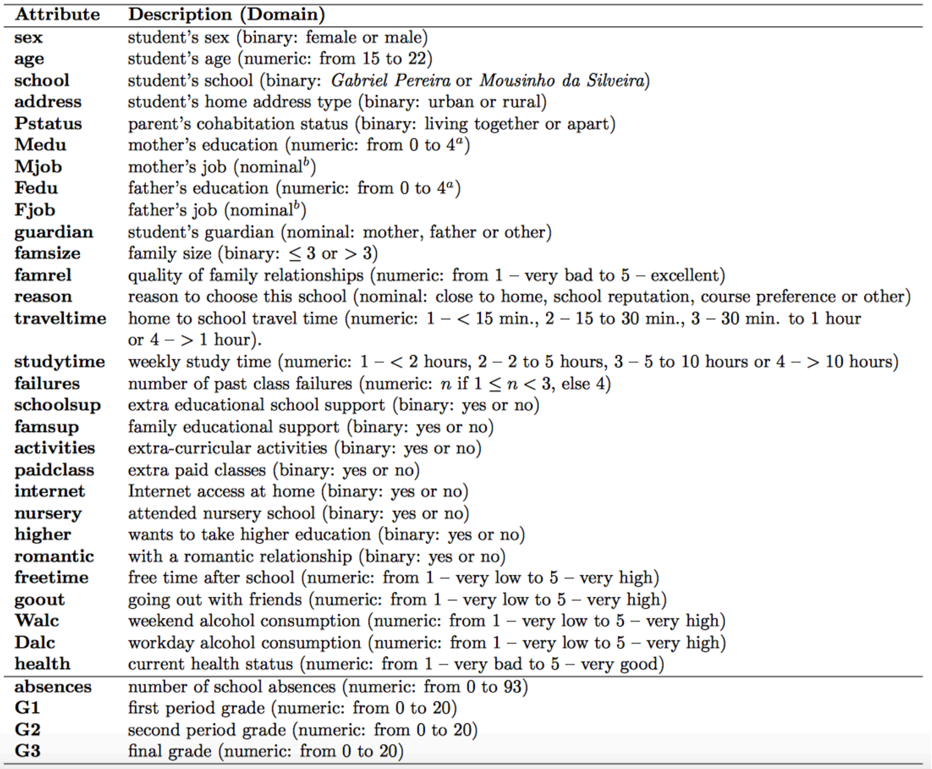
5. Empirical Results

In this section, we discuss our experiments that demonstrate the improved predictive accuracy using our ensemble filtering approach as compared to single model filtering. We tested our approach on two datasets: (a) UCI Student Performance dataset [11] and (b) New York City College of Technology CST introductory course dataset. For each dataset, we compare the accuracies after filtering using the following techniques:

1. Single Model: We used decision trees (J48) as our single filtering base model.
2. Online Bagging: We implemented online bagging as illustrated by Oza [12] using Naïve Bayes as the base model.
3. Ensemble Filtering: Our algorithm (shown in Fig 1) uses the following classifiers: J48, RandomForest and Naïve Bayes. We use *consensus vote* for Student performance dataset and *majority vote* for the dataset from New York City College of Technology.

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 Portugal [11]. 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) classes [11].



Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4^a)
Mjob	mother's job (nominal^b)
Fedu	father's education (numeric: from 0 to 4^a)
Fjob	father's job (nominal^b)
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
studytime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

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 Cortez [11]. The results are shown in Table 3.

16-20	14-15	12-13	10-11	0-9
A	B	C	D	F

Table 2. Five level classification of the final grade G3

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

Table 3. Predictive accuracies after using the different classification techniques

We find the following commonly voted association rules using 3 association mining techniques: Apriori, Filtered Associator and Tertius. We found that using ensemble voting provided stronger factors that determine student achievement, than using any individual algorithm. The top 5 voted rules are as shown in Fig 2.

```
schoolsup=no AND paid=no AND internet=yes AND G2=Fail ==> class=Fail conf:(0.95)

schoolsup=no AND internet=yes AND Dalc=1 AND G2=Fail ==> class=Fail conf:(0.94)

Pstatus=T AND schoolsup=no AND paid=no AND G1=Fail ==> class=Fail conf:(0.93)

famsize=GT3 AND Pstatus=T AND internet=yes AND G2=Fail ==> class=Fail conf:(0.92)

traveltime=1 AND schoolsup=no AND paid=no AND G2=Fail ==> class=Fail conf:(0.91)
```

Fig 2. Ensemble Association Rules generated by Apriori, Filtered Associator and Tertius

As shown in Fig 2, we find factors that cause a student to fail the finals such as: (a) no extra educational support from school (schoolsup=no), (b) Daily alcohol consumption (Dalc=1), (c) Large Family size (famsize=GT3) (d) Internet access at home (internet=yes) and (e) failed the previous test (G2=fail). As expected, we find a strong correlation between failing G2 and failing the final exam.

5.2 First year college student performance dataset

First year Computer Systems Technology students from the New York City College of Technology (CUNY) enrolled in 6 different semesters (Fall 2013, Fall 2014, Fall 2015 Spring 2013, Spring 2014 and Spring 2015) taking an introductory computer systems course was used for this study. The same professor taught all the semesters. Data from students who dropped the class or stopped attending the class were excluded from the study. The class has two tests, a midterm and a final. We attempt to predict the final grade given the two test scores and the midterm score. The five level classification for the final grade is as shown in Table 4.

>=80	60-80	40-60	30-40	<30
A	B	C	D	F

Table 4. Five level classification of the final grade G3

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. We use a majority vote 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		
	Decision Tree (J48)	*Online Bagging*	*Ensemble Filtering*
CST Course	0.63	0.75	**0.91**

Table 5. Predictive accuracies after using the different classification techniques

6. Discussion and Conclusion

Resolving data quality issues in predicting student academic performance is often one of the biggest efforts in Educational Data Mining. Prior work in this area has focused on using single classifiers and no filtering on student data has been performed.

In this work, we show that student data when filtered can show a huge improvement in predictive accuracy. We compare using a single filters with ensemble filters and show that using ensemble filters works better for identifying and eliminating noisy instances. We show that both types of voting (majority and consensus) can show improvements. We have shown that this ensemble technique works for two different settings: high school data and first year college data. Although we have used decision trees, random forest and naïve bayes, other base classifier models can also be used.

In our future work, we would like to explore other data mining techniques such as clustering to identify groups of students who have similar academic performance.

7. References:

1. J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann, 2000.
2. Alaa el-Halees, “Mining students data to analyze e-Learning behavior: A Case Study”, 2009.
3. U . K. Pandey, and S. Pal, “Data Mining: A prediction of performer or underperformer using classification”, (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.
4. S. T. Hijazi, and R. S. M. M. Naqvi, “Factors affecting student‟s performance: A Case of Private Colleges”, Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.
5. Hyuk Kwang, et al.,*Conceptual Modeling with Neural Network for Giftedness Identification and Education*, Lecture Notes in Computer Science, Volume 3611, pp. 560-538,2005.
6. Minaei-Bidgoli, B., et al., *Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA*,Proceedings of ASEE/IEEE Frontiers in Education Conference, Boulder, CO: IEEE, 2003.
7. Superby, J.F., Vandamme, J-P., Meskens, N., *Determination of factors influencing the achievement of the first-year university students using data mining methods*, Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006), Jhongli, Taiwan, Pages 37-44, 2006.

8. Naplava, P. and Snorek N., *Modeling of student's quality by means of GMDH algorithms*, Modelling and Simulation 2001, 15th European Simulation Multiconference 2001, ESM'2001, Prague, Czech Republic, 2001.
9. Luan, J. and Serban, A. M., *Data Mining and Its Application in Higher Education*, Knowledge Management: Building a Competitive Advantage in Higher Education, New Directions for Institutional Research, Jossey-Bass, 2002.
10. Massa, S. and Puliafito P. P., *An application of data mining to the problem of the university students' dropout using Markov chains*, Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD'99, Prague, Czech Republic, 1999.
11. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
12. Oza, N. C. (2005, October). Online bagging and boosting. In Systems, man and cybernetics, 2005 IEEE international conference on (Vol. 3, pp. 2340-2345). IEEE.