Grade Prediction Based on Cumulative Knowledge and Co-taken Courses

Zhiyun Ren
Computer Science
George Mason University
4400 University Drive,
Fairfax, VA 22030
zren4@gmu.edu

Xia Ning
Biomedical Informatics
The Ohio State University
Columbus, OH 43210
Xia.Ning@osumc.edu

Andrew S. Lan
College of Information and Computer Sciences
University of Massachusetts Amherst
140 Governors Dr., Amherst, MA 01003
andrewlan@cs.umass.edu

Huzefa Rangwala
Computer Science
George Mason University
4400 University Drive,
Fairfax, VA 22030
rangwala@cs.gmu.edu

ABSTRACT

Over the past decade, low graduation and retention rates have plagued higher education institutions. To help students graduate on time and achieve optimal learning outcomes, many institutions provide advising services supported by educational technologies. Accurate grade prediction is an integral part of these services such as degree planning software, personalized advising systems and early warning systems that can identify students at-risk of dropping from their field of study. In this work, we present next-term grade prediction models based on students' cumulative knowledge and co-taken courses. The proposed models are based on a matrix factorization framework and incorporate a co-taken course interaction function to learn the influence from the co-taken courses on the target course. The co-taken course interaction function is formed by a neural network, which takes the knowledge difference between the co-taken courses and the target course as input, and outputs an influence value that will be used to predict students' grades on the target course. The experimental results on various datasets from a U.S. University demonstrate that the proposed models significantly outperform competitive baselines across different test sets. Furthermore, we analyze the proposed models' performance with different numbers of co-taken courses as well as different numbers of co-taken course subjects, and highlight with an application case study how a student might make decisions related to selection of courses. The codes are available at `https://github.com/Zhiyun0411/EDM`.

Keywords

matrix factorization, next-term grade prediction, cumulative knowledge, co-taken courses

Zhiyun Ren, Xia Ning, Andrew Lan and Huzefa Rangwala "Grade Prediction Based on Cumulative Knowledge and Co-taken Courses" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 158 - 167

1. INTRODUCTION

For over a decade higher education institutions in the United States have been grappling with low graduation rates [9]. The National Center for Education Statistics [1] reports that approximately 59% of students who started college in 2009 were able to graduate and obtain a 4-year college program degree within 6 years. There is a pressing need for data-driven applications and services to guide students through academic pathways and achieve better learning outcomes. Many higher education institutions have implemented programs and services supported by educational technologies to increase overall graduation rates [17]. For example, Academic Advising service [2] provides effective student-centered advising at Purdue University. Graduation Progression Success (GPS) Advising [3] implemented at Georgia State University helps identify at-risk students and have advisors respond alerts. Their reports show a 6% increase of 6-year graduation rate over 4 years. Our work aims to help students select courses for the next term by developing methods that can provide accurate grade prediction for the courses they have not taken yet.

In the past few years, many approaches have been developed for next-term grade prediction. One of the most popular approaches is matrix factorization (MF), which is inspired from the Recommender Systems (RS) literature [2, 3, 7, 15, 18]. Specifically, MF decomposes the student-course grade matrix into two matrices containing student and course latent factors, respectively. The predicted grade of a student on a course is given by the inner product of the corresponding student and course latent factors [4, 10]. There are other extended MF-based models which achieve better grade prediction results than MF. For example, Morsy *et al.* [8] proposed a Cumulative Knowledge-based Regression Model (CK) to tackle the next-term grade prediction problem. CK models each student with cumulative knowledge acquired by the student in the

[1] https://nces.ed.gov
[2] http://www.purdue.edu/advisors/index.html
[3] http://giving.gsu.edu/student-success/

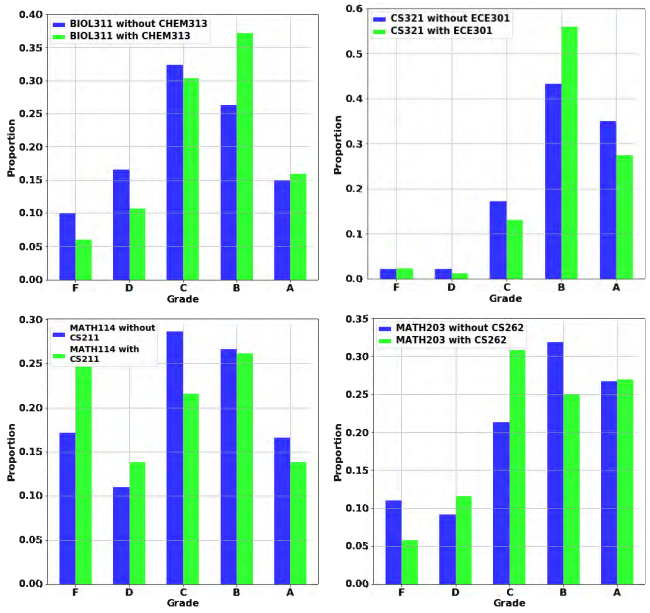

Figure 1: Students' Performance with Different Co-taken Course Pairs. Note: BIOL311 is course "General Genetics". CHEM313 is course "Organic Chemistry". CS321 is course "Software Engineering". ECE301 is course "Digital Electronics". MATH114 is course "Analytic Geometry and Calculus". CS211 is course "Object Oriented Programming". MATH203 is course "Linear Algebra". CS262 is course "Low-level Programming".

past terms. However, among all the existing methods for next-term grade prediction [2,13,14], very few consider the effect of co-taken courses on students' performance.

We conduct a statistical analysis on a dataset collected from George Mason University in order to demonstrate the effects of co-taken courses on students' performance. Figure 1 shows the true grade distribution of students' on a specific course *with* and *without* enrolling in another course in the same term. The course pairs we choose in this analysis are frequently co-occuring in our dataset. For each target course pair, we choose the students who take more than four courses in a term, including the corresponding course pairs. We keep the students if the other co-taken courses only share few topics/material as the target course pairs. Figure 1 shows that students who take BIOL311 (Genetics) with CHEM313 (Organic Chemistry) have fewer "F", "D" and "C" grades, and several more "B" grades than those students who only take BIOL311 in a term. Similar trend has been found for course pairs CS321 (Software Engineering) and ECE301 (Digital Electronics). Moreover, students who take MATH114 (Calculus) with CS211 (Object Oriented Programmming) will have more "F" grades than those students who only take MATH114 in a term. Students who co-take MATH203 (Linear Algebra) and CS262 (Low-level programming) have more "C" grades than those students who only take MATH203 in a term. This shows that it can be challenging for students to take some courses together in a term (e.g., MATH114 and CS211, MATH203 and CS262), while it might not cause grade drop if taking other course pairs together (e.g., BIOL311 and CHEM313, CS321 and ECE301). Thus, we assume that co-taken courses can have substantial effect on student grades in different ways.

In this work, we propose grade prediction models that incorporate both **C**umulative **K**nowledge and **C**o-taken **C**ourses (CKCC) to predict students' performance in the next term. Inspired by Morsy *et al.* [8], the proposed methods model each student's latent factors by cumulating the knowledge provided by the sequence of courses the student has taken in the past terms. Furthermore, we introduce a co-taken course interaction function to model the influence of the co-taken courses on students' performance. The co-taken course interaction function is formed by a neural network which takes the knowledge difference between the co-taken courses and the target course as input, and outputs an influence value from the co-taken courses on the target course. We conduct comprehensive experiments on various datasets collected from George Mason University and thorough analysis on the effect of co-taken courses. Our experimental results show that CKCC significantly outperforms other competitive baselines methods for the task of grade prediction. We also provide detailed case study on how our model can help student in course selection for the next term.

The main contributions can be summarized as follows:

1. We develop CKCC models on next-term grade prediction. The models consider both students' cumulative knowledge and co-taken courses in the target term. To the best of our knowledge, this is the first work that learns and explicitly incorporates influences from co-taken courses for grade prediction.
2. We provide a detailed case study on how our model helps students in course selection for the next term by comparing the performance of CKCC with different sets of co-taken courses.

2. RELATED WORK

2.1 Grade Prediction Approaches

Methods originating from recommender systems research have attracted increasing attention in educational data mining [2,3,13,14, 20]. Sweeney *et al.* [18,19] applied several recommender systems approaches to predict next-term grades. The authors implemented MF-based methods including SVD, SVD-kNN and factorization machine and simple baseline methods including global, student, and course means. The work showed that MF-based methods consistently achieve better grade prediction results over the baselines. Elbadrawy *et al.* [1] developed a domain-aware grade prediction method with student/course-group biases. This method groups students based on majors and academic levels. Additionally, it groups courses based on course levels and course subjects. The method assumes that the students/courses in a same group tend to have similar biases. Accordingly, this method models biases for each student and course group within a MF framework and achieved significant improvement on grade prediction performance over baselines.

2.2 Grade Prediction based on Student Historical Information

Polyzou *et al.* [12] addressed the future course grade prediction problem with different approaches based on sparse linear models and MF approaches. The experimental results showed that the course-specific regression approach achieved the best performance among all approaches. This method predict a student's performance using a sparse linear combination of the grades that the student obtained in past courses. Morsy *et al.* [8] proposed a model named Cumulative Knowledge-based Regression Model (CK) to predict student's grade on a certain course at the next term. CK

models each student with the cumulative knowledge he/she obtained from the sequence of courses he/she took in the past. Then CK calculated the inner product of the cumulative knowledge vector of a student and the required knowledge vector of the target course as the predicted grade. The experimental results showed that CK significantly outperforms MF in grade prediction. Ren *et al.* [13] proposed a matrix factorization model with temporal course-wise influence to predict next term student grades. This model considers two components in predicting a student's grade on a certain course: (i) the student's competence with respect to the target course's topics, content and requirements, etc., and (ii) student's previous performance over other courses. The study concluded that considering temporal influence can significantly improve the next-term grade prediction performance.

2.3 Neural Network in Educational Data Mining

Neural networks have been applied to solve many educational data mining problems. For example, Sharma *et al.* [16] proposed a composite deep neural network to predict whether the educational video is lively or not. The proposed method first used a convolutional neural network to extract the video features, and then used a deep recurrent neural network to predict the human movement label in order to detect video liveliness. Klingler *et al.* [6] presented a semi-supervised classification pipeline that employed deep variational auto-encoders to detect students who are suffering from developmental dyscalculia. Piech *et al.* [11] introduced Deep Knowledge Tracing (DKT) to model student learning with Recurrent Neural Networks. The authors provided experiments on how to use DKT to detect latent structure between the assessments in the dataset. The models proposed in this paper tackle the challenges of next-term grade prediction with students' history information (the sequence of courses the student has taken) and the co-taken courses in the next term. The main contribution of our model is to explicitly incorporate the co-taken courses with in MF framework.

3. PRELIMINARIES AND PROBLEM DEFINITION

3.1 Problem Definition

Formally, student-course grades will be represented by G_1, G_2, ..., G_T for a total of T terms. Each G_t is a matrix, and contains the set of student-course grades for all students enrolled in courses within term t. For all the students, the set of student-course grades up to term t can be represented by $G^t = \bigcup_{i=1}^{t} G_i$. The set of courses that student s has taken in term t is represented by $C_{s,t}$ and the set of grades that student s achieves in term t is represented by $G_{s,t}$. The set of courses that student s has taken up to term t is represented by C_s^t, and the set of grades that student s has achieved up to term t is represented by G_s^t.

In this paper, all vectors are represented by bold lower-case letters and all matrices are represented by upper-case letters. Row vectors are represented by having the transpose superscript $^\mathsf{T}$, otherwise by default they are column vectors. A predicted value is denoted by having a $\tilde{}$ symbol. Table 1 summarizes the key notations used in this paper.

Given student-course grades up to term $t-1$ and the set of courses each student plans to take at term t, the objective of our work is to predict student's grades on a specific course given the set of co-taken courses at term t.

3.2 Grade Prediction based on Matrix Factorization

MF methods factor the student-course grade matrix into two matrices containing latent factors of courses and students in a common knowledge space, respectively [1,12]. The dimension of the knowledge space is much lower than that of the original student-course grade matrix. We use $\mathbf{p}_s$ ($\mathbf{p}_s \in \mathbb{R}^k$) and $\mathbf{q}_c$ ($\mathbf{q}_c \in \mathbb{R}^k$) to represent latent factors of k dimensions for student s and course c, respectively. Thus, the grade of student s on course c can be predicted as

$$\tilde{g}_{s,c} = \mathbf{p}_s^\mathsf{T} \mathbf{q}_c + b_s + b_c. \tag{1}$$

where b_s and b_c are bias terms for student s and course c, respectively.

3.3 Grade Prediction with Cumulative Knowledge

Morsy *et al.* [8] proposed the CK model which learns each student's latent factors with cumulative knowledge acquired by the student in past terms. Specifically, CK uses two vectors to model a course: the provided knowledge by the course and the prerequisite knowledge of the course, respectively. A student's latent factor is given by the knowledge accumulated from the previous course that the student has taken and the corresponding course grades. Formally, the cumulative knowledge acquired by student s up to term t is represented by $\mathbf{p}^t_{ck(s)}$, and is given by:

$$\mathbf{p}^t_{ck(s)} = \sum_{g_{s,c'} \in G_s^{t-1}} (e^{-\lambda(t-t_{s,c'})} \mathbf{k}_{c'} \cdot g_{s,c'}), \tag{2}$$

where $t_{s,c'}$ is the term in which student s took course c', $e^{-\lambda(t-t_{s,c'})}$ is an exponential time decay function with $\lambda > 0$ denoting the decay rate, $\mathbf{k}_{c'}$ is the latent knowledge factor of course c', and $g_{s,c'}$ is the grade of student s on course c'. Given $\mathbf{p}^t_{ck(s)}$, CK predicts student s's grade on course c in term t as follows:

$$\tilde{g}^t_{s,c} = {\mathbf{p}^t_{ck(s)}}^\mathsf{T} \mathbf{q}_c. \tag{3}$$

Note that in prior work, Ren *et al.* [14] have shown that CK can achieve better grade prediction performance when the cumulative knowledge $\mathbf{p}^t_{ck(s)}$ is averaged in Eq 3. Therefore, $\tilde{g}^t_{s,c}$ is presented as follows:

$$\tilde{g}^t_{s,c} = \frac{1}{|G_s^{t-1}|} \sum_{g_{s,c'} \in G_s^{t-1}} (e^{-\lambda(t-t_{s,c'})} \mathbf{k}_{c'} \cdot g_{s,c'})^\mathsf{T} \mathbf{q}_c, \tag{4}$$

We refer to this model as the averaged cumulative knowledge (CK) model and will consider it as one of our baseline methods.

4. METHODS

4.1 Model Overview

In this paper, we propose grade prediction models that incorporate **C**umulative **K**nowledge and **C**o-taken **C**ourses (CKCC). To predict student s's grade on course c in term t, CKCC takes into account two factors: i) cumulative knowledge of student s up to term $t-1$, and ii) the other courses that will be taken together with course c in term t. To model the first factor, we adopt the CK model as in Eq. 4, that is, we cumulate the provided knowledge of the courses which student s has taken in the past, denoted as c', to represent his/her cumulative knowledge, and use a latent factor to represent knowledge required by course c. To model the second factor, we introduce an co-taken course interaction function $f(\cdot)$ to learn the

Table 1: Notations

Notation	Explanation
m	number of courses
n	number of students
k	number of latent dimensions
$\mathbf{p}^t_{ck(s)}$	the cumulative knowledge of student s up to term t
$\mathbf{q}_c$	latent factor of the required knowledge components of course c
$\mathbf{k}_c$	latent factors of the provided knowledge components of course c
b_s	student bias term
b_c	course bias term
$g^t_{s,c}$	the grade of student s on course c at term t
$t_{s,c}$	the academic term when student s takes course c
G_t	student-course grades at term t
G^t	all the student-course grades up to term t
$G_{s,t}$	all the grades student s obtains at term t
G^t_s	all the grades student s obtains up to term t
$C_{s,t}$	the set of courses student s chooses at term t
C^t_s	the set of courses student s chooses up to term t

influence from co-taken courses, denoted as c'', on student s's grade on course c in term t.

Specifically, we use a latent vector $\mathbf{q}_c$ to represent the knowledge components that course c requires. We hypothesize that the difference of the required knowledge between two courses will cause the influence from one course on the other, as shown in Figure 1. Based on this hypothesis, the difference between $\mathbf{q}_c$ of course c and $\mathbf{q}_{c''}$ of a co-taken course c'' can be used in $f(\cdot)$ to learn the influence from c'' to c. We sum up the differences between each co-taken course c'' and c in order to aggregate the influence. Thus, the sum of the absolute values of the differences between each $\mathbf{q}_{c''}$ and $\mathbf{q}_c$, that is, $\sum_{c''\in C_{s,t}\setminus\{c\}} |\mathbf{q}_{c''} - \mathbf{q}_c|$, is used in $f(\cdot)$ to learn the influence from all co-taken courses. Note that the use of absolute values here is to avoid the scenarios in which the influences from different co-taken courses are canceled out. Thus, CKCC predicts student s's grade on course c in term t as follows:

$$\tilde{g}^t_{s,c} = \frac{1}{|G^{t-1}_s|} \sum_{g_{s,c'}\in G^{t-1}_s} (e^{-\lambda(t-t_{s,c'})}\mathbf{k}_{c'} \cdot g_{s,c'})^{\mathsf{T}}\mathbf{q}_c + f(\sum_{c''\in C_{s,t}\setminus\{c\}} (|\mathbf{q}_{c''} - \mathbf{q}_c|)), \tag{5}$$

where $|\mathbf{q}_{c''} - \mathbf{q}_c|$ is the vector of absolute values of entry-wise difference between latent vector $\mathbf{q}_{c''}$ and latent vector $\mathbf{q}_c$, $c'' \in C_{s,t} \setminus \{c\}$ indicates that course c'' is one of courses taken together with c in term t. Note that in Eq. 5, the two terms share a common latent vector $\mathbf{q}_c$.

4.2 Co-taken Course Interaction Function

In CKCC, the co-taken course interaction function $f(\cdot)$ learns the influence on student s's grade on course c from all the other co-taken courses in term t. We hypothesize that such influence can be nonlinear in general. Therefore, we use a feedforward neural network (FNN) [21] as $f(\cdot)$ to model the influence. The FNN takes the input as described in last section, and outputs a scalar influence value on course c. We use hyperbolic tangent (Tanh) as the activation function in each layer of the FNN. Note that when there are no hidden layers and no nonlinearity, the FNN model learns the weights directly from the input layer (i.e., difference of courses) to the output layer (i.e., the influence), and the function $f(\cdot)$ becomes a simple inner product operation (parameterized by a vector). This simplified model is referred to as CKCC-*l*. Figure 2 shows the structure of the CKCC model.

Algorithm 1 *CKCC*: Learn

procedure CKCC_LEARN
  Initialize $\mathbf{k}_c$, $\mathbf{q}_c$ for each c
  $\eta \leftarrow$ learning rate
  $T \leftarrow$ number of terms in training set
  $\lambda \leftarrow$ time decay parameter
  $\alpha_1, \alpha_2, \alpha_3 \leftarrow$ regularization weight
  $t \leftarrow 2$
  $iter \leftarrow 0$
  while iter<maxIter **do**
    for $t \le T$ **do**
      for all $g^t_{s,c} \in G^t_s$ **do** ▷ step 1
        $\hat{g}^t_{s,c} \leftarrow g^t_{s,c} - f(\sum_{c'\in C_{s,t}\setminus\{c\}}(|\mathbf{q}_{c''} - \mathbf{q}_c|))$
        $\mathbf{p}_{ck(s)} \leftarrow 0$
        for all $c' \in C^{t-1}_s$ **do**
          $\mathbf{p}_{ck(s)} \leftarrow \mathbf{p}_{ck(s)} + e^{-\lambda(t_{s,c}-t_{s,c'})}\mathbf{k}_{c'} \cdot g^{t_{s,c'}}_{s,c'}$
        $\tilde{g}^t_{s,c} \leftarrow \mathbf{p}^{\mathsf{T}}_{ck(s)}\mathbf{q}_c$
        $e^t_{s,c} = \hat{g}^t_{s,c} - \tilde{g}^t_{s,c}$
        for all $c' \in C^{t-1}_s$ **do**
          $\mathbf{k}_{c'} \leftarrow \mathbf{k}_{c'} + \eta(\mathbf{q}_c \cdot e^{-\lambda(t_{s,c}-t_{s,c'})} \cdot g_{s,c'} \cdot e^t_{s,c} - \alpha_1 \cdot \mathbf{k}_{c'})$
        $\mathbf{q}_c \leftarrow \mathbf{q}_c + \eta(\mathbf{p}_{ck(s)} \cdot e^t_{s,c} - \alpha_2 \cdot \mathbf{q}_c)$
      for all $g^t_{s,c} \in G^t_s$ **do** ▷ step 2
        $\hat{g}^t_{s,c} \leftarrow g^t_{s,c} - \mathbf{p}_{ck(s)}\mathbf{q}_c$
        $\tilde{g}^t_{s,c} \leftarrow f(\sum_{c'\in C_{s,t}\setminus\{c\}}(|\mathbf{q}_{c''} - \mathbf{q}_c|))$
        $e^t_{s,c} = \hat{g}^t_{s,c} - \tilde{g}^t_{s,c}$
        Update Θ_f with Adam
    $iter \leftarrow iter + 1$
  return $\Theta = \{\{\mathbf{k}_c\}, \{\mathbf{q}_c\}\}, \Theta_f$

4.3 Optimization of CKCC

Given the grade estimation as in Equation 5, we formulate the grade prediction problem for term T as the following optimization problem:

$$\begin{aligned}\underset{\Theta,\Theta_f}{\text{minimize}} \sum_s \sum_{t=1}^{T-1} \sum_{g^t_{s,c}\in G^t_s} (g^t_{s,c} - \tilde{g}^t_{s,c})^2 \\ + \alpha_1(|\mathbf{k}_c| + |\mathbf{q}_c|) + \alpha_2(\|\mathbf{k}_c\|_2^2 + \|\mathbf{q}_c\|_2^2) \\ + \alpha_3\|\text{vec}(\Theta_f)\|_2^2,\end{aligned} \tag{6}$$

where $\Theta = \{\{\mathbf{k}_c\}, \{\mathbf{q}_c\}\}$ represents the set of latent vectors, and Θ_f represents the parameters of $f(\cdot)$. α_1, α_2, and α_3 denote the nonnegative weights on the regularization terms to prevent overfitting.

The optimization process for CKCC is presented in Algorithm 1. It consists of two steps: The first step is to update the course parameters, i.e., Θ, using stochastic gradient descent. The second step is to update $f(\cdot)$ parameters, i.e., Θ_f, with the adaptive moment estimation (Adam) algorithm [5].

5. EXPERIMENTS

5.1 Dataset Description

The data used in this work is obtained from George Mason University. Our dataset contains two student groups: first-time freshmen (FTF; i.e., students who begin their study initially at this Uni-

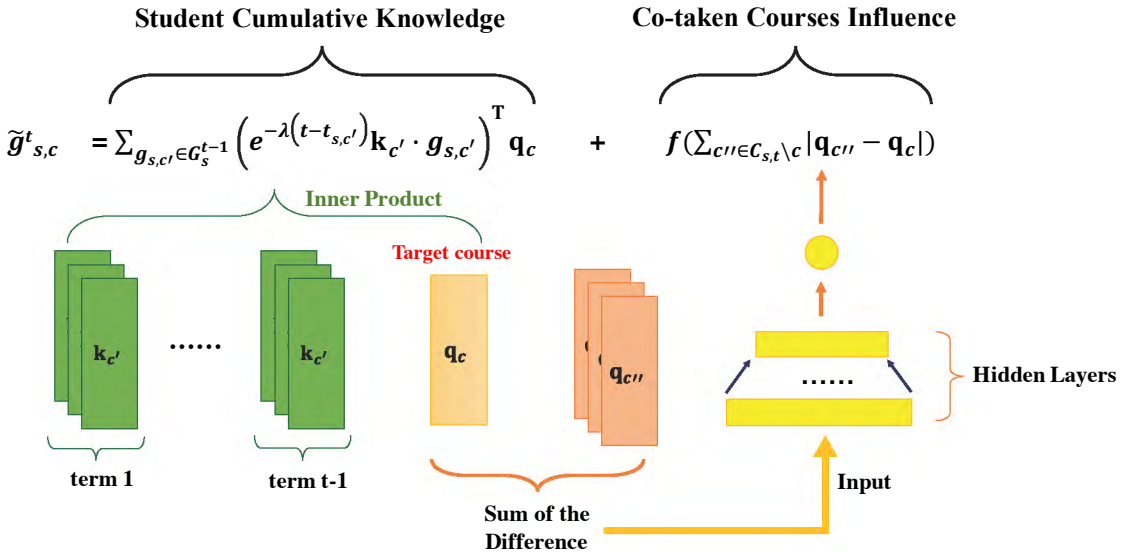

Figure 2: CKCC Model Structure

Table 2: Dataset Statistics

Major	FTF student group			TR student group		
	#S	#C	#S-C	#S	#C	#S-C
MATH	271	693	3,325	243	597	2,031
PHYS	144	488	2,044	73	286	905
CHEM	427	673	4,942	257	473	1,937
IT	430	473	5,984	1,163	487	10,302
CS	819	714	16,955	526	435	7,840
BIOL	1,951	1,197	22,065	1,481	980	10,851

#S, #C and #S-C are the number of students, courses and student-course pairs from Fall 2009 to Spring 2018, respectively.

Figure 3: Different Experimental Protocols

versity), and transfer students (TR; i.e., students who transfer to this University from a different one). The dataset was extracted in the period of Fall 2009 to Spring 2018. It includes information of 23,435 FTF students and 28,470 TR students across 153 majors. For simplicity, we use students from six different majors to evaluate the proposed models. These majors have different numbers of enrolled students, courses, and different major syllabi. We will evaluate these majors on both FTF and TR student groups. The majors in our experiment include: (i) Mathematical Sciences (MATH), (ii) Physics (PHYS), (iii) Chemistry (CHEM), (iv) Information Technology (IT) , (v) Computer Science (CS) and (vi) Biology (BIOL). Table 2 shows the statistics across these majors.

5.2 Experimental Protocols

To assess the performance of our next-term grade prediction models, we trained our models on data up to term $T-1$ and make predictions for term T. We evaluate our method for three test terms, i.e., Spring 2018, Fall 2017 and Spring 2017. As an example, for evaluating predictions for term Fall 2017, data from Fall 2009 to Spring 2017 is considered as training data and data from Fall 2017 is testing data. datasets. Figure 3 shows the three different train-test splits.

5.3 Evaluation Metrics

In our experiments, we use Mean Absolute Error (MAE) to evaluate the predicted results in numbers. MAE is calculated as:

$$MAE = \frac{\sum_{g^t_{s,c} \in G_T} \left| g^t_{s,c} - \tilde{g}^t_{s,c} \right|}{|G_T|} \tag{7}$$

where $g^t_{s,c}$ and $\tilde{g}^t_{s,c}$ are the ground-truth grade and predicted grade for student s on course c at term T, respectively. G_T is the set of student-course grades in the T-th term, which is considered as the test set in our experiment.

Moreover, since a student receives a letter grade for a course, i.e., A, A-, ..., F, we use the Percentage of Tick Accuracy (PTA) [12] as one of our evaluation metrics. During training, we map letter grades “A+" and “A" to the real-valued grade point number 4.0, “A-" to 3.67, “B+" to 3.33, etc. During testing, we map the predicted grade point numbers back to their closest letter grades. Then, we define tick as the difference between two consecutive letter grades (e.g., C+ vs C or C vs C-). We then compute the percentage of predicted grades that match the actual grades (or within 0-ticks of them), and those that are within 1 tick and within 2 ticks of the actual grades as PTA_0, PTA_1, and PTA_2, respectively.

5.4 Compared Methods

Since there is no prior research on the influence of co-taken courses within a same term, we use the two following methods and three other variants of CKCC as baselines in our experiments:

- **MF** The MF model is described as Eq. 1.
- **CK** The CK model is described as Eq. 4.
- **MFCC** We add the co-taken course influence to the MF model, and obtain the Matrix Factorization with Co-taken Courses (MFCC) model. Specifically, the predicted grade of student s on course c at term t is defined as

$$\tilde{g}^t_{s,c} = \mathbf{p}_s^\top \mathbf{q}_c + f(\sum_{c'' \in C^t_s \setminus c} (|\mathbf{q}_{c''} - \mathbf{q}_c|)), \quad (8)$$

  where $\mathbf{p}_s$ denotes the latent factors of for student s. Similar to the CKCC model, we optimize the MFCC model with two steps by alternately updating the latent factors and the model parameters in the mapping function $f(\cdot)$.
- **MFCC-*l*** The MFCC-*l* model is a special case of the MFCC model where $f(\cdot)$ is simply an inner product (parameterized by a vector) instead of an FNN.
- **CKCC-*l*** The CKCC-*l* model is described in Section 4.2.

5.5 Parameter Learning

The set of parameters in the optimization problem (Eq 6) includes the number of latent dimensions (i.e., k), regularization parameters (i.e., α_1, α_2, and α_3) and the decay rate (i.e., λ). We performed a grid search over all the parameters with $k \in \{5, 10, \ldots, 25\}$, and $\alpha_1, \alpha_2, \alpha_3, \lambda \in \{1e-3, 1e-2, 0.1\}$. Note that for the CKCC and MFCC models, the optimal neural network structure (e.g., number of layers, the size of each layer) depends on the value of k. Thus, we swept different neural network structure parameters for every k value in our grid search. The neural network structures that consistently achieve good performance contain one hidden layer with 2 or 3 hidden units.

6. RESULTS AND DISCUSSION

6.1 Overall Performance

Table 3 and 4 shows the overall performance for all methods for both FTF and TR student groups, respectively.

Table 3 shows that for FTF students, CKCC and CKCC-*l* outperform the baseline methods over most datasets. Specifically, CKCC outperforms the other compared methods across different experimental protocols by 4.39%, 7.01%, 3.50%, 3.87% in terms of MAE, PTA_0, PTA_1, and PTA_2, respectively. Furthermore, CK based methods outperform MF based methods on all experimental protocols. This table also shows that co-taken course based methods (MFCC, MFCC-*l* and CKCC, CKCC-*l*) outperform their baseline methods (MF and CK) on all experimental protocols, respectively. This illustrates that for FTF students, both cumulative knowledge and co-taken courses have great influence on student's performance, and the proposed methods can capture such influence accurately.

Table 4 shows that CK has competitive results over TR students. Moreover, for MF based methods, MFCC and MFCC-*l* outperform MF for all the experimental protocols. This illustrates that co-taken courses are likely to have influence on student's performance, but the influence may not be as strong as it is of cumulative knowledge for TR students.

6.2 Analysis on Individual Majors

In order to understand the proposed methods' performance on each major, we have tested all the aforementioned methods on different majors separately. We conducted this group of experiments for both FTF and TR students. And we use Spring 2018 as test set. We provide detailed experimental results in Table 5 and 6.

Table 5 shows that the CKCC model outperforms other compared methods for some majors (e.g., PHYS, CS) on all metrics, but has weak performance on some metrics for other majors (e.g., MATH, CHEM). Especially for MATH major, CKCC has the highest MAE result while MFCC and MFCC-*l* have the best MAE result. The reason might be that the performance of CKCC relies on the student historical information, and it tends to have good performance on the students with rich historical information. However, in the test set, some students in certain majors do not have much historical information and thus drag down the model performance. Table 6 shows that, for TR students, there is no method that consistently outperforms others across different metrics. The reason might be that the diversity in student characteristics (many TR students have different backgrounds) leads to diverse course selection plans among them. Such diversity greatly influences the performance of the different models.

6.3 Linear versus Nonlinear Mapping Function

As aforementioned, we have two forms of co-taken course interaction function: FNN model and linear model (parameterized by a vector). Specifically, we compare the results for MFCC versus MFCC-*l*, and CKCC versus CKCC-*l*, respectively, in order to understand how different mapping functions $f(\cdot)$ influence grade prediction performance. Table 3 shows that for FTF students, MFCC-*l* has slightly better performance than MFCC, and CKCC-*l* has competitive performance as CKCC across different experimental protocols. Same trend has shown in table 4 for TR students. Furthermore, table 5 shows that MFCC and CKCC consistently outperform MFCC-*l* and CKCC-*l* across different majors for FTF students. This illustrates that the influence of co-taken courses for FTF student group can be better captured by a nonlinear model (i.e., FNN) than a simple linear model. Table 6 shows that for TR students, MFCC and CKCC don't always outperform MFCC-*l* and CKCC-*l* for different majors. The reason might be that some TR students will have fewer co-taken courses than those of FTF students, and the influence from co-taken courses can be well captured by a linear model.

6.4 Performance on Different Numbers of Co-taken Courses

In this section, we test the CKCC model on different data subgroups with different number of co-taken courses in a term. Specifically, we take the students in the test set and divide them into five groups: students who take $\{2, 3, 4, 5, 6+\}$ courses (6+ refers to six and more). We perform this experiment on each major for both FTF and TR students, respectively. For the sake of page limit, we only show the results for FTF students. Figure 4 shows the experimental results in terms of PTA_0, PTA_1 and PTA_2. The results show that different majors exhibit different trends when the number of co-taken courses varies. For example, for CHEM and BIOL majors, the performance of the CKCC model on PTA improves with more

Table 3: Performance Comparison for All Methods on FTF students

Method	Spring 2018				Fall 2017				Spring 2017			
	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2
MF	0.762	0.172	0.303	0.549	0.759	0.168	0.303	0.556	0.772	0.162	0.306	0.540
MFCC-*l*	0.756	0.180	0.320	0.565	0.745	0.186	0.331	0.574	0.757	0.181	0.331	0.564
MFCC	0.763	0.175	0.317	0.573	0.753	0.188	0.322	0.573	0.760	0.173	0.317	0.565
CK	0.726	**0.190**	0.330	0.575	0.724	0.184	0.336	0.575	0.727	0.186	0.333	0.575
CKCC-*l*	**0.711**	0.189	**0.338**	0.589	0.712	0.191	**0.343**	**0.589**	0.717	0.182	0.332	0.587
CKCC	0.716	0.187	0.332	**0.593**	**0.709**	**0.195**	0.334	**0.588**	**0.710**	**0.196**	**0.339**	**0.594**

Table 4: Performance Comparison for All Methods on TR students

Method	Spring 2018				Fall 2017				Spring 2017			
	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2
MF	0.775	0.184	0.316	0.537	0.760	0.157	0.300	0.565	0.773	0.168	0.299	0.550
MFCC-*l*	0.763	0.178	0.315	0.543	0.748	0.187	0.326	0.571	0.755	0.185	0.328	0.563
MFCC	0.761	0.174	0.321	0.544	0.754	0.177	0.330	0.580	0.761	0.177	0.316	0.569
CK	0.753	**0.268**	**0.400**	**0.586**	0.770	**0.259**	**0.389**	0.570	0.750	**0.273**	**0.397**	**0.583**
CKCC-*l*	**0.733**	0.182	0.324	0.560	0.743	0.180	0.313	0.558	**0.739**	0.172	0.310	0.563
CKCC	0.735	0.181	0.323	0.562	**0.728**	0.175	0.335	**0.571**	0.740	0.169	0.318	0.553

Table 5: Performance Comparison for All Methods on FTF students on Different Majors

Method	MATH				PHYS				CHEM			
	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2
MF	0.762	0.234	0.336	0.523	1.099	0.106	0.206	0.383	0.684	0.262	0.399	0.601
MFCC-*l*	**0.758**	0.195	0.333	0.568	0.960	0.113	0.213	0.447	0.678	0.221	0.374	0.589
MFCC	**0.758**	0.206	0.322	0.559	0.998	0.163	0.248	0.433	**0.663**	0.249	0.380	0.592
CK	0.782	0.267	0.378	**0.569**	0.910	0.135	0.270	0.468	0.680	0.249	0.393	0.595
CKCC-*l*	0.784	0.184	0.316	0.535	0.978	0.238	0.294	0.437	0.734	**0.312**	**0.449**	0.611
CKCC	0.842	**0.309**	**0.413**	0.562	**0.842**	**0.254**	**0.373**	**0.508**	0.697	0.290	0.411	**0.620**

Method	IT				CS				BIOL			
	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2
MF	0.655	0.201	0.36	0.623	0.723	0.190	0.346	0.595	0.687	0.253	0.411	0.626
MFCC-*l*	0.664	0.181	0.365	0.630	0.715	0.177	0.326	0.603	0.777	0.317	0.439	0.599
MFCC	0.627	0.231	0.381	0.659	0.704	0.209	0.362	0.605	0.676	0.274	0.429	0.638
CK	0.606	0.299	**0.466**	0.681	0.722	0.244	**0.395**	0.597	**0.643**	0.316	0.464	**0.653**
CKCC-*l*	0.693	0.288	0.460	0.632	0.784	0.242	0.376	0.578	0.771	**0.341**	0.461	0.605
CKCC	**0.600**	**0.310**	0.465	**0.692**	**0.696**	**0.256**	**0.395**	**0.612**	0.660	0.329	**0.467**	0.649

Table 6: Performance Comparison for All Methods on TR students on Different Majors

Method	MATH				PHYS				CHEM			
	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2
MF	0.608	0.270	0.433	0.617	0.675	0.235	0.431	0.569	0.749	0.219	0.325	0.553
MFCC-*l*	0.637	0.270	0.418	0.610	0.669	0.216	0.353	0.588	**0.634**	0.281	0.412	0.649
MFCC	0.621	0.241	0.397	0.645	**0.577**	**0.353**	**0.471**	**0.667**	0.675	0.228	0.404	0.649
CK	**0.573**	0.394	0.545	0.677	0.741	0.200	0.275	0.550	0.679	0.368	0.491	0.623
CKCC-*l*	0.641	0.384	0.515	0.677	0.694	0.325	0.450	0.625	0.651	0.377	0.500	0.667
CKCC	0.613	**0.404**	**0.576**	**0.707**	0.805	0.200	0.350	0.600	0.642	**0.404**	**0.518**	**0.675**

Method	IT				CS				BIOL			
	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2	MAE	PTA_0	PTA_1	PTA_2
MF	0.614	0.217	0.405	0.662	0.836	0.175	0.302	0.538	0.711	0.200	0.341	0.559
MFCC-*l*	0.610	0.227	0.419	0.665	0.818	0.189	0.325	0.541	0.670	0.213	0.366	0.617
MFCC	0.608	**0.243**	0.415	0.658	0.796	0.193	0.333	0.578	0.674	0.206	0.367	0.604
CK	0.608	0.223	0.406	0.659	**0.737**	0.212	0.369	0.577	0.695	0.226	0.370	0.600
CKCC-*l*	**0.598**	0.235	**0.426**	0.659	0.756	0.184	0.343	**0.599**	0.679	0.228	0.384	0.600
CKCC	0.602	0.231	0.412	**0.672**	0.773	**0.234**	**0.371**	0.563	**0.643**	**0.260**	**0.393**	**0.629**

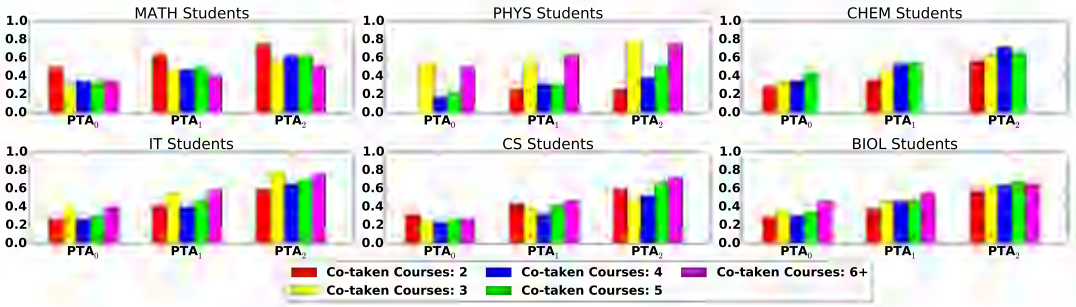

Figure 4: PTA Results for Different Number of Co-taken Courses on FTF students

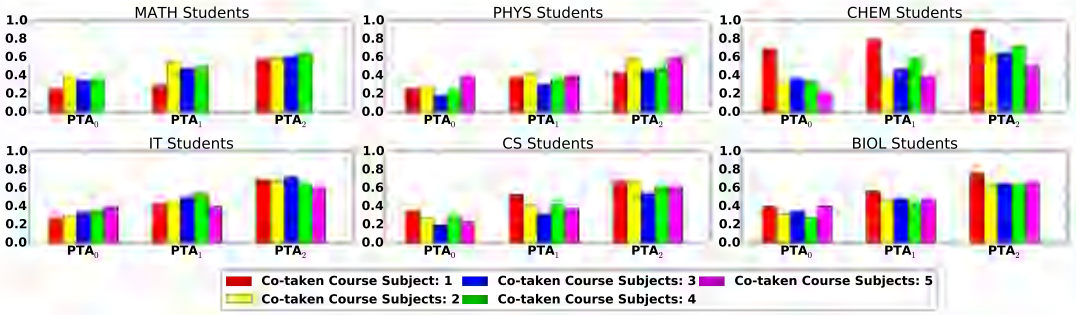

Figure 5: PTA Results for Different Number of Co-taken Course Subjects on FTF students

co-taken courses. This observation suggests that CKCC is able to leverage stronger influence of co-taken courses to improve its performance. However, for PHYS and CS majors, CKCC achieves better performance with 2, 3 or 6+ co-taken courses than with 4 or 5 co-taken courses. We postulate that this is due to the characteristics of courses chosen within a term and their content. These results also indicate that CKCC is able to model co-taken courses' influence despite of the number of the co-taken courses.

6.5 Performance on Different Numbers of Co-taken Course Subjects

In this section, we extract each course's subject and test the CKCC model on different data subgroups with different number of co-taken course subjects in a term. The reason we conduct this experiment is because we assume that courses with the same subject tend to have relevant knowledge components. Students who have co-taken courses from many different subjects may have wide knowledge diversity. This experiment aims to test the performance of CKCC in terms of co-taken course subjects.

Specifically, we take the students in the test set and divide them into five groups: students who take courses from $\{1,2,3,4,5\}$ subjects in a term. Since there are few students co-taking courses from 6+ subjects, we exclude these students in our experiment. We perform this group of experiment on each major for both FTF and TR students, respectively. For the sake of page limit, we only show the results for FTF students. Figure 5 shows the experimental results in terms of PTA_0, PTA_1 and PTA_2. The results show that CKCC have different prediction results regarding the number of co-taken course subjects for different majors. For example, for CHEM, CS and BIOL majors, the performance of the CKCC model on PTA has the best performance with 1 co-taken course subject than other subgroups. This observation suggests that CKCC is able to model co-taken courses' influence better with less knowledge diversity in a term. However, for IT major, CKCC achieves better performance with more co-taken course subjects. And for MATH and PHYS majors, CKCC has better performance on 2 or 5 co-taken course subjects than other subgroups. We assume that this is affected by the characteristics of different majors. Moreover, for MATH and IT major, the PTA results don't vary much comparing to CHEM and BIOL majors. This illustrates that for some majors, students may take courses from several subjects at a term, and the CKCC model can still well capture the co-taken courses' influence.

7. SIGNIFICANCE AND IMPACT

To highlight the use-case scenario of the developed next term grade prediction approach using co-taken courses, we ran a simulated case study. Having demonstrated the prediction accuracy of these proposed models, the objective of this case study is to highlight the strengths of the proposed models in helping students to select courses in the future term. Implicitly we want to provide students information about their workload (or change in their overall grades) by addition of one or more courses within the next term.

Specifically, we extract two pairs of popular co-taken courses: BIOL311 ("General Genetics") and CHEM313 ("Organic Chemistry"), MATH213 ("Analytic Geometry and Calculus II") and PHYS260 ("University Physics"), and conduct a study to illustrate how our model can help plan students' course selections or allocate the necessary study time. Take the course pair BIOL311 and CHEM313 as an example. We extract the students who take course BIOL311 and CHEM313 together in a term. We predict students' performance on course BIOL311 using the CKCC model. We then eliminate course CHEM313 from our data set and predict the grade on course BIOL311 again using the CKCC model. Comparing the

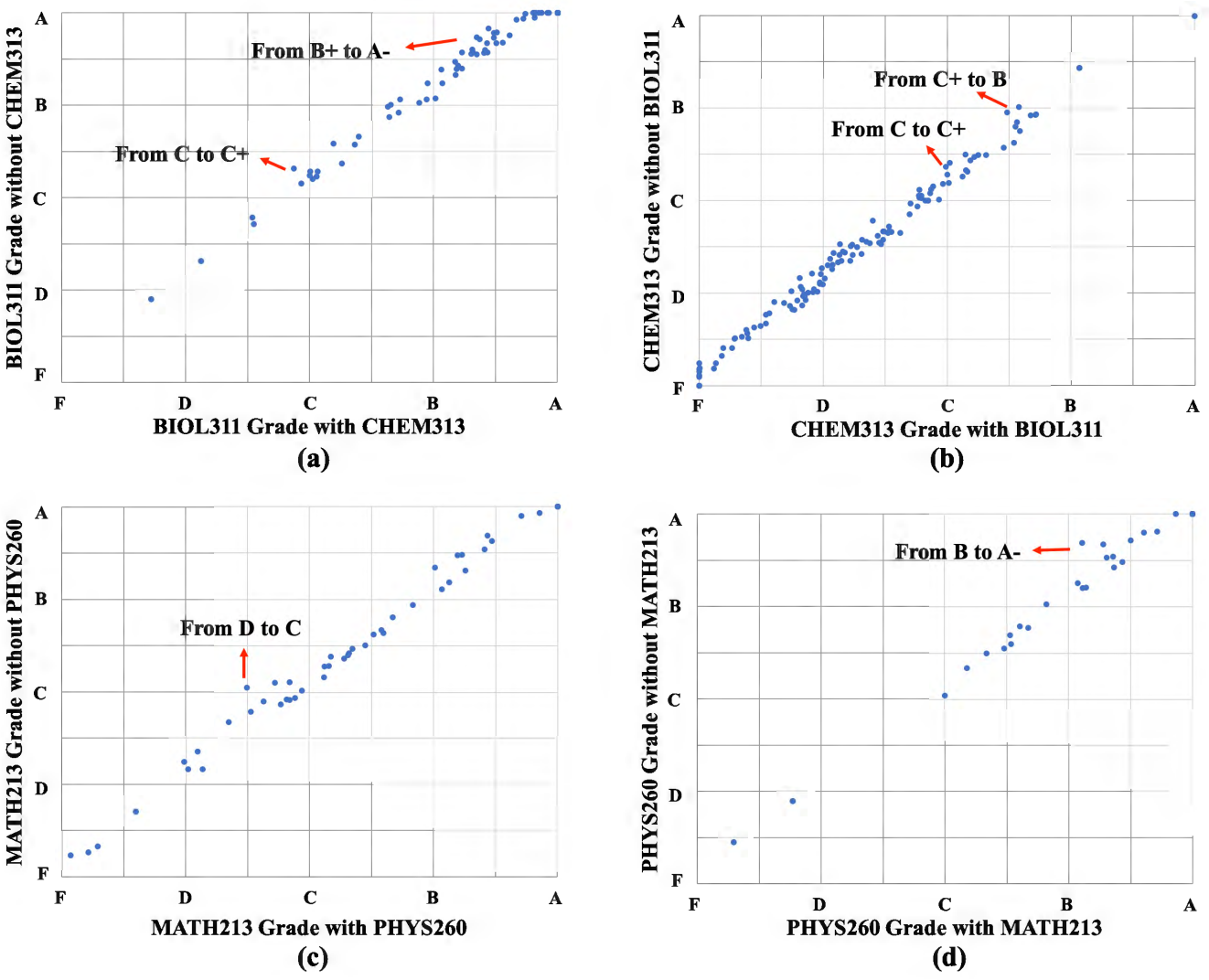

Figure 6: Comparison Results on the Co-taken Course Influence

predicted grades helps determine if the two courses should be taken together within the same term or not. The sampled students have a total of five courses that they are enrolled in for the particular term. The comparison results are shown in Figure 6 (a). It is a scatter plot of predicted grades for a student where the x-axis shows the performance on course BIOL311 co-taken with the CHEM313 and the y-axis is the performance on course BIOL311 with course CHEM313 removed. We have conducted the same experiments for other course pairs using the same protocol and shown these results in Figure 6 (b), (c) and (d).

In general, students' performance will get better with the other course eliminated due to the reduction in workload. However, different students get affected differently by the additional course. For students who take BIOL311 and CHEM313, some of them will have improvement in BIOL311 grades if they do not enroll for CHEM313 in the same semester. On the other hand, some students will not have any change in their grades for BIOL311 based on course CHEM313 (the plotted results along the diagonal). Similar trends can be observed in Figure 6 (b), (c) and (d) as well. In the Figure 6, we also highlight different cases where students grade changes with the removal of the particular course. Using this information, students can plan the set of courses that they might enroll for in the next term, and allocate study time accordingly.

8. CONCLUSION AND FUTURE WORK

In this work, we propose grade prediction models that incorporate both cumulative knowledge and co-taken courses (CKCC) to predict students' performance in the next term. The proposed models consider both cumulative knowledge a student has acquired after taking a series of courses in the passing terms, and the co-taken courses the student plans to take in the next term. Our experimental results on a dataset from George Mason University shows that the proposed models significantly outperform other competitive baselines over most the datasets for the task of next-term grade prediction. Moreover, our experimental results show that the proposed model is able to capture strong influence of co-taken courses to improve its grade prediction performance. Furthermore, we ran a simulated case study to illustrate how our proposed model can help students in course selection for the future term.

In the future, we plan to take into account additive factors, such as instructor, student's academic level and course's difficulty level along with co-taken course information, in order to achieve more accurate grade prediction results. We hope such a grade prediction system can not only help students select courses, finish their study at college but also guide them in career planning in the future.

9. ACKNOWLEDGMENTS

Funding was provided by NSF Grant, 1447489.

10. REFERENCES

[1] Asmaa Elbadrawy and George Karypis. Domain-aware grade prediction and top-n course recommendation. *Boston, MA, Sep*, 2016.

[2] Asmaa Elbadrawy, Agoritsa Polyzou, Zhiyun Ren, Mackenzie Sweeney, George Karypis, and Huzefa Rangwala. Predicting student performance using personalized analytics. *Computer*, 49(4):61–69, 2016.

[3] Asmaa Elbadrawy, Scott Studham, and George Karypis. Personalized multi-regression models for predicting students performance in course activities. *UMN CS 14-011*, 2014.

[4] Chein-Shung Hwang and Yi-Ching Su. Unified clustering locality preserving matrix factorization for student performance prediction. *IAENG Int. J. Comput. Sci*, 42(3):245–253, 2015.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Severin Klingler, Rafael Wampfler, Tanja Käser, Barbara Solenthaler, and Markus Gross. Efficient feature embeddings for student classification with variational auto-encoders.

[7] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

[8] Sara Morsy and George Karypis. Cumulative knowledge-based regression models for next-term grade prediction. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 552–560. SIAM, 2017.

[9] Michelle Parker. Advising for retention and graduation. 2015.

[10] Štefan Pero and Tomáš Horváth. Comparison of collaborative-filtering techniques for small-scale student performance prediction task. In *Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering*, pages 111–116. Springer, 2015.

[11] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pages 505–513, 2015.

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

[13] Zhiyun Ren, Xia Ning, and Huzefa Rangwala. Grade prediction with temporal course-wise influence. *arXiv preprint arXiv:1709.05433*, 2017.

[14] Zhiyun Ren, Xia Ning, and Huzefa Rangwala. Ale: Additive latent effect models for grade prediction. *arXiv preprint arXiv:1801.05535*, 2018.

[15] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B Kantor. Recommender systems handbook., 2011.

[16] Arjun Sharma, Arijit Biswas, Ankit Gandhi, Sonal Patil, and Om Deshmukh. Livelinet: A multimodal deep recurrent neural network to predict liveliness in educational videos. In *EDM*, pages 215–222, 2016.

[17] Jill M Simons. *A National Study of Student Early Alert Models at Four-Year Institutions of Higher Education.* ERIC, 2011.

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

[19] Mack Sweeney, Huzefa Rangwala, Jaime Lester, and Aditya Johri. Next-term student performance prediction: A recommender systems approach. *arXiv preprint arXiv:1604.01840*, 2016.

[20] Nguyen Thai-Nghe, Lucas Drumond, Artus Krohn-Grimberghe, and Lars Schmidt-Thieme. Recommender system for predicting student performance. *Procedia Computer Science*, 1(2):2811–2819, 2010.

[21] Andreas Zell. *Simulation neuronaler netze*, volume 1. Addison-Wesley Bonn, 1994.