Exploratory Clustering approach of Students Learning Behavior

ashish.akshantal[1]

[1]Affiliation not available

January 10, 2018

Introduction

In recent years, there has been increasing interest in Learning Analytics (LA) and Educational Data Mining (EDM) among both researchers and practitioners of the field. By emerging the Computer-assisted learning systems and automatic analysis of educational data, many efforts have been carried out in order to enhance the learning experience .

There are three main groups for any Computer-assisted learning systems : Educators, Learners, and Administrators. Educators are responsible to design and plan the educational systems, and they are the most aware of the students' learning process, their needs, and common mistakes.

Educators providing real-time feedback into the performance of learners is a great help for this group to adapt their teaching activities to the students' needs.

Learners benefit from recommendation and feedback on their learning activities, resources, and paths. The type of feedback given to the students can be motivating and encouraging.

Finally, administrators are dealing with decision-making and budget allowance, and can influence the process of improving the systems and learning resources

LA and EDM are both two emerging fields that have a lot in common, although they have differences in their origins and applications. LA is a multi-disciplinary field that involves ML, artificial intelligence, information retrieval, statistics, and visualization. Additionally, it contains the Technology Enhanced Learning (TEL) areas of research such as EDM, recommended systems, and personalized adaptive learning . EDM is concerned with: "developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist". While, LA initially was defined as: "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs".

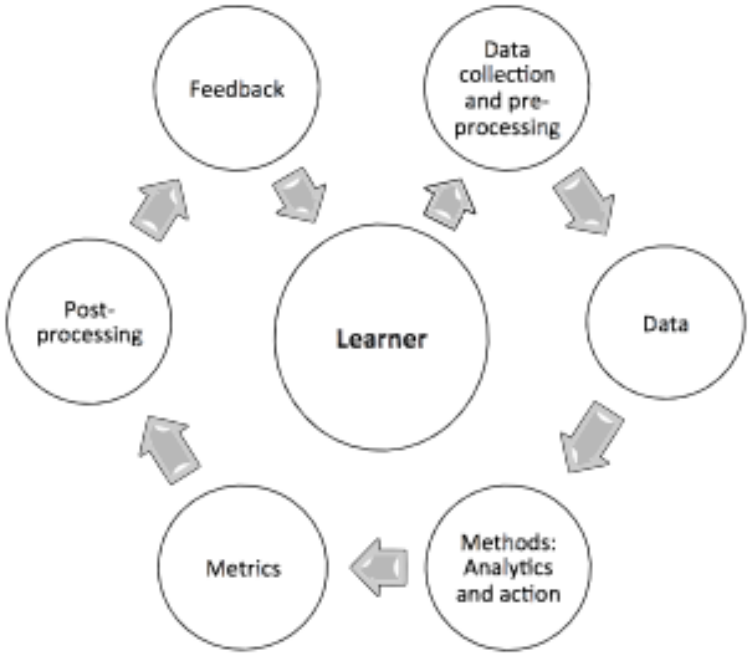

Figure 1: **An LA/EDM process starts with learner whose data is collected and analyzed, and after post processing, feedback and interventions are made in order to optimize learning**

Problem Statement

Identifying how Students of University of Genova have reacted to the Computer Assisted Learning Systems[DEEDS(Digital Electronics Education and Design Suite)] and has this system helped students in improving their performance.

Data Set Information

The Data set is about the experiments which were carried out with a group of 115 students of first-year, undergraduate Engineering major of the University of Genoa. The study was carried over a simulation environment named Deeds (Digital Electronics Education and Design Suite) which is used for e-learning in digital electronics. The environment provides learning materials through specialized browsers for the students and asks them to solve various problems with different levels of difficulty.

The data set includes the following files :

features_info.txt': contains information about the variables used on the feature vector.

'activities_info.txt': contains information about the variable 'activity'.

'exercises_info.txt': contains information about the variable 'exercise'.

'grades_info.txt': contains information about the grade data.

Data:

'Processes': contains the data files from Session 1 to 6.

- 'logs.txt': shows information about the log data per student Id. It shows whether a student has a log in each session (0: has no log, 1: has log).

- 'final_grades.xlsx': contains the results of the final exam in two sheets.

- 'intermediate_grades.xlsx': contains the grades for the students' assignments per session.

- 'final_exam.pdf': shows the content of the final exam (original in Italian).

- 'final_exam_ENG.pdf': shows the content of the final exam translated in English.

Data Integration

The Data was not in a single file and there was no extension for the files.

Totally there were 594 files of the Data set.

Created a Batch file to add extensions(.csv) to all the files.

Used another Batch file to merge all the files into one single file.

Data Preprocessing

Clustering is done taking in consideration the distance between two records. There are few distances for specific data types, like we have Euclidean Distance, Manhattan Distance and Minkowski distance for Numeric Data Type and Hamming Distance and Jaccard Distance for Categorical Data Type.

Depending on the Problem, one chooses any of the above distance metric to find the distance between two records.Before applying any clustering models on the data, we have to standardize the data to bring all the attributes to a common unit, so that the distance metric will not be affected.I have used z-score standardization on the data.

Initially converted all the categorical variables to dummies so that the distance can be calculated but, dummifying the categorical variables have increased the variables to 255 from 15(originally).

This is very high dimensions (curse of dimensionality) in clustering and performing a clustering algorithm on this data will not give any good clusters results.

Dropped a variable Activities from the data which contained 99 levels and reduced the dimensions to 156.

Feature Engineering

From End time and start time, calculated the time difference and added to the features. Removed the Start time and End time features from the data set, because distance metrics would not work on time data type.

Model Building

Run K-Means clustering on the data in R, but was getting a memory error.

Error: cannot allocate vector of size 197.6 G. Used H2o and applied K-Means on the Data.

From the Final Grades data set, I have created a new column "Grade" by assigning Grades to each student based on the total they got.

Changed the problem to a classification task by assigning grades to Students based on their Final Grades Total.Total number of records obtained after filtering are 98. There were 16 questions and each question had different weight-age. Total marks for the exam was 100.

The exam was held in two times (in two sheets) and some students took the exam two times. In both times, the exams addressed the same concepts but with different details. Some students who attended the course did not take the final exam, therefore, some Ids are missing in final grades.

The questions of the final exam addressed the concepts of sessions of the course. So, we provide the grades per question based on their reference to the sessions topics in addition to the total final grade. The column names indicate ES # of session. # of exercise (the total points dedicated to exercise).

Used Binning (manual) to bin the students into 3 categories (A, B, C)

Models Built

1. Logistic regression – Target (2 class) Removed the Student ID, combined (Session, Activity, Exercise) and the data was not Standardized.

Threshold chosen – 0.4 (After the ROC Curve) . For different threshold, (I have tried 0.50) the accuracy was further decreasing).Data set Used for this model is 'All_Students_with_grades.csv'. Here I have converted the milliseconds to minutes.

Accuracy = 54.91(threshold = 0.4)

Accuracy = 43.16(threshold = 0.5)

Accuracy Test = 48.78

2. Converted the milliseconds in Idle_time to seconds and again ran a logistic regression model. I have predicted on the validation using the above model

For threshold 0.40 I was getting Accuracy – 48%

For threshold according to the ROC curve – 0.48 - accuracy – 42%

Test Accuracy = 50.15

Here I have separately used the (Session, Activity and Exercise attributes)

AIC: 157403

3. Applied the Step-AIC model for the above. There was no change in the Step-AIC value.

Test Accuracy = 45.97

4. Applied the GLM on the same data, but this time, I have Standardized it and applied Step-AIC.

Accuracy = 49.01%

Test Accuracy = 47.24%

5. Random forest on the Standardized data (dropped activity column)

Accuracy = 60.71 %

Test Accuracy = 58.27%

When plotted the variable importance for the model, it showed that only 5 variables are more important.

Selected

Mouse movement,

time_diff,

exercise,

mouse_click_left,

idle time.

6. RF_model_imp_variable s:

Variables used: mouse_movement + idle_time + time_diff + exercise + mouse_click_left

accu_val_rf_imp = 0.568

Test Accuracy = 54.15%

7. SVM model – standardized and without activity

Accuracy_svm_val = 0.5682113

Test Accuracy = 52.78%

8. Using Polynomial kernel, accuracy on validation

accu_val_svm_poly = 56.8245

Accuracy Test = 53.24%

9.Random forest with standardized and three classes, [dropped activity, student ID but session is present] (here only for students present for all sessions)

accuracy on the validation set - 0.4879224

accuracy on test = 49.75%