An Application of Collaborative Filtering in Student Grade Prediction

Chaloemphon Sirikayon[1] **and Panita Thusaraon**[2]

College of Innovative Technology and Engineering
Dhurakij Pundit University
Bangkok, Thailand
E-mail: chaloemphon.sir@dpu.ac.th, panita.thu@dpu.ac.th
www.dpu.ac.th

Abstract

This research presents the process of student performance prediction by using the collaborative filtering (CF). The benefit of this research includes assist instructor to identify student performance, personalized advising, and student degree planning. The CF technique composes of similarity calculation and prediction. In our experiments, a prior course clustering with heuristic knowledge is adopted and different techniques of similarity calculation are compared. The performance of each student has been predicted by using existing grades available at that time.

Keywords: grade prediction, collaborative filtering, educational data mining

1. Introduction

Student performance prediction in future course is important as it provides valuable information to facilitate student success. In this paper, we present the process of student performance prediction by using the collaborative filtering: CF [1], which is one of the most popular techniques wildly used for student performance prediction. The performances that students achieved in the earlier courses are used to predict grade that they will obtain in future courses. The algorithm is based on the idea of finding the most similar students. We have performed various methods to calculate students' similarity, i.e. Pearson correlation, cosine similarity, and Euclidian distance. The performance of each method is experimentally evaluated on a dataset obtained from Dhurakij Pundit University with enrollments of 200 undergraduate students between 2012 and 2016 from the Faculty of Information Technology. Our experiments shows that finding students' similarity with Pearson correlation achieves the lowest prediction error and a prior course clustering with heuristic knowledge can enhance predictability.

The rest of this paper is organized as follows: related work and fundamental concept is given in Section 2. The proposed method is described in Section 3. Experiments are conducted in Section 4. Conclusions are summarized in Section 5.

2. Preliminaries

This section summarizes related work and briefly defines the fundamental concept needed to facilitate the presentation of the proposed algorithm.

2.1. *Related work*

Different models have been developed in order to predict student's performance and many approaches rely on collaborative filtering methods. The similarities of students are calculated utilizing their study results, represented by the grades of their previously passed courses. A recommendation tool called the personalized Grade Prediction Advisor (pGPA) was proposed in [3]. The system allows user to set parameters such as number of similarity students used for prediction. Another course recommender system for University College Dublin's on-line enrolment application was proposed in [2]. The system recommends elective modules to students based on the core modules that they have selected by using item-based collaborative filtering.

On the other hand, [5] presented future course grade prediction methods that utilize approaches based on linear regression and matrix factorization. Hybrid methods and content features are also used in [6].

2.2. *User-based collaborative filtering*

Collaborative Filtering algorithm is based on the main idea that people have similar preferences and interests. One user's behavior is compared with other user's behavior to find his/her nearest neighbors, and according to his/her neighbor's preferences or interest to predict his/her preferences or interest. Suppose that U = {u_1, u_2,..., u_m} is a list of m users and I = {i_1, i_2,..., i_n} is a list of n items. Each user U_i gives rating scores for a list of items I_{ui}. The prediction problem is to predict the rating active user U_a will give to an item I_{ua} from the set of all items that U_a has not yet rated. The CF technique composes of 3 steps as follows: 1) users similarity calculation 2) top N nearest neighbors selection and 3) prediction.

2.2.1. *Similarity and distance*

Various methods can be used to find similarity between users such as Pearson correlation and cosine similarity. On the other hand, dissimilarity calculation, i.e. Euclidean distance can be converted to similarity.

2.2.1.2 *Pearson correlation*

Let the set of items rated by both users u and v be denoted by I, then similarity coefficient $sim(u,v)$ between them is calculated as

$$sim(u,v)=\frac{\sum_{i\in I_u\cap I_v}(r_{u,i}-\bar{r}_u)(r_{v,i}-\bar{r}_v)}{\sqrt{\sum_{i\in I_u\cap I_v}(r_{u,i}-\bar{r}_u)^2}\sqrt{\sum_{i\in I_u\cap I_v}(r_{v,i}-\bar{r}_v)^2}} \quad (1)$$

Here $r_{u,i}$ denotes the rating of user u for item i , and $\bar{r}_u$ is the average rating of all items given by user u. Similarly, $r_{v,i}$ denotes the rating of user v for item i, and $\bar{r}_v$ is the average rating of all items given by user v.

2.2.1.2 *Cosine similarity*

The similarity $sim(u,v)$ between user u and v is calculated as

$$s(u,v)=\frac{r_u\cdot r_v}{\| r_u \|^2 \| r_v \|^2}=\frac{\sum_i r_{u,i}r_{v,i}}{\sum_i r_{u,i}^2 \sum_i r_{v,i}^2} \quad (2)$$

where $r_{u,i}$ denotes the rating of user u for item i , and $r_{v,i}$ denotes the rating of user v for item i.

2.2.1.3 *Euclidean distance*

Euclidean distance for two user u and v is calculated by

$$d(u,v)=\sqrt{\sum_{i\in I_u\cap I_v}(r_{u,i}-r_{v,i})^2} \quad (3)$$

Here $r_{u,i}$ denotes the rating of user u for item i , and $r_{v,i}$ denotes the rating of user v for item i . Then, obtained distance scores are converted to similarities by

$$sim(u,v)=\frac{1}{1+d(u,v)} \quad (4)$$

2.2.2. *Prediction*

Once similarities are calculated, a set of top-k users most similar to the active user u are selected and their rating scores are used for the prediction $P_{u,i}$ of the specific item i for user u as follow:

$$P_{u,i}=\bar{r}_u+\frac{\sum_{v\in N}s(u,v)(r_{v,i}-\bar{r}_v)}{\sum_{v\in N}| s(u,v) |} \quad (5)$$

3. Dataset and Method

For each student whose grade needs to be predicted, a set of similar students are identified by using their grades from courses that they have already taken. The data used for this study obtained from Dhurakij Pundit University with enrollments of 200 undergraduate students between 2012 and 2015 from the Faculty of Information Technology. The dataset comprised of 200 students and their 12,000 grades. The A-F letter grades were converted to the 4–0 scale. The performance of each student who enrolled in semester 2, 2015 has been predicted by using grades available at that time.

3.1. *Prediction without prior courses clustering*

Our study have comprised of 3 steps in user-based CF to make a prediction for each student as follows:

Step 1: Calculate similarity between the active student S_a and every other user by using Pearson correlation, cosine similarity, and Euclidian distance.

Step 2: Based on their similarity scores, various set of k students, most similar to active student S_a is then selected.

Step 3: Prediction for grade student S_a will receive from the course i is generated by using grades of course i that k similar neighbors have already taken.

Table 1. Example of similarity scores for student S_1

Top N neighbors	Similarity Calculation Method		
	Pearson	Cosine	Euclidean
1	025: 0.3033	025: 0.9466	023: 0.2450
2	010: 0.2843	013: 0.9429	041: 0.2297
3	013: 0.2393	010: 0.9427	055: 0.2240
4	073: 0.2205	023: 0.9408	067: 0.2240
5	102: 0.2121	065: 0.9400	036: 0.2222
6	064: 0.2102	068: 0.9396	040: 0.2188
7	022: 0.2072	073: 0.9396	052: 0.2188
8	023: 0.2065	022: 0.9395	004: 0.2171
9	041: 0.2040	014: 0.9392	082: 0.2171
10	075: 0.2033	019: 0.9386	046: 0.2139

Table 1. shows an example of similarity scores for student S_1 obtained by various approach. The result is slightly different for Pearson correlation and cosine similarity methods. For Pearson correlation method, top 3 most similar students are student S_{25}, S_{13}, and S_{10} with similarity scores 0.9466, 0.9249 and 0.9427, respectively. The result from cosine similarity method depicts that top 3 most similar students are student S_{25}, S_{10} and S_{13}, respectively. On the other hand, top 3 most similar students form Euclidean distance are student S_{23}, S_{41}, and S_{55}.

Once similarities for each student are obtained, the performance of each student who enrolled in semester 2, 2015 has been predicted as show in Table 2.

Table 2. Example of grade prediction for student S_1 with different similarity approach

Method	Neighbors Size (N)				
	25	30	35	40	45
Subject: IS203 Real Grade: 2(C)					
Pearson	2 (1.85)	2 (1.85)	2 (1.85)	2 (1.85)	2 (1.85)
Cosine	2 (1.74)	2 (1.74)	2 (1.77)	2 (1.78)	2 (1.79)
Euclidean	2 (1.86)	2 (1.86)	2 (1.86)	2 (1.87)	2 (1.87)
Subject: IS306 Real Grade: 2.5(C+)					
Pearson	2 (1.90)	2 (1.88)	2 (1.87)	2 (1.87)	2 (1.87)
Cosine	2 (1.93)	2 (1.87)	2 (1.83)	2 (1.86)	2 (1.83)
Euclidean	2 (1.87)	2 (1.86)	2 (1.85)	2 (1.86)	2 (1.85)

3.2. *Performance evaluation*

The performance evaluations were conducted using accuracy measure and root mean square error (RMSE) obtained by Eq.(6) and Eq.(7), respectively.

$$Accuracy = \frac{\sum Corrected \ \ Answer}{|Subjects| \times |Users|} \times 100 \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{7}$$

where $\hat{Y}_i$ is a predicting value of subject i and Y_i is a real value of subject i.

4. Experiment

We compared the predicted grades with the actual grades of students who enrolled in semester 2, 2015. The