Using Data Mining to Predict Student Academic Performance

Ashwin Satyanarayana and Mariusz Nuckowski

Department of Computer Systems Technology

14th Annual City Tech Poster Session

Student grade prediction

- Can we predict student grades?
- Can we determine the most important factors that determine academic success?

Overview

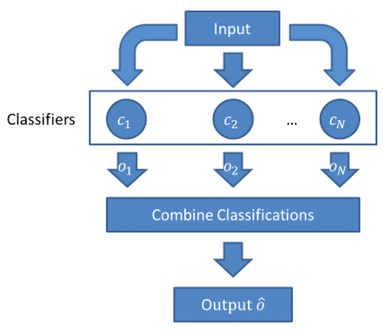
- Data Mining (DM) techniques, which allow a *high level extraction of knowledge* from raw data, offer interesting possibilities for the education domain (Educational Data Mining).
- Predicting academic performance of students is challenging since the students' academic performance depends on diverse factors such as:
  - personal,
  - socio-economic,
  - psychological and
  - other environmental variables.

Our contributions are as follows:

1. To use data mining filtering techniques on student data to improve the quality of the data.
2. To use ensemble (i.e more than one classifier) techniques to create a more accurate prediction of student performance
2. To use ensemble association rules to create more accurate association mining rules.

Background: Ensemble Classifiers

- Instead of using one base classifier, we use multiple classifiers with voting between them to identify bad records (instances).
- 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.



Predictive accuracies of the different techniques

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**

Factors that determine Student Grades

```
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)
```

Weka: Data Mining Software

- Collection of machine learning algorithms
  - open-source package written in Java
- Used for research, education and application
- Main features:
  - data pre-processing tools
  - learning algorithms
  - evaluation methods
  - graphical inference
  - environment for comparing learning algorithms

Dataset 1: UCI Student Performance

- 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. 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 data was integrated into two datasets related to Mathematics (with 395 examples) and the Portuguese language (649 records).

Dataset 1: UCI Student Performance

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)

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

Predictive accuracies of the different techniques

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**

Factors that determine Student Grades

```
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)
```

Dataset2: CST 1100

- 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. 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.

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

Predictive accuracies for CST1100

As was done with the previous dataset, we used ensemble classifier (J48, Random Forest and Naïve Bayes) 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 below:

Dataset	Predictive accuracy of student academic performance		
	Decision Tree (J48)	*Online Bagging*	*Ensemble Filtering*
CST Course	0.63	0.75	**0.91**

Why does it work?

- Bayesian Analysis of Ensemble Filtering:
  - Ensemble Filtering: An ensemble classifier detects mislabeled instances by constructing a set of classifiers (*m* base level detectors). In this paper, we use **Decision Trees, Random Forest and Naïve Bayes** as our base classifiers.
  - A majority vote filter tags an instance as mislabeled if more than half of the *m* classifiers classify it incorrectly

$$\Pr(y = t \mid x, \Theta^*) = \frac{1}{|\Theta^*|} \sum_{\theta \in \Theta^*} \delta(t, \theta(x))$$

Conclusion

- We show that using data mining (with ensemble classifiers) with majority voting can show very high predictive accuracies of student grades.
- We show empirically that this techniques works for two different settings: high school data and first year college data.

References

Ensemble Noise Filtering for Streaming Data using Poisson Bootstrap Model Filtering
Ashwin Satyanarayana, Rosemary Chinchilla
13th International Conference on Information Technology : New Generations (ITNG 2016), Las Vegas, NV, April 11th-13th, 2016

Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance
Ashwin Satyanarayana, Mariusz Nuckowski
ASEE Mid-Atlantic Section Spring 2016 Conference, George Washington University, Washington D.C, April 8-9, 2016.