Data mining and DSP analysis of social media for student performance prediction

Student: Andreu Grimalt Reynes

Supervisor: Matthew Yee-King

1 Final project proposal

In my final project I would like to explore the fields of Data Mining, Machine Learning and Neural Networks to try to provide answers to questions about education.

The growth in the use of VLEs (Virtual learning environments), MOOCs (Massive Online Open Courses) and other online education systems provides an increasing stream of data regarding student behaviour. It is now possible to use data mining algorithms to analyse student habits and try to find correlations between student behaviour and academic performance.

Teachers and researchers have begun to explore the possibility of using social media in teaching and learning. In general, the use of social networks generate data sets that accurately describe their users behaviour. When social media is used for learning, the data set generated can be used to parametrise the learning experience and therefore to optimise it. This enables us to think about social networks, education and student behaviour as a classification problem. That is to say, to use the data generated in social media and apply data mining algorithms to it in order to classify students in groups according to their academic performance. Once the algorithms are trained in recognising these groups, they can predict future performance and allow the optimisation of the learning.

Code Circle is a social network for sharing and discussing programming code. Its aim is to make it easy to share code with a community of users and to provide tools that encourage feedback and the discussion of work in progress. Therefore, being able to give specific feedback on the code (i.e being able to select particular bits on the code and attach comments to them) and being able to modify the code and run it on the browser are key features. Code circle was developed at Goldsmiths as part of a group summer project, partly using software components developed within the PRAISE research project. It is aimed at programming students, it can run Processing sketches on the browser and display

the code of a particular sketch. It is currently used in a case study involving approximately 200 undergraduate students from an introductory programming course.

In Code Circle all the user interaction is logged. This data set contains information about the user, the time, the type and the element that received the user action. The data has this structure: {*user_id: Number, time: Number, type: String, id: Number, logotype: String*}

- user_id: The user id that performed the action
- time: Timestamp when the action occurred
- type: The type of action (click, mouseover, etc...)
- id: Id of the element that received the action
- logtype: Text description of the element that received the action

In my project I would like to apply data mining techniques to the analysis of the data set generated within Code Circle.
The data analysis will address the following research questions:

- Can data mining and neural networks algorithms classify student academic performance?
  Try to train a system to recognise "good" and "bad" behaviour in students. This is a supervised learning problem which will use the grades of the students as ground truth. Several data mining algorithms can be used and compared using performance measures.
- Can DSP be used to pre-process the data and perform feature reduction on it? How does it perform compared to other dimensional reduction techniques?
  Code Circle's log data depends on the time and it can be treated as a signal. I would like to explore how DSP can be used with data mining in order to pre-process the data and perform dimensional reduction. Also, I would like to measure how it performs compared to other feature reduction techniques.

2 References

Ian Witten, Eibe Frank, Mark Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier Science & Technology, 2011.

Romero, Cristóbal and López, Manuel-Ignacio and Luna, Jose-María and Ventura, Sebastián, *Predicting students' final performance from participation in on-line discussion*

forums, Computers & Education, vol. 68, p458-472, Elsevier Ltd, 2013.

Hundhausen, Christopher D and Carter, Adam S, *Facebook me about your code: An empirical study of the use of activity streams in early computing courses*, Journal of Computing Sciences in Colleges, vol. 30, p151-160, 2014.

Kulkarni, Chinmay and Wei, Koh Pang and Le, Huy and Chia, Daniel and Papadopoulos, Kathryn and Cheng, Justin and Koller, Daphne and Klemmer, Scott R., *Peer and self assessment in massive online classes*, ACM Transactions on Computer-Human Interaction (TOCHI), vol. 20, 2013.

J.P. Vandammea, N. Meskensa, J.F. Superbya, *Predicting Academic Performance by Data Mining Methods*, Education Economics, p405-419, 2007.