Collaborative Multi-Regression Models for Predicting Students' Performance in Course Activities

Asmaa Elbadrawy
Computer Science
University of Minnesota
Twin Cities, MN
asmaa@cs.umn.com

R. Scott Studham
University of Minnesota
Twin Cities, MN
studham@umn.edu

George Karypis
Computer Science
University of Minnesota
Twin Cities, MN
karypis@cs.umn.com

ABSTRACT

Methods that accurately predict the grade of a student at a given activity or course can identify students that are at risk in failing a course and allow their educational institution to take corrective actions. Though a number of prediction models have been developed, they either estimate a single model for all students based on their past course performance and interactions with learning management systems (LMS), or estimate student-specific models that do not take into account LMS interactions; thus, failing to exploit fine-grain information related to a student's engagement. In this work we present a class of collaborative multi-regression models that are personalized to each student and also take into account features related to student's past performance, engagement and course characteristics. These models use all historical information to estimate a small number of regression models shared by all students along with student-specific combination weights. This allows for information sharing and also generating personalized predictions. Our experimental evaluation on a large set of students, courses, and activities shows that these models are capable of improving the performance prediction accuracy by over 20%. In addition, we show that by analyzing the estimated models and the student-specific combination functions we can gain insights on the effectiveness of the educational material that is made available at the courses of different departments.

Keywords

Collaborative multi-regression models, Analyzing student behavior, Predicting student performance

1. INTRODUCTION

The problem of identifying students that are at risk of failing a course in order to allow for taking corrective actions can be addressed through analyzing historical data for students' academic performance that is collected by the various Colleges and Universities. Two general approaches have been developed for solving this problem. The first uses features related to the student's past course performance and interactions with online learning management systems (LMS) to estimate a general predictive model for all students (single regression) [2]. The second uses factorization models, initially developed in the context of recommender systems, to predict students' grades for different course activities. Specifically, multi-relational models were used to predict students' performance by learning student, task and skill latent factors that satisfy student-task and task-skill relations [3] while approaches based on tensor factorization were used to account for the temporal aspect in order to model student's knowledge acquisition over time [5]. Unlike single linear regression, factorization models can achieve better prediction accuracy as their predictions are *personalized* to each student. However, these approaches only use students' past performance and entirely ignore how the students interact with the information provided in the LMS, which can potentially be used to improve the overall prediction accuracy.

In this work we investigate the effectiveness of a class of collaborative multi-regression models for predicting the students' performance at various course activities (e.g., quizzes and assignments). These models are inspired by recommendation approaches [6, 1, 4] and they estimate a small number of linear regression models along with a student-specific linear function to combine them. In this approach the regression models are estimated using the historical information of all students. This allows for cross-student information sharing and thus overcome data sparsity issues while providing accurate modeling of each student's unique characteristics via the user-specific linear combination function. The regression models utilize a wide-range of features including the students' past performance, their interactions with the LMS, as well as course and activity-related information. We experimentally evaluated the performance of these models on a large dataset extracted from the University of Minnesota's Moodle installation [1]. The collaborative multi-regression models were able to achieve an RMSE of 0.147 whereas the RMSE of the corresponding single regression model was 0.177. The features related to viewing of course material and previous student performance showed high contributions to the predicted grades.

An advantage of the collaborative multi-regression model is that by clustering the students based on their combination weights we can segment them into groups whose prediction

[1]http://moodle2.umn.edu

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA

http://dx.doi.org/10.1145/2723576.2723590

models are quite similar. Analyzing these groups gives insights on the factors relating to the students' performance. Towards this end, we analyzed the combination weights for the collaborative multi-regression model consisting of just two regression models and identified three groups of students. The underlying regression models for two of these groups were different from each other in how much they rely on the LMS interaction features. In addition, some of the departments of the courses that these student took showed a high specificity to one of these groups. These results may suggest that the information that is provided in the LMS for certain departments may not be beneficial and as such accessing it does not lead to better understanding and thus grades.

2. MODEL DESCRIPTION

We wish to learn a model that predicts the student grades within the different course activities, like assignments and quizzes, given some input features. We achieve this using a collaborative multi-regression model inspired by [6] and [1]. This approach learns a small number of linear models that capture the performance patterns of the different student groups and thus it has an advantage over learning a different model per student as it makes use of the similarities among the students (with respect to performance) and can better handle data sparsity issues. Moreover, unlike single linear regression, the collaborative multi regression model achieves personalization through student-specific bias terms as well as student-specific membership weights which determine how much each linear model contributes to the grades estimated for each student. It also utilizes course bias terms that capture the grade patterns within the different courses.

In this model, grade $\hat{g}_{s,a}$ of student s in activity a is estimated as

$$\hat{g}_{s,a} = b_s + b_c + \boldsymbol{p}_s^t W \boldsymbol{f}_{sa} = b_s + b_c + \sum_{d=1}^{l} \Big(p_{s,d} \sum_{k=1}^{n_F} f_{sa,k} w_{d,k} \Big),$$

where b_s and b_c are student and course bias terms, respectively, $\boldsymbol{f}_{sa}$ is a vector of length n_F that holds the input features (the predictors), l is the number of linear regression models, W is a matrix of dimensions $l \times n_F$ that holds the coefficients of the l linear regression models, and $\boldsymbol{p}_s$ is a vector of length l that holds the memberships of student s within the l different regression models. The term $w_{d,k}$ represents the weight of feature k under the d^{th} regression model, whereas the term $p_{s,d}$ represents the membership of student s in the d^{th} regression model; that is, how much the d^{th} regression model contributes to grade estimations for student s. Throughout the rest of the paper, we will refer to the model parameters as the bias terms, the regression models' feature weights (referred to by the vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_l$) and the students' memberships within the different regression models (referred to as $\boldsymbol{p}_s$ for each student s).

The parameters of the model are estimated by solving a minimization process of the form

$$\underset{(W,P,B)}{\text{minimize}} \;\; \mathcal{L}(W,P,B) + \lambda(||P||_F^2 + ||W||_F^2), \qquad (1)$$

where $\mathcal{L}(.)$ is the Root Mean Squared Error loss function and W, P and B are the feature weights, students memberships and bias terms, respectively. The term $\lambda(||P||_F^2 + ||W||_F^2)$ controls the magnitude of the feature weights and the student memberships and thus prevents over-fitting. The scalar λ is fine-tuned at estimating the model parameters.

In addition to accurately predicting students performance, the collaborative multi-regression model can be used to analyze how the different features contribute to the predicted grades. For proper analysis of the estimated model parameters, it is more convenient that all the parameters have non-negative values so that they additively contribute to the predicted grades. This is achieved by adding these constraints to the optimization problem:

$$w_{d,c} \geq 0,\ p_{s,d} \geq 0,\ b_c \geq 0,\ b_s \geq 0,\ \forall s,\ \forall c,\ \forall d. \qquad (2)$$

The optimization problems are solved using stochastic coordinate descent.

3. DATASET AND EVALUATION

The dataset was extracted from the University of Minnesota's Moodle installation; which is one of the largest Moodle installations world wide. The dataset spans two semesters including 11,556 students, 832 courses belonging to 157 departments. Each student has registered in at least 4 courses. There is a total of 114,498 assignment submissions, 75,143 quiz submissions, and 251,348 forum posts. Assignments and quizzes are referred to as *activities*. Activity grades are normalized to be in the range $[0, 1]$.

3.1 Feature Description

Each student-activity pair (s, a) is associated with a feature vector $\boldsymbol{f}_{sa}$. Features fall into three categories:

1) Student performance-specific features:
– *cumGPA*: the GPA accumulated over the courses previously taken by the student.
– *cumGrade*: the average grade achieved over all of the pervious activities in the course. For the first activity in the course, cumGrade is set to the cumGPA.

2) Activity and course-specific features:
– *activity type:* is either quiz or assignment. We define one indicator variable for each activity type.
– *course level:* describes the course's difficulty and takes an integer value of 1, 2, 3 or 4 with 4 being the most difficult.
– *department:* the department to which the course belongs. We define one indicator feature per department.

3) Moodle interaction features: These were extracted from Moodle's log files and they describe the student's interaction with Moodle prior to the activity's due date:
– *n-init-disc*: number of discussions initiated by the student.
– *n-engaged-disc*: number of times that the student has posted to an open discussion.
– *n-read-posts*: number of forum discussions that has been read by the student.
– *n-viewed-mater*: The number of times the student has viewed some course material.
– *n-add-contrib*: number of times the student has contributed to the course by adding something to the course page (e.g., a wiki-page).
– *n-other-accesses*: number of times the student has made any other kind of access to the course pages. This feature is concerned with the student's interaction with the other Moodle modules (e.g., surveys).

For each of the six Moodle interaction features, we created five different features that measured the specified interaction

at various time intervals prior to the activity due date. Four of them measure the interaction at $[0,1)$, $[1,2)$, $[2,4)$, and $[4,7)$ days prior to the due date, whereas the fifth measures the interaction up to the due date of the previous assignment. These features will be denoted by appending "-x" to the feature name, where x is the intervals' upper bound (e.g., *n-init-disc-1*, *n-init-disc-2*, *n-init-disc-4*, and *n-init-disc-7*), and the fifth will be denoted without the "-x" suffix. Note that for the forum interaction features, the collected numbers were normalized with respect to the total number of available forum discussions.

3.2 Evaluation

The dataset was randomly split into 80%-20% train-test subsets. The model was trained on the training set and evaluated on the test set. This process was repeated 4 times and the obtained results on the test set were averaged and reported. The model is evaluated in terms of the root mean squared error (RMSE) between the actual and predicted grades of the test set.

3.3 Baseline Approach

We compare the collaborative multi-regression model against a linear regression model that estimates the student grades as

$$\hat{g}_{sa} = w_0 + \sum_{k=1}^{n_F} w_k f_k, \tag{3}$$

where f_k is the value of feature k and the w_k's are the regression coefficients of the linear regression model. Note that this is different from a collaborative multi-regression model with one linear model since the latter estimates the student grade as $\hat{g}_{sa} = b_s + b_c + m_{s,1} \sum_{k=1}^{n_F} w_{1,k} f_k$, where $m_{s,1}$ is the student-specific membership term.

4. RESULTS AND ANALYSIS

The results and analysis are presented in three parts that address three questions: (1) how the collaborative multi-regression model performs given different features and how it performs against a single linear regression model, (2) how the different bias terms affect the performance of the collaborative multi-regression model, and (3) whether analyzing the estimated model parameters give insights about the different students.

4.1 Collaborative Multi-Regression Prediction Accuracy

Figure 1 shows the RMSE achieved by collaborative multi-regression model that was trained using different feature combinations [2]. These results shows that the prediction accuracy improves as the number of regression models increases. A larger number of linear models with student-specific memberships allow the models to capture relations among the features that better describe different subsets of students. Using ten regression models, the obtained RMSE falls to 0.145. The results also show that the Moodle interaction features do provide predictive signals about students' performance. Comparing the performance of the model that only uses student and activity features against the model that uses student, activity and Moodle features, it is obvious that the incremental gains achieved by the model that does not use the Moodle features saturate faster than the incremental gains achieved by the other model. We believe this is because the model that uses the Moodle features have more information to learn from as the number of regression models increases.

The baseline linear regression model described by Equation 3 gives an RMSE of 0.223 when trained using all features. This is worse than a collaborative multi-regression model with one linear model trained using all features which gives an RMSE of 0.168 as shown in Figure 1. This is due to the student-specific membership and bias terms which enable the collaborative multi-regression model to better capture individual student performances. Moreover, the course-specific bias terms can capture the grade distribution within the different courses.

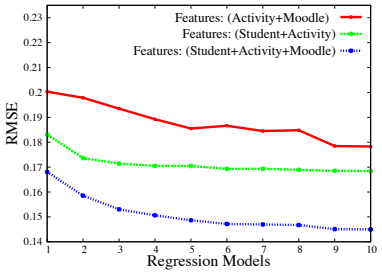

Figure 1: Change in RMSE as the number of regression models increases.

4.2 Effect of the Bias Terms

In order to understand how the different bias terms contribute to the prediction accuracy, we trained the collaborative multi-regression model using each of the student and course bias terms separately. Figure 2 shows the performance of collaborative multi-regression model that uses different bias terms and different number of linear models. The course bias contributes more than the student bias to the prediction accuracy. We believe this is because the contributions of the student bias can be captured by the membership weights.

4.3 Analyzing Feature Weights

To analyze the contributions of the different features, we learn the collaborative multi-regression model while applying the non-negativity constraints of Equation 2 which returns non negative model parameters. These non-negativity

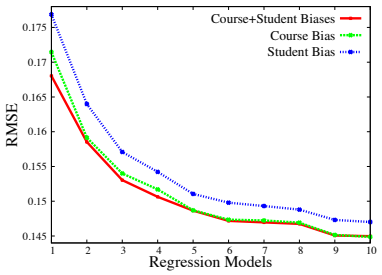

Figure 2: Effect of the different bias terms on the model performance.

[2]These results were generated by learning the model without the non-negativity constrains according to Equation 1.

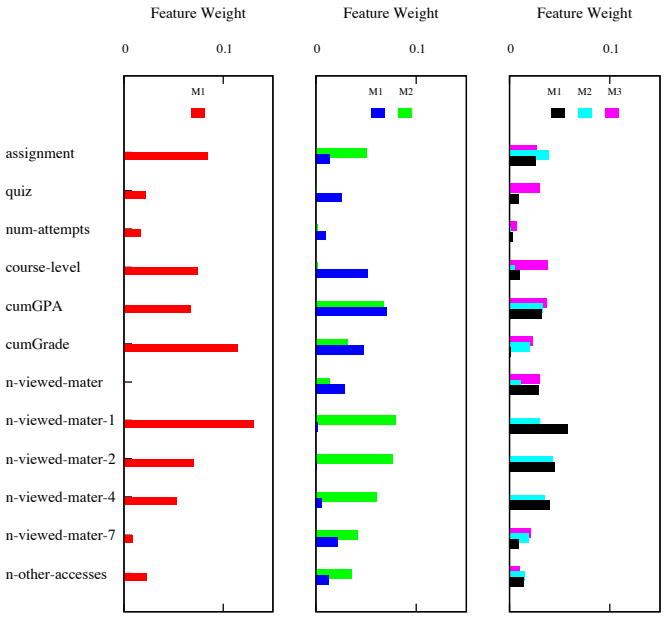

Figure 3: Feature weights for learning a collaborative multi-regression model with one (left), two (center) and three linear models (right).

constrains did not degrade the RMSE in a significant way. For the case of two linear models, the RMSE obtained with and without the non-negativity constrains are 0.165 and 0.160, respectively.

4.3.1 *Determining Importance of Model Parameters*

We estimate for each feature weight $w_{d,k}$ how much it contributes to all the predicted grades. Given a grade $g_{s,a}$, and according to Equation 1, the weight $w_{d,k}$ contributes to $\hat{g}_{s,a}$ by $(m_{s,l} w_{d,k} f_{sa,k})$. Accordingly, the importance $i_{d,k}$ of the feature weight $w_{d,k}$ is accumulated using all the predicted grades as $i_{d,k} = \frac{\sum_{g_{s,a} \in \mathcal{G}} (m_{s,d} w_{d,k} f_{sa,k})/\hat{g}_{s,a}}{|\mathcal{G}|}$, where $\mathcal{G}$ is the set of all grades in the test set and $|\mathcal{G}|$ is the size of $\mathcal{G}$.

Similarly, we estimate the importance of the student and course biases by how much they contribute to all the predicted grades. The student bias importance is estimate as $i_S = \frac{\sum_{g_{s,a} \in \mathcal{G}} b_s/\hat{g}_{s,a}}{|\mathcal{G}|}$, while the course bias importance is estimate as $i_C = \frac{\sum_{g_{s,a} \in \mathcal{G}} b_c/\hat{g}_{s,a}}{|\mathcal{G}|}$.

4.3.2 *Results*

We analyze the estimated feature weights for learning a collaborative multi-regression model with one, two and three linear models. Figure 3 shows the estimated feature weights for one (left), two (center) and three (right) regression models. The binary features representing the departments were omitted as well as the features with zero or very low importance values. The features related to the forum activities had very low importance values as they only appear in a small fraction of the training data (10 to 25% of the training instances).

In all three cases of Figure 3 the features describing viewing of the course material and students previous performance contribute the most to the predicted grades. In the case of two regression models, which we will refer to as M1 and M2, the quiz, number of attempts and course level are important under M1 and not M2. Another interesting point is that the features related to viewing the course material have higher importance under M2. In the case of three regression models, which we will refer to as M1, M2 and M3, the quiz, number of attempts and course level are important under M3 but not under M1 or M2, whereas the assignment and most of the features related to viewing the course material have higher importance under M1 and M2. The fact that we have some models concerned with assignments and others concerned with quizzes reflects that the activity type and properties can impact student performance.

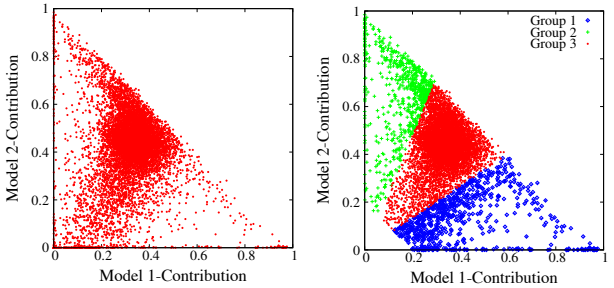

Figure 4: Left: student memberships for a collaborative multi-regression model with two linear models. Every point represents a student. Right: Students are divided into three groups based on their memberships.

4.4 Analyzing Student Memberships

Analyzing student memberships can give insights about the different student populations. We focus on the case in which we learn two linear models since this case is easy to visualize. Moreover, we have seen from Figure 3 that the features related to viewing the course material are more important under one of the two models. This indicates that viewing course material does not similarly impact all students.

4.4.1 *Models' Contributions to Student Grades*

Given a collaborative multi-regression model with two linear models M1 and M2, we estimate for each student s how much M1 and M2 contributes to the grades predicted for s. Given a grade $g_{s,a}$, then according to Equation 1 model d, where $d \in \{1, 2\}$, contributes to $\hat{g}_{s,a}$ by $(m_{s,d} \sum_{k=1}^{n_F} w_{d,k} f_{sa,k})$. Accordingly, the contribution of model d to the grades of student s is estimated as

$$j_{s,d} = \frac{\sum_{g_{s,a} \in \mathcal{G}_s} m_{s,d} \sum_{k=1}^{n_F} w_{d,k} f_{sa,k}/\hat{g}_{s,a}}{|\mathcal{G}_s|},$$

where $\mathcal{G}_s$ is the set of all grades of student s, and $|\mathcal{G}_s|$ is the size of $\mathcal{G}_s$. The value $j_{s,d}$ lies in the range $[0, 1]$, where $j_{s,d} = 0$ means model d does not contribute at all to the grades predicted for s, and $j_{s,d} = 1$ means that the grades of s are only estimated via model d. When we only have two models, then $0 \leq (j_{s,1} + j_{s,2}) \leq 1$.

We have plotted $j_{s,1}$ against $j_{s,2}$ for each student s as shown in Figure 4-left. Each point corresponds to a student, and the x- and y-axis represent $j_{s,1}$ and $j_{s,2}$, respectively. Some students have almost the same contributions by the two models, whereas other students have a higher contribution by one of the two models. Since the two models differ in how much viewing course material influences the predicted grades, this can indicate that students with high contribu-

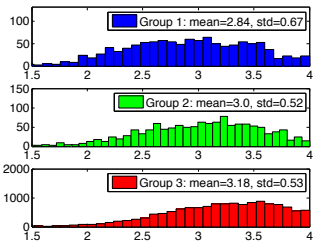

Figure 5: GPA distributions of the student groups.

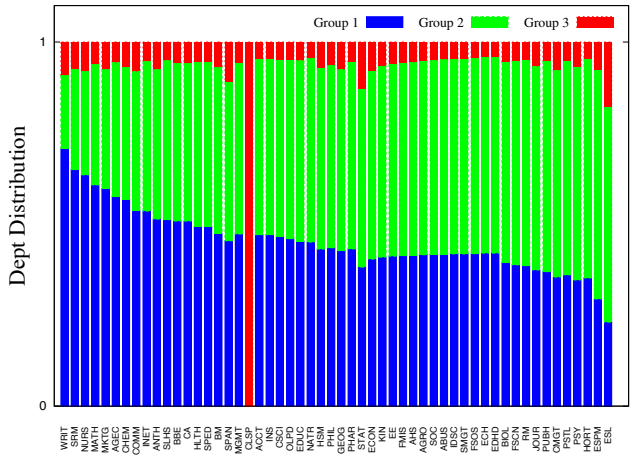

Figure 6: Relative appearance for the different departments within the three student groups.

tion by one model can be different from students with high contribution by the other model.

In order to explore for student differences, we divided the students into three different groups based on their $(j_{s,1}, j_{s,2})$ values using the following procedure: First, we normalized $j_{s,1}$ and $j_{s,2}$ for each student to have a one-norm of 1. Second, we estimated the mean and standard deviation of $j_{s,1}$ for all students, and we got $(\mu, \sigma) = (0.453, 0.142)$. Finally, we clustered the students into three different groups as follows:

– Group 1: all students with $j_{s,1} > \mu + \sigma$.
– Group 2: all students with $j_{s,1} < \mu - \sigma$.
– Group 3: all students with $\mu - \sigma \leq j_{s,1} \leq \mu + \sigma$.

Student Groups 1, 2 and 3 are shown in Figure 4-right using different colors (blue, green and red). We have omitted all students with $(j_{s,1} + j_{s,2}) < 0.2$, that is, students with both models contributing to their grades by less than 0.2. Groups 1, 2 and 3 contain 847, 797 and 7824 students, respectively. The GPA distributions of the three student groups are shown in Figure 5. Group 1 has a lower GPA average and a higher GPA standard deviation than the other two groups.

We have investigated the appearance of the thee student groups within the different departments. For each (department, group) pair (d, g), we compute the group appearance as $q_{d,g} = \frac{n_{d,g}/n_g}{\sum_{h=1}^{3} n_{d,h}/n_h}$, where $n_{d,g}$ is the number of students belonging to group g and are enrolled in courses belonging to department d, and n_g is the number of students belonging to group g. The $q_{d,g}$ metric can be interpreted as the probability of group g given department d. Figure 6 shows how the three student groups tend to appear within each department. Each department is represented by a vertical line. That vertical line has three parts that correspond to $q_{d,G1}$, $q_{d,G2}$ and $q_{d,G3}$, that is the appearance of student groups 1, 2 and 3, within that department. An interesting finding is that some departments are primarily dominated by students of group 1 (the departments towards the left of the figure like Writing, Nursing and Math). These departments tend to have students whose access to course material is less influencing in the sense that it does not additively contribute to their predicted grades. This may suggest that the information provided in the LMS for the departments that intensively appear with group 1 may not be addressing the right student needs and therefore are not beneficial in improving the grades of the students. Accordingly, students' access to such non-beneficial material does not lead to better understanding and thus does not lead to better grades.

5. CONCLUSIONS

In this work, we have used a collaborative multi-regression model to predict student performance in course activities and analyze the resulting student populations. Our results showed that a collaborative multi-regression model performs better than single linear regression as it captures personal student differences, and using the Moodle interaction features lead to better prediction accuracy. Analyzing the estimated model parameters showed that the features related to viewing the course material and previous performance highly contribute to the predicted grades. Moreover, the analysis of the different student populations showed that the features relating to viewing course material contribute to the predictions of a certain student subpopulation higher than other students and that some departments tend to have students whose viewing of course material does not contribute much to their predicted grades.

6. REFERENCES

[1] D. Agarwal and B.-C. Chen. Regression-based latent factor models. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD, 2009.

[2] R. Barber and M. Sharkey. Course correction: Using analytics to predict course success. In *Conference on Learning Analytics and Knowledge*, LAK, 2012.

[3] T. H. Nguyen Thai-Nghe, Lucas Drumond and L. Schmidt-Thieme. Multi-relational factorization models for predicting student performance. In *KDDinED Workshop*, KDD. ACM, 2011.

[4] S. Rendle. Factorization machines with libFM. *ACM Trans. Intell. Syst. Technol.*, 3(3), May 2012.

[5] N. Thai-Nghe, T. Horváth, and L. Schmidt-Thieme. Factorization models for forecasting student performance. In *International Conference on Educational Data Mining*, pages 11–20, 2011.

[6] L. Zhang and Y. Zhang. Discriminative factored prior models for personalized content-based recommendation. In *ACM International Conference on Information and Knowledge Management*, CIKM, 2010.