Enriching Course-Specific Regression Models with Content Features for Grade Prediction

Qian Hu
Department of Computer Science
George Mason University
Fairfax, VA
qhu3@gmu.edu

Huzefa Rangwala
Department of Computer Science
George Mason University
Fairfax, VA
rangwala@cs.gmu.edu

ABSTRACT

An enduring issue in higher education is student retention and timely graduation. Early-warning and degree planning systems have been identified as a key approach to tackle this problem. Accurately predicting a student's performance can help recommend degree pathways for students and identify students at-risk of dropping from their program of study. Various approaches have been developed for predicting students' next-term grades. Recently, course-specific approaches based on linear regression and matrix factorization have been proposed, which achieved better performance than existing approaches based on traditional methods. To predict a student's grade, course-specific approaches utilize the student's grades from courses taken prior to that course. However, there are a lot of factors other than student's historical grades that influence his/her performance, such as the difficulty of the courses, the quality and teaching style of the instructor, the academic level of the students when taking the courses and so on. In addition to that, course-specific models show poor performance if the program has flexible degree plans i.e., several electives. In this paper, we propose a course-specific regression model enriched with features about students, courses and instructors. Our proposed models were evaluated on a dataset from a public university for departments with varying flexibility. The experimental results showed that incorporating content features can boost the performance of the course-specific model. For some degree programs with high flexibility, our experiments showed that predicting the grades with only content features can give better results.

KEYWORDS

Educational data mining, Grade Prediction

ACM Reference format:
Qian Hu and Huzefa Rangwala. 2017. Enriching Course-Specific Regression Models with Content Features for Grade Prediction. In *Proceedings of ACM SIGKDD, Halifax, Nova Scotia, Canada, August 13-17 2017 (KDD'17),* 7 pages.
https://doi.org/10.475/123_4

KDD'17, August 13-17 2017, Halifax, Nova Scotia, Canada

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

The past few years have seen the rise of technologies that capture and leverage massive quantities of education-related data to deliver and improve all levels of learning and education in our society. The Department of Education Report [4] specifically highlighted the current successes of learning analytics and critical need for further research focused on development of robust applications that lead to better student outcomes, improved instructor pedagogy, enhanced curriculum and higher graduation rates for all students irrespective of their backgrounds from kindergarten through college. Currently, higher education institutions face a critical challenge of retaining students and ensuring their successful graduation [16]. Towards this end, several universities seek to deploy accurate and effective *degree planners* that assist students in choosing academic pathways towards a successful and timely graduation; and *early-warning systems* that aid academic advisors in identifying students who are at the risk of failing or dropping out of a program for timely intervention. In this paper we present solutions that analyze in a systematic and careful manner, the large and diverse type of education-related data collected at George Mason University with the objective of assisting students to make informed decisions about their future course selections. Specifically, we develop methods that perform next-term grade prediction i.e., predict the grade for students in future courses that they have not taken yet.

In this work, based on course-specific models we proposed a model which not only uses the grades of prior courses but also different kinds of content features. The course-specific models have been applied to predict student's next-term grades by using grades of prior courses, which better addresses problems associated with the reliable estimation of the low-rank models [14]. However, course-specific models that use the grades of prior courses can only capture the information of student's knowledge evolution. There are some other factors that can influence student's grades, such as his/her academic level when taking a certain course, instructor's teaching quality and courses' difficulty. In addition, course-specific models also suffer from inaccurate prediction if the degree program is flexible (i.e., has several electives). To solve this problem we incorporated content features, which can capture diverse information about students, courses and instructors.

We evaluated our proposed method on a dataset from George Mason University collected from Fall 2009 to Spring 2016. The results showed that our proposed method outperformed competing methods to some degree. Another conclusion was that when the prior-course information was sparse, the included content features were more likely to help.

The paper is organized as follows. Section 2 investigates the related work in the area of student's performance prediction. Section 3 describes the notations we used in the paper. Section 4 discusses our proposed method and other comparison methods. Section 5 is about protocol. In Section 6, we presents our experimental results and analysis. And the last section gives some conclusions and future direction.

2 RELATED WORK

In recent years, data mining and machine learning techniques have been applied to improve educational quality including areas related to learning and content analytics [9, 10], knowledge tracing [6, 20], learning material enhancement [1] and early warning systems [3, 11]. Based on the scope of this paper, we only review approaches for next-term student grade prediction.

Knowing student's performance in advance can help instructors catch at-risk students early and advise them in choosing appropriate courses that fit their current knowledge state better. As such, several methods have bee developed to tackle the next term prediction problem. Most of the methods are inspired from recommender system literature [17, 18], such as matrix factorization [14] and collaborative filtering [5, 19]. Approaches based on standard classification approaches such as random forests trees have also been applied [2, 18]. A majority of the algorithms proposed are "one-size-fits-all", namely, trying to model all the students with one model. To model students with different characteristics, personalized grade prediction approaches have been proposed [12, 15]. Using features mined from student interaction with learning management systems, Elbadrawy *et. al.* proposed a personalized mutli-regression model [7] for in-class grade prediction.

Recently course-specific models proposed by Polyzou *et. al.* [14] achieved better prediction accuracy than existing approaches, assuming that students acquire knowledge in an cumulative manner. Course-specific models are cumulative, in the sense that to predict a student's grade in a target course, the students' grades from courses taken prior to the target course are utilized.

However, one of the drawbacks of course-specific models is that they show poor performance if the degree program is flexible [14]. In addition, the grades of the prior courses can not completely capture all the factors that affect students' performance. In this paper, based on course-specific models, we proposed a hybrid model to predict students' next-term performance by taking some informative factors into consideration.

3 PROBLEM FORMULATION AND NOTATIONS

Formally, we assume that we have records of n students and m courses, comprising a $n \times m$ sparse grade matrix $\mathbf{G}$, where $g_{s,c} \in [0-4]$ is the grade a student s earned in course c. The objective of next-term grade prediction problem is to estimate the grade $\hat{g}_{s,c}$, a student s will achieve in course c in the next term. Besides the grade matrix $\mathbf{G}$, we have information that can be associated with the student (e.g., academic level, previous GPA, major) and course offering (e.g., discipline, course level, prior courses frequently taken, instructor, etc) that can be combined to extract a feature vector per dyad. We denote this feature vector as $\mathbf{x}$ of p dimensions. As a convention, bold uppercase letters are used to represent matrices (e.g., $\mathbf{X}$) and bold lowercase letters represents vectors (e.g., $\mathbf{x}$).

4 METHODS

4.1 Course-Specific Regression (Prior Courses)

Polyzou et.al. [14] motivate the use of course-specific regression models that leverage the sequential structure of undergraduate degree programs. These regression models assume that the performance of a student in a future course is strongly correlated with past performance on a subset of courses taken earlier. Specifically, this regression model estimates the grades for a future class as a sparse linear combination of grades obtained on prior courses. For a course c the grades that students obtained on courses taken prior to c are extracted from the grade matrix $\mathbf{G}$, and denoted by $\mathbf{G}_c^{pr}$. Each row of this matrix corresponds to students that have taken the course c. Assume that n_c students have taken the course c so far and m_c represents the union set of courses taken by students prior to c, then the dimensions of $\mathbf{G}_c^{pr}$ is $n_c \times m_c$. $\mathbf{g}_{:,c}$ is the vector representing the grades that students obtained for course c. We learn the parameters of this Course-Specific Regression (CSR) model by solving the least square regression problem enforcing $\ell 1$ and $\ell 2$ norms. The optimization problem is given below:

$$minimize \underbrace{||\mathbb{1}w_{c,0} + \mathbf{G}_c^{pr}\mathbf{w}_c^{pr} - \mathbf{g}_{:,c}||_2^2}_{\text{loss}} + \underbrace{\lambda_1||\mathbf{w}_c^{pr}||_2^2}_{\ell 2} + \underbrace{\lambda_2||\mathbf{w}_c^{pr}||_1}_{\ell 1} \tag{1}$$

where $\mathbb{1}$ is a vector of ones of dimension n_c, $\mathbf{w}_c^{pr} \in R^{m_c}$ denotes the weight vectors associated with each course c and $w_{c,0}$ is the bias term. The $\ell 1$ norm ensures sparsity and $\ell 2$ avoids overfitting.

Having learned the weight vectors and bias terms, the grade estimate for a student s enrolling in course c is given by:

$$\hat{g}_{s,c} = w_{c,0} + \mathbf{x}_{s,c}^T \mathbf{w}_c^{pr} \tag{2}$$

where $\mathbf{x}_{s,c} \in \mathbb{R}_c^m$ is a feature vector representing the grades on prior courses that the student has taken so far. We denote this Course-Specific Regression model with Prior Courses as CSR_{PC}.

In this approach, prior to estimating the model using equation 1, we row-centered each row of matrix $\mathbf{G}_c^{pr}$ and $\mathbf{g}_{:,c}$, which is done by subtracting the GPA of corresponding students from the non-zero entries in each row of $\mathbf{G}_c^{pr}$ and $\mathbf{g}_{:,c}$ [14]. We found that row-centering gives better performance by mitigating the negative influence of missing grades from prior courses.

4.2 Course-Specific Regression (Content Features)

The CSR_{PC} model described above is able to provide accurate estimates of student performance provided a course has sufficient number of prior courses. We seek to extract key features associated with students and courses and incorporate them within the prediction formulation. Based on course-specific idea, instead of training one global model for all the courses, we propose to train independent course-specific regression models with content features. We refer to this model by CSR_{CF}. In terms of formulation, the proposed CSR_{CF} is similar to CSR_{PC} except that the feature vector is a composite of student, course and instructor-related features as described in Section 4.2.1. We denote the weight vector learned by

this formulation as $\mathbf{w}_c^f$ and the feature vectors $\mathbf{x}_{s,c} \in \mathbb{R}^p$ where p is the total number of features. The predicted grade estimate is then given by:

$$\hat{g}_{s,c} = w_{c,0} + \mathbf{x}_{s,c}^T \mathbf{w}_c^f \tag{3}$$

The CSR_{CF} model is estimated in a similar manner as CSR_{PC} and given by:

$$minimize \underbrace{||\mathbb{1} w_{c,0} + \mathbf{X}_c^f \mathbf{w}_c^f - \mathbf{g}_{:,c}||_2^2}_{\text{loss}} + \underbrace{\lambda_1 ||\mathbf{w}_c^f||_2^2}_{\ell 2} + \underbrace{\lambda_2 ||\mathbf{w}_c^f||_1}_{\ell 1} \tag{4}$$

where $\mathbf{X}_c^f$ is a matrix of stacked feature vectors from the different students who have taken the course c in the past. Each row of this matrix is a feature vector for a student enrolled in the course c.

4.2.1 Content features. Student-related features include their demographic data, such as their age, race, gender, high school GPA and so on. For each term, we have the GPA of the student from the previous term and the accumulative GPA as of last term. As students might take courses from other departments which has less influence than those from their own departments, we can extract GPA of courses only from their own departments. When taking a course, different students might come from different academic level, therefore, it might be beneficial to incorporating their academic level into the model.

The features relating to a course include its discipline, the credit hours it's worth of and its course level (e.g. 100, 200, 300, 400-level). As the difficulty of a course can influence the performance of the students, it's helpful to include the course difficulty information into the model. We use the GPA of the course from last term to represent the difficulty of the course.

As the factors from instructors can also influence the performance of the students, we extract content features about the instructors which include his rank, tenure status and the GPA of the courses he has taught.

We one-hot-encoded categorical features in $\mathbf{X}_c^f$ and standardized the continuous features.

4.3 Hybrid Model

We also combine the feature vectors $\mathbf{X}_c^f$ and $\mathbf{G}_c^{pr}$ obtained from the student-course content and prior grades and learn weight vectors per course, respectively. We refer to this hybrid model as CSR_{HY} and learn a course-specific regression model as discussed above.

4.4 Baseline Methods

In the experiments, we compare the proposed methods with the following competing approaches.

(1) Matrix Factorization (MF): The use of MF for grade prediction is based on the assumption that the students and courses' knowledge space can be jointly represented in low-dimensional latent feature space [14]. Each component in the latent feature space corresponds to knowledge components. The grade of student s in a future course c is estimated as:

$$\hat{g}_{s,c} = b_0 + b_s + b_c + \langle \boldsymbol{p}_s, \boldsymbol{q}_c \rangle \tag{5}$$

where b_0, b_s and b_c are the global bias, student bias and course bias respectively and $\boldsymbol{p}_s$, $\boldsymbol{q}_c$ are the latent vectors representing student s and course c.

(2) Course-specific Matrix Factorization (CS_{MF}): CS_{MF} is similar to MF except that the grade matrix $\mathbf{G}_c$ for CS_{MF} only includes the grades of students taking the course and their grades of courses taken prior to the course we are going to predict [14].

(3) BiasOnly (BO): BiasOnly method only takes into consideration student's bias, course's bias and global bias which are estimated using Equation 5 by setting the dimension of the latent factors as 0 [14].

Table 1: Information about the different majors

Major	#Students	#Courses	#Grades	Flexibility
CS	988	53	21,880	0.283
ECE	396	69	161,70	0.272
BIOL	1629	105	20,602	0.339
PSYC	1114	60	14,851	0.429

5 EXPERIMENTAL PROTOCOL

5.1 Dataset description and preprocessing

We evaluated our proposed methods on dataset obtained from four departments: (i) Computer Science (CS), (ii) Electrical and Computer Engineering (ECE), (iii) Biology (BIOL) and Psychology (PSYCH) at George Mason University. The data was collected from Fall 2009 to Spring 2016. According to the University Catalog [8], we kept the courses that were required by the degree program and electives within the same major. The statistics of the four majors are shown in Table 1.

We removed any courses whose grades were pass/fail. If a course was taken more than once by a student, only the last grade was kept. To form the test and training dataset, we use the data extracted from last term (i.e., Spring 2016) as test dataset and data from all the terms before Spring 2016 as training. The training dataset was split into 80/20, of which 80% was training data, 20% was validation data.

As the flexibility of a degree program can influence the course-specific models' performance, a flexible parameter associated with each department is computed according to [13]. The major's flexibility is the average course flexibility over all courses belonging to that major. We computed the flexibility of a course as one minus the average Jaccard coefficient of the courses that were taken by the students that took c prior to taking this course. The flexibility of a course will be low if the students have taken very similar prior courses and high otherwise.

5.2 Evaluation Metrics

To assess the performance of the models, we used three kinds of metrics, namely mean absolute error (MAE), root mean squared error (RMSE) and tick error. MAE and RMSE are computed by pooling together all the grades across all the courses. As each course has different number of students, we also computed the average MAE and RMSE denoted as AvgMAE and AvgRMSE, respectively,

Table 2: AverageMAE and MAE of different methods

Method	AvgMAE				MAE			
	CS	ECE	BIOL	PSYC	CS	ECE	BIOL	PSYC
BO	0.7551	0.6965	0.5755	0.5390	0.7359	0.7285	0.5853	0.5882
MF	0.7866	0.8346	0.6112	0.5182	0.8150	0.8447	0.6169	0.5648
CS_{MF}	0.7647	0.6936	0.5355	0.4961	0.7609	0.7015	0.5579	0.5240
CSR_{PC}	0.6870	0.6451	0.5421	0.4928	0.6805	0.6739	0.5372	0.4933
CSR_{CF}	0.6929	0.6599	**0.4673**	**0.4836**	0.7183	0.6775	**0.4769**	**0.4743**
CSR_{HY}	**0.6606**	**0.6289**	0.4987	0.4858	**0.6693**	**0.6630**	0.5057	0.4859

Table 3: AverageRMSE and RMSE of different methods

Method	AvgRMSE				RMSE			
	CS	ECE	BIOL	PSYC	CS	ECE	BIOL	PSYC
BO	0.9443	0.8911	0.7372	0.7148	0.9622	0.9748	0.7794	0.7829
MF	1.0228	1.0296	0.7849	0.6998	1.0879	1.1104	0.8173	0.8035
CS_{MF}	0.9921	0.8738	0.7345	0.6840	1.0126	0.9623	0.8045	0.7372
CSR_{PC}	0.8982	0.8570	0.7488	0.7092	0.9288	0.9699	0.7943	0.7348
CSR_{CF}	0.8943	0.8470	**0.6464**	**0.6588**	0.9539	0.9680	**0.7205**	**0.6732**
CSR_{HY}	**0.8773**	**0.8380**	0.7200	0.7058	**0.9199**	**0.9542**	0.7679	0.7283

Table 4: Prediction performance of different methods based on Ticks

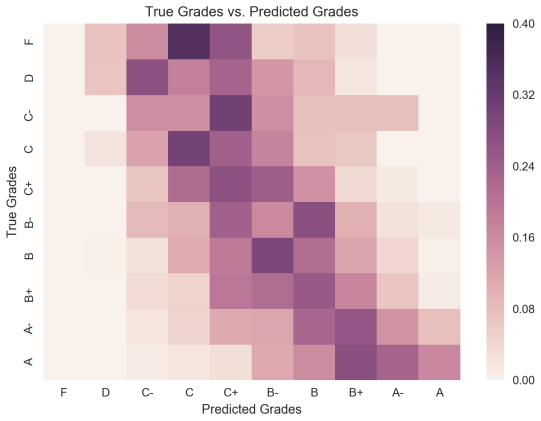
#Ticks	Method	CS	ECE	BIOL	PSYC
Percentage of Grades predicted with no error	BO	15.02	18.58	19.41	19.75
	MF	13.04	9.84	19.95	23.89
	CS_{MF}	15.22	18.58	24.53	23.25
	CSR_{PC}	19.57	20.77	28.84	34.08
	CSR_{CF}	13.44	16.39	28.03	27.39
	CSR_{HY}	**19.76**	**22.40**	**30.73**	**35.35**
Percentage of grades predicted with an error of at most one tick	BO	44.27	44.26	55.26	53.82
	MF	42.29	39.34	51.75	54.46
	CS_{MF}	43.08	40.44	58.76	61.78
	CSR_{PC}	48.22	**55.19**	62.80	61.15
	CSR_{CF}	44.66	51.37	**70.89**	**64.97**
	CSR_{HY}	**49.80**	**55.19**	67.38	61.78
Percentage of grades predicted with an error of at most two ticks	BO	69.17	66.67	77.63	75.80
	MF	64.82	63.38	76.82	77.07
	CS_{MF}	67.59	72.68	82.21	78.66
	CSR_{PC}	74.31	73.22	81.40	79.62
	CSR_{CF}	73.52	**75.96**	**87.87**	**83.44**
	CSR_{HY}	**75.10**	74.32	82.75	78.66

which are computed by averaging the MAE and RMSE for each course.

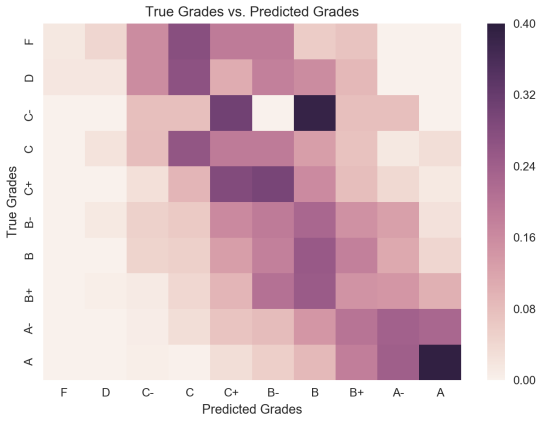
MAE, RMSE, AvgMAE and AvgRMSE are all averaged errors between the predicted grades and the actual grades. To gain a better insight into the quality of the predictions, we also report the tick error as done in [13, 14]. The grading system of George Mason University has 11 letter grades (A+, A, A-, B+, B, B-, C+, C, C-, D, F) which correspond to (4, 4, 3.67, 3.33, 3, 2.67, 2.33, 2, 1.67, 1, 0). We refer to the difference between two successive letter grades as a tick. The performance of a model is assessed based on how many ticks away the predicted grade is from the actual grade. We first converted the predicted grades into their closest letter grades and then computed the percentages of each of the x ticks [13, 14].

6 RESULTS AND DISCUSSION

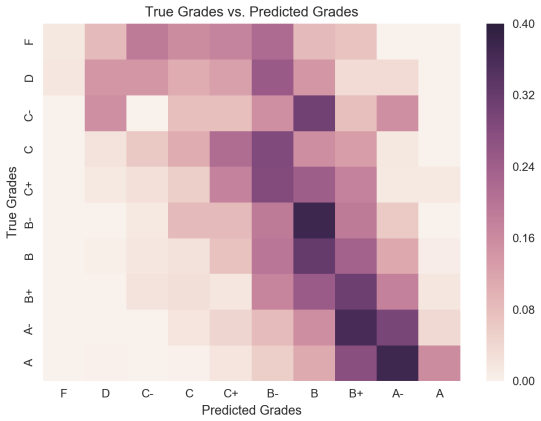
Table 2 and 3 shows the comparative performance of different methods on four different departments by using metrics average MAE, MAE and average RMSE, RMSE. Generally, all course-specific models outperform non-course-specific models, which means focusing on a course-specific subset of data can result in better performance. For departments with less flexibility such as Computer Science and Electrical Engineering, we observe that the hybrid model has the



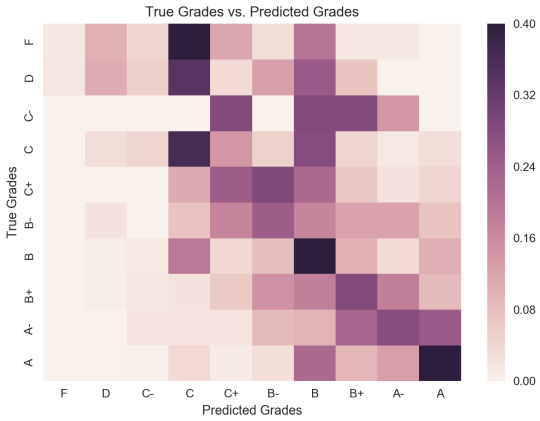
(a) True vs. Predicted Grades for BO



(b) True vs. Predicted Grades for CSR_{PC}



(c) True vs. Predicted Grades for CSR_{CF}



(d) True vs. Predicted Grades for CSR_{HY}

Figure 1: True vs. Predicted Grades for BiasOnly and Course-specific Models

best performance. Thus incorporating content features into course-specific model further improves its performance. The model with only grades of prior courses performs better than model with only content features. For departments with high flexibility such as Biology and Psychology, the model with only content features shows the best performance, which suggests that if a department has a flexible degree program, content features might be more informative than the grades of prior courses.

To gain deeper insights into the types of errors made by different methods, Table 4 reports the percentage of grades predicted with no error, with an error of at most one tick and with an error of at most two ticks. Comparing the performance achieved by the methods we notice that the course-specific models have relatively better performance than non-specific approaches. In terms of the exact prediction (i.e., no error), the hybrid model has the best performance. For departments with rigid degree program, such as Computer Science and Electrical Engineering, the hybrid model has better performance than other methods. If minor errors are allowed (i.e., one or two ticks), for flexible departments, model with only content features gives better performance.

The distribution of true (ground truth) and predicted grades are also plotted in Figures 1a, 1b, 1c and 1d for BiasOnly, CSR_{PC}, CSR_{CF} and CSR_{HY}, respectively. Each row of the figure represents the ratio of the predicted grades. For example, in Figure 1b the bottom row represents that a high proportion of A's are predicted as such. We see that BiasOnly tends to smooth the predicted grades i.e., it predict most of the grades around the average GPA (around B-). However, for high grades most of the predicted grades are around the true grades in course-specific models and for lower grades all the models tend to over predict.

Table 5 shows the detailed statistics of the courses from two departments CS and PSYC with strict and flexible degree program, respectively, and the errors (RMSE) made by three course-specific regression models. From Table 5, we can see that if the grades in test set have high standard deviation or higher than that of training set,

Table 5: Per course statistics and errors

Course	#training	#testing	density	Mn Tr	StD Tr	Mn Te	StD Te	CSR_{PC}	CSR_{CF}	CSR_{HY}
CS-2xx	322	76	0.766	2.640	1.249	2.548	1.455	1.179	1.226	1.176
CS-2xx	303	66	0.623	2.915	1.062	2.899	0.941	0.686	0.755	0.735
CS-3xx	138	19	0.748	3.049	0.803	3.158	0.597	0.463	0.417	0.434
CS-3xx	285	62	0.638	2.634	1.155	2.694	1.236	1.037	1.156	1.037
CS-3xx	181	41	0.711	3.063	0.779	3.041	0.617	0.527	0.465	0.539
CS-3xx	42	13	0.802	3.104	1.140	3.360	0.591	0.748	0.668	0.752
CS-3xx	189	35	0.754	2.783	1.032	2.657	1.053	0.876	0.949	0.876
CS-3xx	19	8	0.885	2.719	1.072	2.959	1.368	1.152	1.035	1.253
CS-3xx	156	29	0.768	3.088	0.762	2.897	1.175	1.072	1.045	1.066
CS-4xx	92	8	0.867	2.859	1.103	2.917	1.090	1.006	1.119	1.006
CS-4xx	29	15	0.868	2.426	1.181	2.311	1.341	1.243	0.972	0.830
CS-4xx	35	7	0.378	2.667	0.983	2.713	0.629	0.711	0.609	0.736
CS-4xx	105	36	0.909	3.137	0.810	3.297	0.965	0.951	0.913	0.994
CS-4xx	43	10	0.912	2.923	1.001	2.567	1.383	1.072	1.063	1.042
CS-4xx	46	19	0.896	2.725	1.111	1.983	1.111	1.090	1.081	1.143
CS-4xx	32	8	0.897	3.083	0.866	3.041	1.207	0.964	1.106	0.964
CS-4xx	115	32	0.868	3.018	0.914	3.229	0.659	0.655	0.643	0.655
CS-4xx	26	22	0.868	3.525	0.668	3.333	0.841	0.669	0.870	0.610
PSYC-2xx	195	24	0.608	3.165	0.802	3.639	0.429	0.709	0.604	0.694
PSYC-2xx	204	23	0.635	3.144	0.726	3.435	0.788	0.678	0.746	0.678
PSYC-3xx	247	23	0.670	3.263	0.813	3.580	0.654	0.796	0.656	0.799
PSYC-3xx	223	24	0.724	3.262	0.870	3.390	0.875	0.759	0.578	0.756
PSYC-3xx	44	5	0.825	3.212	0.943	3.600	0.490	0.507	0.829	0.653
PSYC-3xx	112	8	0.613	3.310	0.858	3.292	0.715	0.878	0.726	0.873
PSYC-3xx	86	7	0.558	3.535	0.758	3.620	0.516	0.696	0.467	0.678
PSYC-3xx	258	21	0.586	3.263	0.936	3.778	0.428	0.760	0.801	0.728
PSYC-3xx	69	14	0.718	3.251	0.667	3.357	0.672	0.481	0.475	0.467
PSYC-3xx	227	26	0.687	3.333	0.728	3.270	0.883	0.776	0.729	0.776
PSYC-3xx	94	9	0.723	3.394	0.617	3.630	0.618	0.600	0.521	0.602
PSYC-3xx	52	6	0.714	3.378	0.911	3.280	1.027	0.978	1.033	0.956
PSYC-3xx	216	22	0.731	3.048	0.951	2.803	1.013	0.940	0.642	0.940
PSYC-3xx	66	12	0.710	3.525	0.802	3.168	0.977	1.020	0.865	1.021
PSYC-3xx	121	18	0.715	3.488	0.705	3.371	0.745	0.692	0.627	0.700
PSYC-4xx	182	21	0.672	3.564	0.716	3.540	0.442	0.549	0.346	0.550
PSYC-4xx	48	5	0.789	3.771	0.409	4.000	0.000	0.424	0.253	0.424
PSYC-4xx	105	30	0.661	3.445	0.884	3.778	0.489	0.588	0.627	0.590
PSYC-4xx	34	12	0.798	3.657	0.521	3.112	0.736	0.809	0.893	0.802

The second and third column stand for the number of training and testing instances, respectively
density means the density of the prior course matrix
Tr train, *Te* test, *Mn* mean, *StD* standard deviation

the prediction error is high. The reason might be that the course-specific models used in this work and previous works are linear. In the future, we would like to explore non-linear course-specific models. We also noticed that the content features don't improve the performance on some courses. The reason might be that the flexibility of these courses are low so that the content features don't help much for these courses and the grades of the prior courses can reflect students' knowledge evolution well.

7 CONCLUSIONS

In this paper, we proposed a hybrid model to further improve the performance of the course-specific models. The performance of course-specific models are greatly influenced by the flexibility of the degree program. For departments with less flexible degree program, the hybrid model achieves better performance than traditional course-specific models. However, for departments with more flexible degree program, the grades of prior courses are less informative than content features, therefore, it is more appropriate to include only content features.

REFERENCES

[1] Rakesh Agrawal, Maria Christoforaki, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi, and Adith Swaminathan. 2014. Mining videos from the web for electronic textbooks. In *International Conference on Formal Concept Analysis*. Springer, 219–234.

[2] Mashael A Al-Barrak and Muna Al-Razgan. 2016. Predicting students final GPA using decision trees: a case study. *International Journal of Information and Education Technology* 6, 7 (2016), 528.

[3] Hall P Beck and William D Davidson. 2001. Establishing an early warning system: Predicting low grades in college students from survey of academic orientations scores. *Research in Higher education* 42, 6 (2001), 709–723.

[4] Marie Bienkowski, Mingyu Feng, and Barbara Means. 2012. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *US Department of Education, Office of Educational Technology* (2012), 1–57.

[5] Hana Bydžovská. 2015. Are Collaborative Filtering Methods Suitable for Student Performance Prediction?. In *Portuguese Conference on Artificial Intelligence*. Springer, 425–430.

[6] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[7] Asmaa Elbadrawy, R. Scott Studham, and George Karypis. 2015. Collaborative Multi-Regression Models for Predicting Students' Performance in Course Activities. *LAK, '15* (2015).

[8] GMU. 2017. George Mason University Catalog. (2017). http://catalog.gmu.edu/

[9] Andrew S Lan, Christoph Studer, and Richard G Baraniuk. 2014. Time-varying learning and content analytics via sparse factor analysis. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 452–461.

[10] Andrew S Lan, Andrew E Waters, Christoph Studer, and Richard G Baraniuk. 2014. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research* 15, 1 (2014), 1959–2008.

[11] Leah P Macfadyen and Shane Dawson. 2010. Mining LMS data to develop an âĂIJearly warning systemâĂİ for educators: A proof of concept. *Computers & education* 54, 2 (2010), 588–599.

[12] Yannick Meier, Jie Xu, Onur Atan, and Mihaela van der Schaar. 2015. Personalized Grade Prediction: A Data Mining Approach. In *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, 907–912.

[13] Sara Morsy and George Karypis. 2017. Cumulative Knowledge-based Regression Models for Next-term Grade Prediction. (2017).

[14] Agoritsa Polyzou and George Karypis. 2016. Grade prediction with models specific to students and courses. *International Journal of Data Science and Analytics* (2016), 1–13.

[15] Mark Sheehan and Young Park. 2012. pGPA: a personalized grade prediction tool to aid student success. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 309–310.

[16] Robert Stillwell and Jennifer Sable. 2013. Public School Graduates and Dropouts from the Common Core of Data: School Year 2009-10. First Look (Provisional Data). NCES 2013-309. *National Center for Education Statistics* (2013).

[17] Mack Sweeney, Jaime Lester, and Huzefa Rangwala. 2015. Next-term student grade prediction. In *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, 970–975.

[18] Mack Sweeney, Huzefa Rangwala, Jaime Lester, and Aditya Johri. 2016. Next-Term Student Performance Prediction: A Recommender Systems Approach. *arXiv preprint arXiv:1604.01840* (2016).

[19] A Toscher and Michael Jahrer. 2010. Collaborative filtering applied to educational data mining. *KDD cup* (2010).

[20] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized bayesian knowledge tracing models. In *International Conference on Artificial Intelligence in Education*. Springer, 171–180.