Predicting Student Performance with Data from an Interactive Learning System

Ana Gonçalves[1], Ana Tomé[2], Luís Descalço[1]

[1]Center for Research & Development in Mathematics and Applications, University of Aveiro, Aveiro, PT
[2]Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Aveiro, PT

Introduction

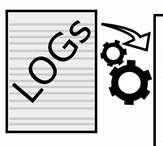
Nowadays Interactive Learning Systems have been developed to provide students with new forms of practicing concepts. In this work we propose to predict if the student fails or succeeds in the introductory mathematics course based on the information collected by an interactive learning platform. The predicting models are based on binary support vector machines (SVM). As some of the collected data sets are unbalanced the study was conducted with suitable strategies to train this binary classifier.

SIACUA and Feature extraction

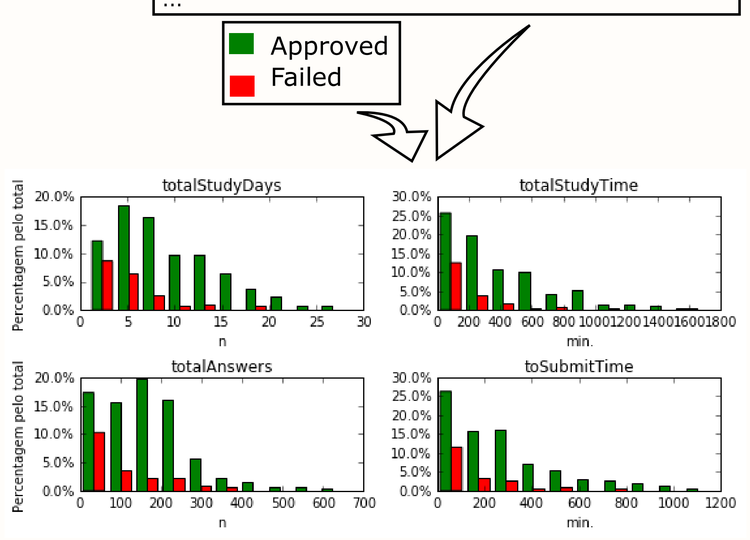
SIACUA - Sistema Interativo de Aprendizagem por Computador, Universidade de Aveiro - is a web application designed to support autonomous study. The information about student interaction with system comprises

- student ID
- time step for each question
- elapsed time between question and answer
- correct, incorrect or solution visualization.

The LOG files are processed and 42 features are calculated



totalStudyDays: number of days the student submited at least one answer
totalStudyTime: amount of time the student spent answering questions
totalAnswers: number of answers submited by the student
toSubmitTime: total amount of time the student spent submiting answers
...



Features concerning evaluation periods and features concerning class periods have low correlation.

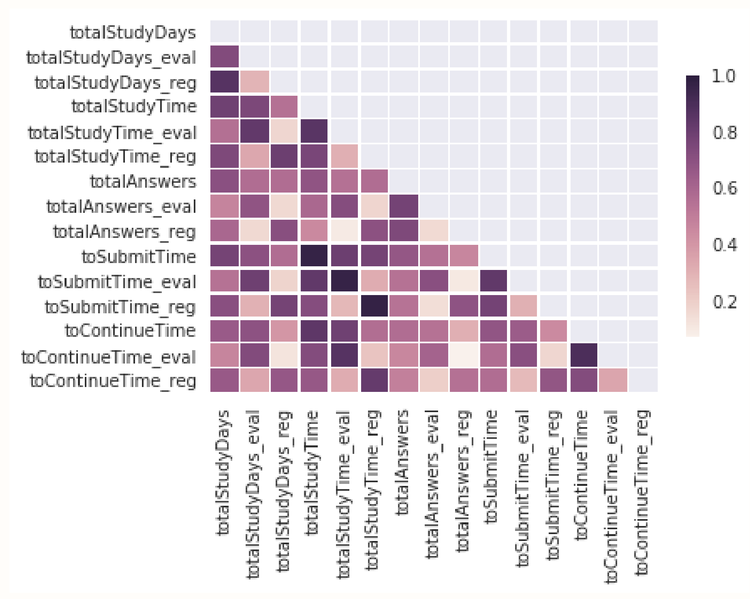


Figure 1: Features heat map.

Funding

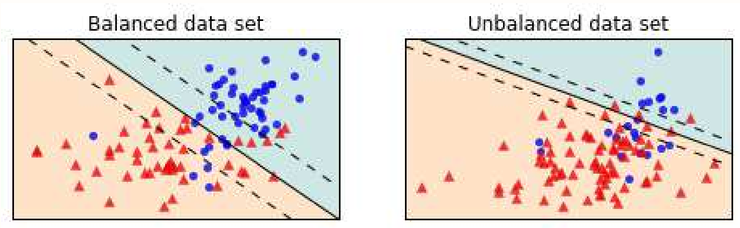
This work is funded by National Funds through the FCT - Foundation for Science and Technology, in the context of projects UID/CEC/00127/2013 and UID/MAT/04106/2013 .

Linear SVM and unbalanced data

SVM Learning

Support Vector Machine (SVM) is a reliable two class classifier whose training goal is to find the decision surface $\mathbf{w}$ such that

$$g(\mathbf{z}) = \mathbf{w}^T\mathbf{z} \Rightarrow \begin{cases} g(\mathbf{z}) > 0, & y = 1 \\ g(\mathbf{z}) < 0, & y = -1 \end{cases}$$



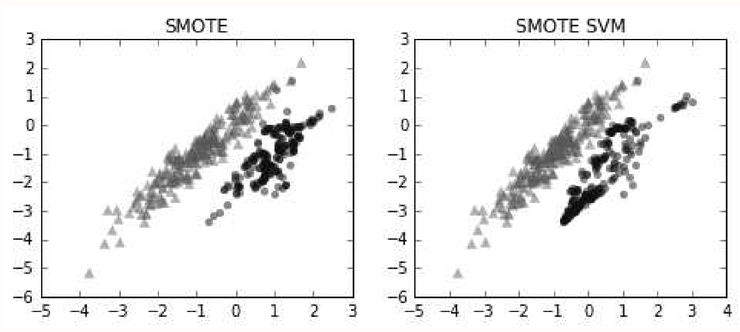
With unbalanced training data the decision surface (solid line) is in most of the cases closer to the majority class.

Learning strategies for unbalanced data

Different Error Costs (DEC): during training misclassifications of minority have more weight than the ones of the majority class.

Synthetic Over-sampling + DEC: Training with synthetic data

- SMOTE:
  - Synthetic instances of minority class are generated.
  - The new data is created on line segments that connect nearest neighbors of minority.
- SMOTE SVM
  - Train SVM.
  - Synthetic instances of minority class are generated based on support vectors of the minority class.



The illustration shows

- SMOTE: the minority class is denser having its instances uniformly distributed over the original instances area.
- SMOTE SVM: the new minority class instances are concentrated along the decision boundary

z-SVM: SVM classifier is trained, afterwards the decision boundary is adjusted in order to correct the bias towards the majority class.

Results and Discussion

Data analysis was performed in Python with software package *scikit-learn* for

- Feature normalization (z-score and min-max).
- Training SVM classifier (with linear and RBF kernel).
- Training Linear SVM using unbalanced strategies.
- Cross-validation: to optimize classifier parameters and performance evaluation.
- Performance measures: geometric mean g and accuracy acc.

Datasets: concerning two disciplines

Table 1: Students by discipline and group.

Discipline	Approved	Failed	Total
Cálculo 2	187	140	327
Cálculo 3	250	62	312

Table 2: Cálculo 3: Accuracy and geometric mean g for unbalanced data.

	none		min-max		z-score	
Classifier	**acc.**	***g***	**acc.**	***g***	**acc.**	***g***
SVM linear	0.785	0.304	0.801	0	0.795	0.218
SVM RBF	0.801	0	0.801	0	0.801	0
DEC	**0.644**	**0.656**	**0.670**	**0.667**	0.587	0.655
SMOTE+DEC	0.603	0.644	0.628	0.651	0.487	0.579
S.SVM+DEC	0.641	0.643	0.715	0.602	**0.663**	**0.656**
z-SVM	0.385	0.422	0.401	0.344	0.494	0.344

With default training the unbalanced data set shows poor performance in minority class.

- Linear SVM model is much biased towards the majority class.
- SVM RBF the minority class is not learned.

For the studied data sets, the approaches for unbalanced data based on different error costs for missclassified objects present better results.

- DEC and SMOTE SVM + DEC classifiers are the ones with highest accuracy and geometric mean.
  - DEC: original data and min-max normalized data
  - SMOTE SVM + DEC: z-score normalized data
- These results suggest that is not always worth it the extra processing for data generation, before applying DEC classifier.

Table 3: Cálculo 2: Accuracy and geometric mean g for balanced data.

	none		min-max		z-score	
Classifier	**acc.**	***g***	**acc.**	***g***	**acc.**	***g***
SVM linear	0.719	0.712	0.749	0.740	0.728	0.724
SVM RBF	0.731	0.723	0.752	0.735	0.740	0.733
DEC	0.716	0.712	0.725	0.726	0.734	0.740
SMOTE+DEC	0.713	0.715	0.722	0.722	0.621	0.627
S.SVM+DEC	0.694	0.698	0.716	0.716	0.618	0.623
z-SVM	0.719	0.713	0.728	0.725	0.706	0.707

There is no advantage to use unbalanced strategies.

- RBF SVM is slightly better than linear SVM.
- With unbalanced strategies, z-SVM included,the performance decreases (less than 3 %).

Conclusion

The preliminaries results presented in this work show that the features are relevant for decision making and strategies for unbalanced data sets improve the classification of linear SVM. In what concerns the prediction of student achievements it is too premature to draw conclusions because data sets are too small to be considered representative of the population.