MOOC Data Analytics: Social Network Analysis of Discussion Forum Data

Xu Lanxiao

Supervisor: Dit-Yan Yeung

Introduction

A massive open online course (MOOC) is an online course aimed at unlimited participation and open access via the web. Currently, MOOC and its several platforms have gained great popularity by providing diverse online courses and learning-oriented services such as forum discussion, assignments and exams. However, because users of online courses can access course resources freely and generally receive zero penalties when dropping out of a course, the number of active users often tends to decrease greatly during the whole course period.

Under this circumstance, it is desirable that a system can be designed to find the potential dropout users based on their past behaviors so that administrators of the course will be able to send reminders promptly to encourage them to stay active. Moreover, many MOOC courses use grades to rate the performances of their users. By conducting grade prediction during the course period, we can see whether earlier behaviors offer clues to the ultimate evaluation of performances, which offers interesting insights in social analysis. This project mainly focus on mining the user patterns though forum social analysis and utilize them with other features in dropout and user performance prediction. Machine learning techniques, including both supervised learning and unsupervised learning algorithms are used in this work.

Dataset

The MOOC data used in this project is extracted from the datasets of a Coursera's July 2013 offering of the Hong Kong University of Science and Technology's "the Science of Gastronomy, a six-week course with 85314 registered users. The datasets were obtained from the Coursera data coordinator in HKUST. Table 1 offers some statistics about this course.

Number of users	85314
Number of dropout users	68058
Number of forum posters	3761
Number of forum voters	2701
Number of threads in the forum	1594

Table 1: Statistics of the MOOC course – the Science of Gastronomy

User Patterns

Now we are trying to reveal these user patterns by utilizing principal component analysis (PCA). PCA is an unsupervised learning approach for dimension reduction. It can reduce high-dimensional data into low-dimensional data while maintaining as much the original information as possible. Suppose in forum data there are n threads and m users, we will obtain a user vector of length n if we record the activities of one user in each thread. By recording the activities of all the users, we will obtain an m × n activity matrix, where each entry contains a value representing the activities of a certain user within a certain thread (In our case, m equals to 3761 and n equals to 1594). By conducting PCA, we are able to reduce the m × n activity matrix into an m × 2 matrix whose values can be plotted on a 2-D plot. Then we can visually see whether there are any patterns for user activities. Figure 2 shows the visualization of the PCA results computed on binary activity matrix, where each entry of the activity matrix is a binary code '0' or '1' representing whether a user has contributed to a certain thread.

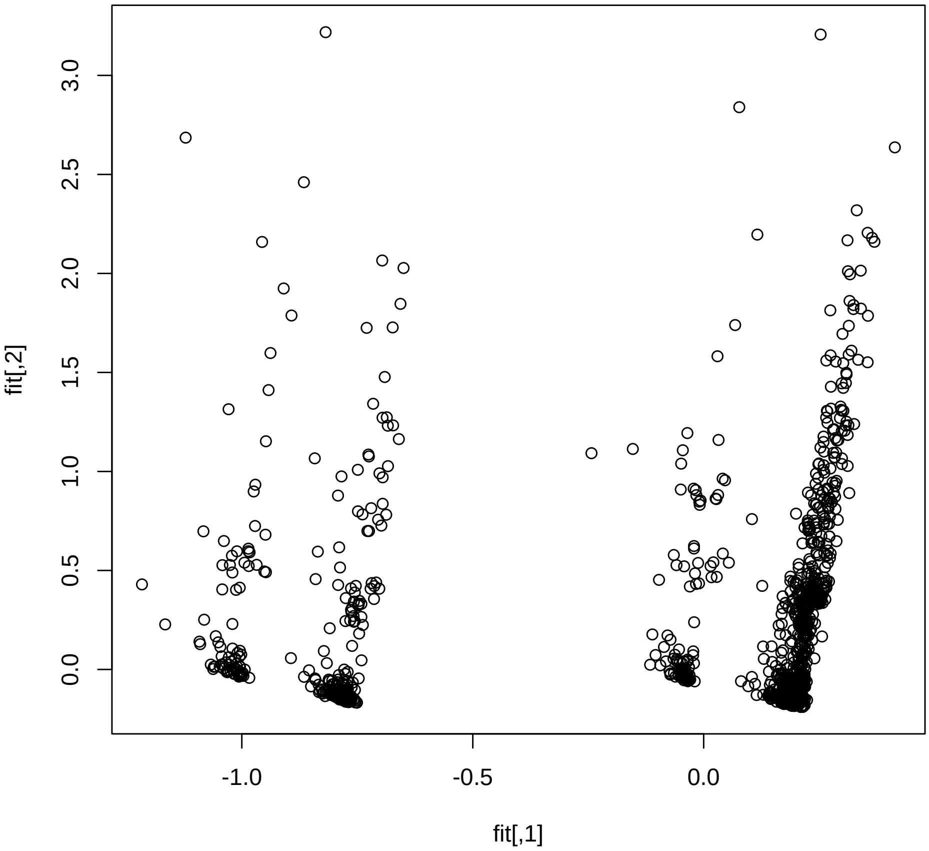

Figure 2: PCA results from the binary representation

Prediction Model

Compared with single predictor, supervised learning models have the advantage of combining different measures to obtain higher prediction outcome. To achieve this goal, a classification model in support vector machines (SVMs) with radial basis function (RBF) kernels was selected.

Given training vectors $x_i \in R^n$, $i = 1,..., l$ and a label vector $y \in R^l$ such that $y_i \in \{1, -1\}$, the model solve the following optimization problem.

$$\min_{\boldsymbol{w},b,\zeta} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\zeta_i$$

$$\text{Subject to } y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \zeta_{i,}$$

$$\zeta_i \geq 0$$

Prediction Result

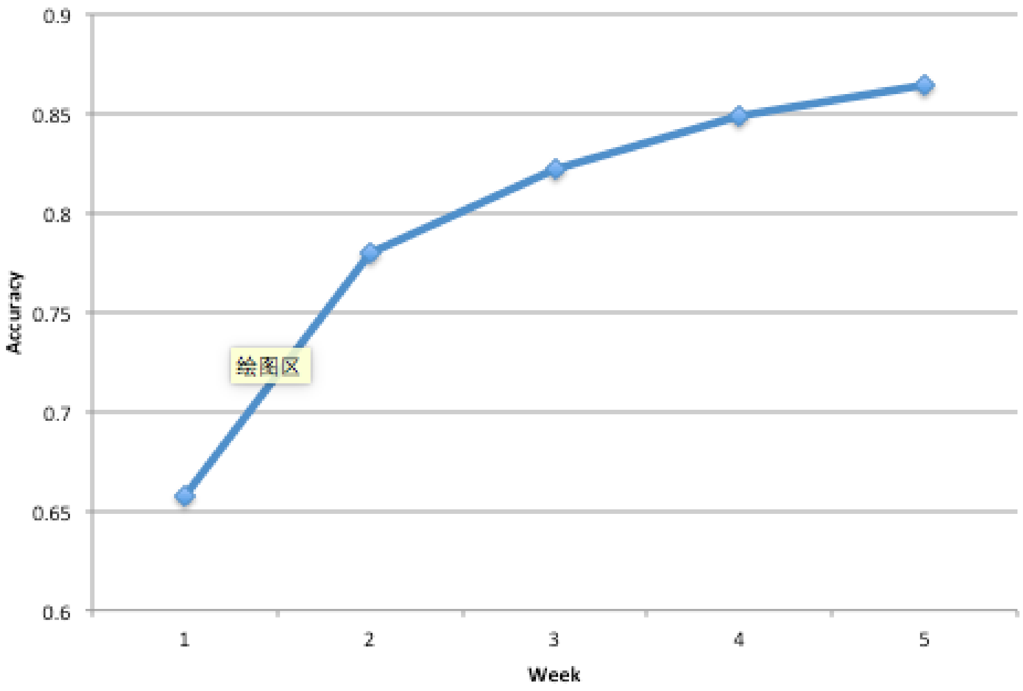

Figure 3: Accuracies of the Dropout Prediction

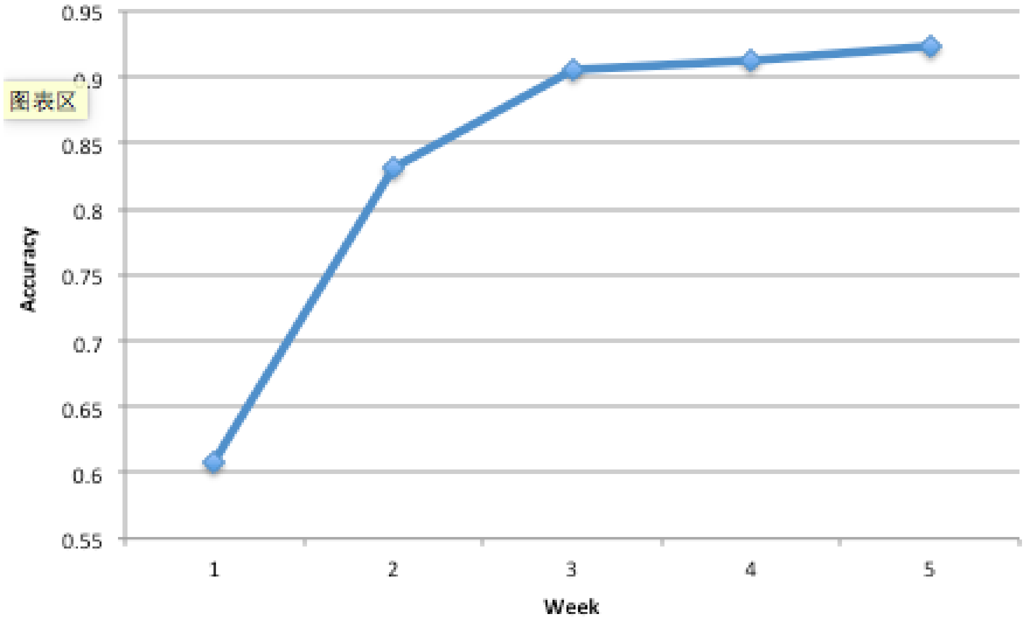

Figure 4: Accuracies of the User Performance Prediction, which is a pass prediction guessing whether a student will pass this course