Next-Term Grade Prediction: A Machine Learning Approach

Audrey Tedja Widjaja, Lei Wang, Truong Trong Nghia,
Aldy Gunawan, Ee-Peng Lim
Singapore Management University
80 Stamford Road
Singapore, 178902
{audreyw, lei.wang.2019, ntruongtrong, aldygunawan, eplim}@smu.edu.sg

ABSTRACT

As students progress in their university programs, they have to face many course choices. It is important for them to receive guidance based on not only their interest, but also the "predicted" course performance so as to improve learning experience and optimise academic performance. In this paper, we propose the next-term grade prediction task as a useful course selection guidance. We propose a machine learning framework to predict course grades in a specific program term using the historical student-course data. In this framework, we develop the prediction model using Factorization Machine (FM) and Long Short Term Memory combined with FM (LSTM-FM) that make use of both student and course attributes as well as past student-course grade data. Our experiment results on a real-world data of an autonomous university in Singapore show that both methods yield better prediction accuracy than the baseline methods. Our methods are also robust to handle cold start courses with the average prediction error can be as low as three quarter grade difference from the ground truth.

Keywords

Grade prediction, factorization machine, long short term memory

1. INTRODUCTION

Predicting student grades has recently gained attention as it benefits not only students, but also instructors [3]. Students face many course courses in every new term. They need some guidance based on their "predicted" performance in future courses so as to improve their course selection and overall academic performance. Instructors, on the other hand, can also adjust their course delivery methods to the predicted student grade performance.

We consider a university setting where students are required to choose courses at the beginning of each program term. The predicted grades of the selected courses is then evaluated against the grades received at the end of that term. This task is called the *next-term student grade prediction* and it requires the past student-course grade data to provide useful features to predict grades of courses taken in the following term.

Our next-term student grade prediction task is different from the previous student grade prediction works [2, 3] which focused on predicting grades of a calendar term where students from different admission years are predicted together. Since different program terms are included in the prediction task, it is difficult to train the model to specialize on courses in the specific program term of the students.

In this paper, we develop FM and long short term memory combined with FM (LSTM-FM) models that are trained on student's program terms instead of calendar terms. The proposed models are evaluated on a real-world data collected from an autonomous university in Singapore. We further make use of both static and dynamic student and course attributes to derive features that improve the prediction results. Additionally, our proposed models could perform well on predicting both existing and cold-start courses.

2. PROBLEM FORMULATION

Given a set of students $S = \{s_1, s_2, ..., s_{|S|}\}$, where each student belongs to a certain cohort, denoted by $cohort(s_i)$ (i.e. batch of students admitting to the university in the same year). To graduate from their programs, students must complete $T = \{t_1, t_2, ..., t_{|T|}\}$ program terms and register one or more courses in each program term. Let $C = \{c_1, c_2, \cdots, c_{|C|}\}$ be the set of all courses taken by students from S. We denote the grade obtained by student s_i in course c_j by $g_{i,j} \in \{A+, A, \cdots, F\}$. Our task is then to predict $g_{i,j}$ for every student s_i from a target student cohort S in a target program term t_k for every course students have registered in the program term t_k. We assume that the course grades for earlier program term(s) by the same students are available, and the course grades for students from previous cohorts in the earlier and target terms can be observed.

We define the feature representation of a student-course pair (s_i, c_j) as a feature vector $X_{i,j}$. A prediction model for the above problem is thus a function $F : X \rightarrow Y$ where $Y \in \mathbb{R}^2$. F is learned from a training data (t_k, X^{trg}, Y^{trg}). For each

Audrey Tedja Widjaja, Lei Wang, Nghia Trong Truong, Aldy Gunawan and Ee-Peng Lim "Next-Term Grade Prediction: A Machine Learning Approach" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 700 - 703

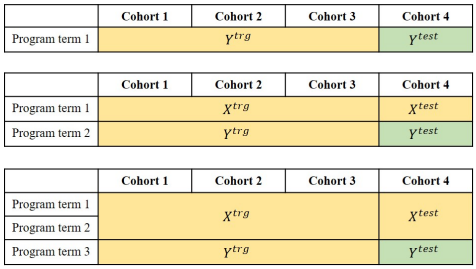

Figure 1: Training and testing instances for program term-specific grade prediction involving data from cohorts 1 to 3 as training, and data from cohort 4 as test.

student $s_i \in S$, $Y^{trg}_{i,j}$ is unknown for courses c_j's registered by s_i during the target program term t_k. For each student s'_i of earlier cohorts, $Y^{trg}_{i,j} = g_{i,j}$ for courses c_j's registered by s_i in the target program term t_k. For all students, $X^{trg}_{i,j}$ are features derived from student s_i and course c_j using data from earlier program terms. The testing data $(t_k, X^{test}, Y^{test})$ consists of $X^{test}_{i,j} = X^{trg}_{i,j}$ and $Y^{test}_{i,j} = g_{i,j}$ when s_i received the grade $g_{i,j}$ in the program term t_k.

Figure 1 illustrates the training and testing instances of the next term grade prediction for students of cohort 4 in target program terms 1 to 3. For target program term 3 (see the last table of the figure), the training data include the student-course data of students from cohorts 1 to 3. The feature representation of a student-course pair is derived from program terms 1 to 2 of these students, or from the non-program term student and course attribute data (e.g., student education background, course major, etc.).

This program term-specific grade prediction approach is more intuitive than previous works that focused on the grade prediction for students taking courses in the same calendar term which could involve different program terms for students from different cohorts [2, 3]. Since student grades of different program terms refer to different sets of courses, our problem definition and solution approach ensure that dyad features and ground truth labels for the testing data of a target program term follow the same data distribution as that of the training data.

3. DATASET AND FEATURES

3.1 Dataset Description

The dataset was collected from an autonomous university in Singapore that covers four consecutive cohorts (2011- 2014) of undergraduate students from the same degree program. Students are required to complete 8 program terms.

Table 1 shows the dataset statistics. It consists of 618 students and 691 courses. In total, we have 19,655 student-course pairs that involve grades, known as the student-course dyads. Students from cohort 4 are used as the test cohort to allow more data to be used in training. The university implements 12 grading letters that are mapped to numeric values for grade prediction as follows. A+, A, A-, B+, B, B-, C+, C, C-, D+, D, and F are mapped to 4.3, 4.0, 3.7, 3.3, 3.0, 2.7, 2.3, 2.0, 1.7, 1.3, 1.0, and 0.0 respectively.

Table 1: Dataset Statistics

	Cohorts				Total
	1	2	3	4	
Num. Students	115	145	157	201	618
Num. Courses	169	160	170	192	691
Num. Dyads	3748	4471	4850	6586	19,655

Table 2: Student-Course Dyads of Target Cohort 4 (CSS: cold start students, CSC: cold start courses, NCS: non-cold start dyads)

Program term	#dyads	#NCS	CS	
			#CSC	#CSS
t_1	986	0	0	986
t_2	955	952	3	0
t_3	856	850	6	0
t_4	919	907	12	0
t_5	801	789	12	0
t_6	704	677	27	0
t_7	699	676	23	0
t_8	666	638	28	0

Cold start dyads. The cold start student-course dyads of a target program term are ones with new students or courses with respect to the program term. They do not appear in the training set, but appears in the testing set. As shown in Table 2, program term t_1 sees all cold start dyads with new students (denoted by CSS). The other program terms however hardly encounter new students. Dyads involving cold start/new courses (denoted by CSC) are relatively fewer as not many new courses are introduced in each program term. Most of the new courses are observed in the program terms t_6 to t_8, the last 3 terms of the program. The other dyads are the non-cold start (NCS) dyads.

3.2 Student-Course Features

We consider five categories of features for representing the student-course dyads (s_i, c_j):

Static student features. These are features of a student which do not change with time as they are not associated with any target program term, such as student's *major*, *gender*, *alma_mater*, and *cohort*.

Dynamic student features. These are student features derived from the data and their values may vary in different target program terms. These features are particularly useful to determine the latest performance and academic load of the student, such as student's average grade in the previous program term (*lterm_gpa*) and up to previous program term (*lterm_cum_gpa*), number of credit units (CUs) a student received up to previous program term (*total_chrs*) and registered in the target program term (*term_chrs*), average CUs per program term taken by a student (*speed*), number of courses taken by a student in every course discipline up to target program term (*disc_distrib*), relative CUs gained by a student compared to all students in the same cohort (*rel_total_chrs*), and relative *lterm_cum_gpa* of a student compared with that of the cohort (*rel_lterm_cgpa*).

Static course features. These are features of a course c_j that do not change with time: course's discipline (*disc*), CUs (*chrs*), and level (*clevel*).

Dynamic course features. These are features of a course c_j that change with time: instructor of c_j (*iid*), number of students taking c_j in the target program term (*num_enrolled*) and in all previous program terms (*total_enrolled*), average grade (*term_cgrade*) and grade distribution (*term_dgrade*) obtained by students of the previous cohort when they took c_j in the target program term, average grade (*lterm_cum_cgrade*) and grade distribution (*lterm_cum_dgrade*) obtained by students of the same and previous cohorts when they took c_j in any program terms in the past.

Student-course interaction features. As we know which student s_i takes which course c_j in the target program but not the grade, we can exploit this information to derive some features that capture the *indirect* interaction between s_i and c_j for us to determine if s_i will perform well in c_j. We derive *rel_cterm* that measures the program term s_i registered for c_j relative to the program term other students of the same cohort taking c_j. We also derive *disc_grade* which is the average grade obtained by s_i when taking any courses sharing the same course discipline as c_j in the previous terms.

4. PROPOSED METHODS

Two methods are proposed for the next-term grade prediction task, namely, **Factorization Machine (FM)**, and **Integrated Long and Short Term Memory with FM (LSTM-FM)**. The former is often used for recommendation tasks. The latter is a sequence model combined with FM to predict grades of courses in each program term.

4.1 Factorization Machine (FM)

To use FM for next-term grade prediction, our training data for predicting grades in a target term t_k is represented by a $N_{trg}^{dyads} \times p$ matrix, X, where N_{trg}^{dyads} represents the number of training dyads, $p = |S| + |C| + |F|$, and F represent the set of features. Each row $X(i, j)$ for dyad (s_i, c_j) consists of a one-hot vector of student ids, a one-hot vector of course ids, and the features representing the dyad (s_i, c_j).

Model. FM captures both 1-way and 2-way interactions between all features using factorized interaction parameters, as formulated below.

$$\hat{Y}_{i,j} = w_0 + \sum_{k=1}^{p} w_k X_{i,j,k} + \sum_{k=1}^{p} \sum_{k'=1}^{p} X_{i,j,k} X_{i,j,k'} \sum_{f=1}^{k} v_{k,f} v_{k',f}$$

where w_0 captures the global intercept and together with the $\sum_{k=1}^{p} w_k X_{i,j,k}$ serves as a basic linear regression model. The last part contains all pairwise interactions of the X features, which is modeled as a factorized parameterization $\sum_{f=1}^{k} v_{k,f} v_{k',f}$.

4.2 LSTM-FM Model

In LSTM-FM, we merge a sequence model with FM to both learn the sequence of grades received by a student and predict the grades in the target program term using the observed sequence as well as the feature interaction for the student-course dyads. The LSTM-FM framework (Figure 2) is decomposed into two main components: 1) *Input Layer* that utilizes bidirectional LSTM networks (Bi-LSTM) [1] to model the historical grades of a student and 2) *Interaction Layer* that employs interaction module similar to FM in order to model features interactions. The returned value is

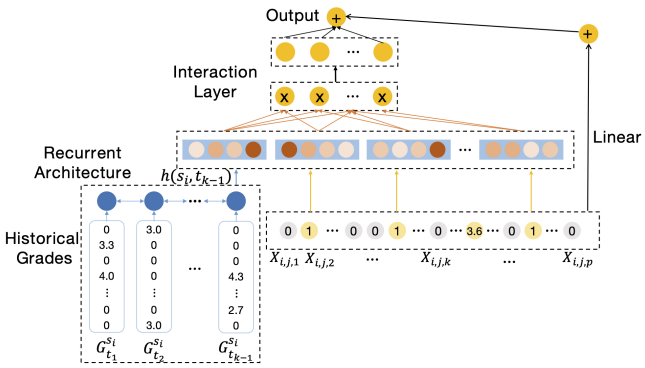

Figure 2: LSTM-FM framework

then transformed into the predicted grade by using 2-layer feed-forward networks with layer normalization [4].

As there can be a number of courses taken by the student in the same program term, we define $G^{s_i}_{t_k,c_j}$ as a |C| dimensional vector keeping the grade score of student s_i gets for course c_j in program term t_k. We then use historical courses-grades of student s_i, $G^{s_i}_{t_k,c_j}$'s, for terms $t_1, \cdots, t_{k-1}$ to learn the hidden states using Bi-LSTM. We subsequently concatenate the hidden states $\overrightarrow{h}(s_i, t_{k-1})$ and $\overleftarrow{h}(s_i, t_{k-1})$ of the bi-LSTM into $h(s_i, t_{k-1})$ which is fed to the interaction layer with the (s_i, c_j)'s features to predict $G^{s_i}_{t_k,c_j}$.

5. EXPERIMENTS

5.1 Evaluation Metrics

Root mean squared error (RMSE) and mean absolute error (MAE) are used to evaluate the accuracy of different grade prediction methods as formulated below. The grades need to be converted to numerical values before using the two metrics. For both RMSE and MAE, the error is defined by the difference between the predicted grade and the actual grade. RMSE is appropriate to penalize methods that yield large errors. MAE, on the other hand, provides the average difference between the predicted and actual grades. For example, for a given actual grade of A- (with numeric score = 3.7), an MAE of 0.3 suggests that the predicted grade differs from the actual grade by an average of half grade, say B+ (with score = 3.4) or A (with score = 4.0).

$$RMSE = \sqrt{\frac{\sum_{Y^{trg}_{i,j} \text{ is defined}} (\hat{Y}^{test}_{i,j} - Y^{test}_{i,j})^2}{|\{(i,j)|Y^{trg}_{i,j} \text{ is defined}\}|}}$$

$$MAE = \frac{\sum_{Y^{trg}_{i,j} \text{ is defined}} |\hat{Y}^{test}_{i,j} - Y^{test}_{i,j}|}{|\{(i,j)|Y^{trg}_{i,j} \text{ is defined}\}|}$$

5.2 Methods for Evaluation

We focus on evaluating FM and LSTM-FM with the features defined in Section 3. There are several variants for both depending on what features are used: FM and LSTM-FM without any features other than student id and course id are also included (**FM and LSTM-FM without features**),

Table 3: Overall Results (CSC: Cold Start Courses)

Method	All dyads		Dyads w/o CSC		Only CSC dyads	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
UR	1.710	1.382	1.735	1.404	1.696	1.378
GM	0.755	0.551	0.757	0.552	0.638	0.506
MoM	0.676	0.488	0.678	0.488	0.583	0.448
Without student-course features						
LR	0.628	0.456	0.629	0.455	**0.577**	**0.434**
FM	0.607	0.428	0.608	0.428	**0.552**	**0.415**
LSTM-FM	0.651	0.464	0.652	0.464	0.618	0.490
With all student-course features						
LR	0.629	0.457	0.630	0.459	0.585	0.446
FM	0.625	0.448	0.622	0.445	0.587	0.457
LSTM-FM	0.628	0.449	0.629	0.449	**0.574**	**0.441**
With selected student-course features						
LR	**0.621**	**0.452**	**0.621**	**0.455**	0.583	0.452
FM	**0.594**	**0.425**	**0.594**	**0.428**	0.601	0.450
LSTM-FM	**0.603**	**0.437**	**0.603**	**0.436**	0.606	0.476

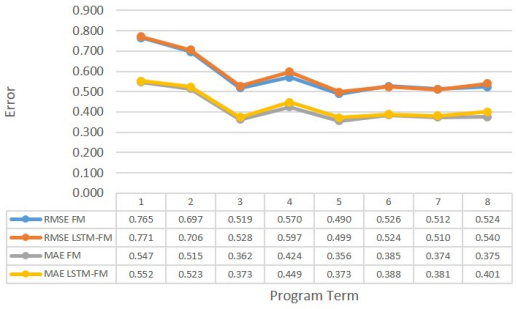

Figure 3: Prediction error per program term

FM and LSTM-FM with all features and **FM and LSTM-FM with only selected features** (Section 5.3).

We include several baseline methods for comparison: **uniform random (UR)** that randomly predicts a grade score from interval [0, 4.3], **global mean (GM)** that predicts a grade score using the average of all observed grades in the training set, **mean of means (MoM)** that returns the average of (a) the predicted grade score of GM; (b) the average observed grades of this student in the training set; and (c) the average observed grades of this course in the training set, and **linear regression (LR)** that uses the first two components of FM ($w_0 + \sum_{k=1}^{p} w_k X_{i,j,k}$) to predict a grade.

5.3 Prediction Results

The overall prediction results are summarized in Table 3. UR yields the highest error. With the use of historical data, GM can predict with smaller errors. MoM further reduces the prediction error with more information used. By implementing a traditional machine learning approach, LR, we can obtain lower prediction error. The results show that the historical data contribute to grade prediction accuracy, and it is worthwhile to explore more machine learning approaches to improve this grade prediction task.

We then analyse the results of our proposed methods. It is interesting to see that FM with only student id and course id predicts grades quite well. It is also applied to LSTM-FM although the latter has a larger error. FM (and LSTM-FM) with all features actually performs worse than the one without features. With selected features (by excluding *cohort*, *disc_distrib*, *iid*, *term_dgrade*, and *lterm_cum_dgrade*), both methods achieve the best results. The overall results show that the lowest error obtained by LR in every scenario is always higher than those of FM and LSTM-FM. This suggests that the 2-way interaction captured in both FM and LSTM-FM can improve prediction accuracy compared to LR that only captures linear model. The results so far are encouraging as an MAE of 0.425 is smaller than a $\frac{3}{4}$ grade difference. We evaluate the methods for dyads that do not involve CSC to see if they are able to improve prediction accuracy. Table 3 shows that CSC dyads do not make significant difference to the prediction results. This suggests that the methods are robust against CSC.

The prediction errors for each program term are illustrated in Figure 3. We observe that both FM and LSTM-FM have similar performance on predicting grades in every program term. The first two program terms t_1 and t_2 have relatively higher errors compared to the latter terms due to lesser training data. t_1 also handles grade prediction for cold start students. As the amount of training data increases, we notice a significant error improvement from term t_3 onwards. The error converges at term t_5 when the model has sufficient training data. For terms t_5 to t_8, both methods can maintain the MAE to be below 0.401.

6. DISCUSSION AND FUTURE WORK

Based on the proposed framework in this paper, we plan to develop a grade prediction API for the university that can be used by both students and instructors. This may help students to select courses that are appropriate to enroll, given their performance in past terms. Instructors then may use this API to understand the class profile, see the predicted performance of their students and use this information to adjust class outline and delivery method. We plan to explore using course description and knowledge graph to improve prediction accuracy. More advanced deep learning models can also be introduced to explain the prediction results.

7. ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

8. REFERENCES

[1] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[2] M. Sweeney, J. Lester, and H. Rangwala. Next-term student grade prediction. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 970–975, Oct 2015.

[3] M. Sweeney, H. Rangwala, J. Lester, and A. Johri. Next-term student performance prediction: A recommender systems approach. *CoRR*, abs/1604.01840, 2016.

[4] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu. On layer normalization in the transformer architecture. *CoRR*, abs/2002.04745, 2020.