IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 17, 2019, accepted February 2, 2019, date of publication February 7, 2019, date of current version February 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2897979

A Score Prediction Approach for Optional Course Recommendation via Cross-User-Domain Collaborative Filtering

LING HUANG[1,2], (Student Member, IEEE), CHANG-DONG WANG[1,2,3], (Member, IEEE), HONG-YANG CHAO[1,2], (Member, IEEE), JIAN-HUANG LAI[1,2], (Senior Member, IEEE), AND PHILIP S. YU[4,5], (Fellow, IEEE)

[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China
[2]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Sun Yat-sen University, Guangzhou 510006, China
[3]Guangdong Engineering Research Center of Smart Vocational Education and Big Data, Guangzhou 511400, China.
[4]Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
[5]Institute for Data Science, Tsinghua University, Beijing 100084, China

Corresponding author: Chang-Dong Wang (changdongwang@hotmail.com)

This work was supported in part by the NSFC under Grant 61672548, Grant U1611461, and Grant 61672313, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2016A030306014, in part by the Tip-top Scientific and Technical Innovative Youth Talents of Guangdong Special Support Program under Grant 2016TQ03X542, in part by the NSF under Grant IIS-1526499, Grant IIS-1763325, and Grant CNS-1626432, and in part by the Guangdong Engineering Research Center of Smart Vocational Education and Big Data under Grant 706049150203.

ABSTRACT Optional course selection is a critical activity for college students due to a large number of available but unfamiliar optional courses. Improper selection of optional courses would seriously affect the students' optional course achievements, which enforces students to drop out the improperly selected optional courses. Therefore, there is an urgent need to develop an optional course recommendation system. In this paper, we develop an optional course recommendation system based on score prediction. In particular, a novel cross-user-domain collaborative filtering algorithm is designed to accurately predict the score of the optional course for each student by using the course score distribution of the most similar senior students. After generating the predicted scores of all optional courses, the top t optional courses with the highest predicted scores without time conflict will be recommended to the student. The extensive experiments have been conducted to evaluate the effectiveness of the proposed method, and the results show that the proposed method is able to accurately recommend optional courses to students who will achieve relatively high scores.

INDEX TERMS Optional course recommendation, score prediction, collaborative filtering, personalized learning.

I. INTRODUCTION

As an important feature of course management system, optional course selection is a critical activity for college students. Before each semester, college students are required to select several optional courses they are interested in from a large number of optional courses for the coming semester [1]. To make the point clear, a survey is conducted on 25 randomly selected universities in China. Figure 1(a) plots the fractions of universities offering different numbers of university-level optional courses in each semester. As shown in the figure, only one university (i.e. 4%) offers less than 300 university-level optional courses in each semester. Most of the universities (i.e. 60%) offer 300-400 university-level optional courses. Moreover, there are 3 universities (i.e. 12%) offer more than 500 university-level optional courses. Figure 1(b) plots the fractions of majors offering different numbers of major-level optional courses in each semester. In this figure, 500 majors are randomly selected from the aforementioned 25 universities. From the figure, we can see that most of the majors (i.e. 31%) offer 20-25 major-level optional courses. Only 15 majors offer less than 10 major-level optional courses. Moreover, there are 25 majors (i.e. 5%) offering more than 30 major-level optional courses. According to the above survey, before each semester, students are facing the issue of selecting only a small number of optional

The associate editor coordinating the review of this manuscript and approving it for publication was Shirui Pan.

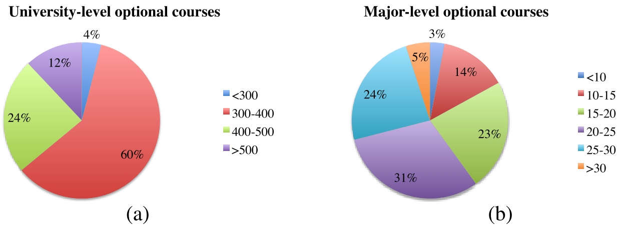

FIGURE 1. A survey on the number of optional courses. (a) Universities. (b) Majors.

courses from a relatively large number of university-level optional courses and major-level optional courses, which is a challenge task due to being unfamiliar with the optional courses [2].

According to [3], the students' achievement goals are the best predictors for success of a course, while improperly selecting optional course would seriously lead to the dropout of optional courses. Figure 2(a) plots the distribution of optional course dropout rate of 1000 optional courses randomly selected from the aforementioned 25 universities. From the figure, we can see that more than 110 optional courses (11%) have the dropout rate as high as 30%, i.e. more than 30% students who have selected each of these courses before the corresponding semester will dropout these courses after a few weeks' trial. Most of the optional courses (42%) have dropout rate between 20% and 30%, which is far larger than that reported in the CS1 course at Helsinki University of Technology [1].

Additionally, we also conduct a survey on 1000 students randomly selected from the aforementioned 25 universities, and analyze the distribution of students dropping out optional courses and the main underlying reasons. Figure 2(b) plots the distribution of students dropping out optional courses. From the figure, we can see that among 1000 students, 48% students have experiences of dropping out at least one optional courses. And there are 13% students having experiences of dropping out more than one optional courses. When analyzing the main reasons why they drop out optional courses, as shown in Figure 2(c), feeling difficulty of getting relatively high scores is the main reason. It coincides with the discovery by Kinnunen and Malmi [1], who conducted a study on the reasons for students' quitting the CS1 course at Helsinki University of Technology. It was discovered that one of the most important factors of quitting the CS1 course is the perceived difficulty of the course.

From the above survey and discussion, one safe conclusion can be drawn that, due to the large number of available but unfamiliar optional courses, optional course selection is a critical activity for college students. Additionally, improperly selecting optional courses would seriously affect the students' optional course achievements, which enforces students to drop out these improperly selected optional courses. Therefore, it is in urgent need of developing an optional course recommendation system that is able to recommend optional courses to students who will achieve relatively high scores.

In the literature of computer assisted education or educational data mining [4], some efforts have been made in analyzing course selection [1], [5], analyzing course achievements [6], [7] and developing recommendation based system for personalized learning [8]–[10]. However, to our best knowledge, there is still a lack of work on optional course recommendation based on score prediction.

In this paper, we develop an optional course recommendation system based on score prediction. The proposed approach consists of two main phases, namely score prediction based on cross-user-domain collaborative filtering and optional course recommendation based on the predicted scores and the curriculum schedule. In the first phase, inspired by the previous work on collaborative filtering [11]–[14], a novel cross-user-domain collaborative filtering (CUDCF) algorithm is designed to accurately predict the score of the optional course for each student by using the course score distribution of the most similar senior students. The underlying rationale is that, students with similar scores in the previous courses will generally obtain similar scores in the subsequent courses. Specifically, for predicting the score of each optional course for each student, a small set of senior students who have already enrolled on the target optional course and have the most similar past score distribution to the student will be discovered by means of Pearson correlation coefficient of the previous course scores. And then the predicted score will be generated based on the deviation-aware weighted average of the scores from the small set of the most similar senior students. Notice that the proposed CUDCF algorithm is quite different from the existing cross-domain collaborative filtering (CDCF) algorithms [12], [15]. In particular, the existing CDCF algorithms usually share the common user domain but adopt different item domains. However, CUDCF shares the common item domain (i.e. the course domain) but adopts different user domains (i.e. the junior student domain and the senior student domain). This may inspire the development of more cross-user-domain collaborative filtering algorithms in different recommendation tasks. In the second phase, for each student, according to the predicted scores of all optional courses and the curriculum schedule, the top t optional courses with the highest predicted scores without time conflict will be recommended to the student.

Extensive experiments will be conducted to evaluate the effectiveness of the proposed method, the results of which show that the proposed method is able to achieve relatively high *average hit rate* and *average accuracy*. That is, the proposed method is able to accurately recommend optional courses to students who will achieve relatively high scores.

In summary, the main contributions of this paper are as follows.

1) We develop an optional course recommendation system based on score prediction, which consists of two main phases, namely score prediction based on cross-user-domain collaborative filtering and optional course

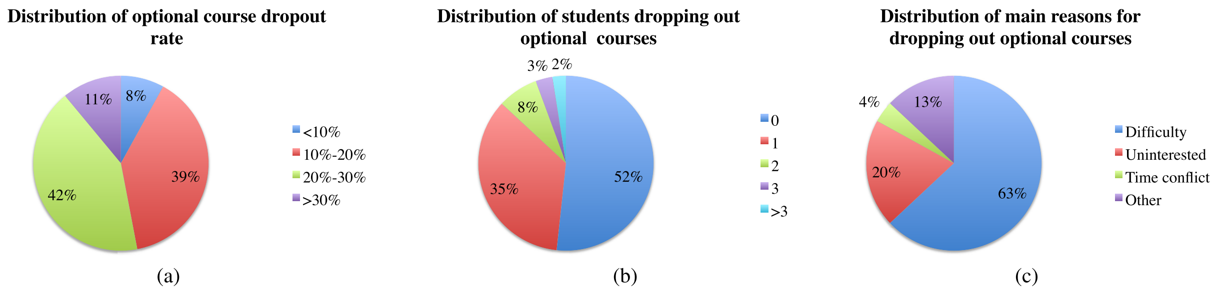

FIGURE 2. A survey on the dropout of optional courses. (a) Optional courses. (b) Students. (c) Main reasons.

recommendation based on the predicted scores and the curriculum schedule.

2) Some analysis is conducted to show the underlying characteristic of the compulsory courses and the optional courses of college students.
3) A new cross-user-domain collaborative filtering (CUDCF) algorithm is designed to accurately predict the score of the optional course for each student.
4) Extensive experiments are conducted to confirm the effectiveness of the proposed approach.

The remainder of this paper will be organized as follows. In section II, we will briefly review the related work, including course selection, course achievement and recommendation based personalized learning. In section III, we will describe the datasets that are used in this paper. In section IV, we will describe in detail the proposed optional course recommendation system. In section V, extensive experiments will be conducted to evaluate the effectiveness of the proposed method, by which some insightful discovery will be discussed. Finally, we will draw the conclusion and describe the future work in section VI.

II. RELATED WORK

A. COURSE SELECTION

Some work has been done on analyzing the college students' course selection [1], [5], [16]. Kinnunen and Malmi [1] conducted a study on the reasons for students' quitting the CS1 course at Helsinki University of Technology. It was discovered that one of the most important factors of quitting the CS1 course is the perceived difficulty of the course. This study suggests that it is a suitable strategy to recommend courses that student will obtain relatively higher course achievement. In [16], a survey was conducted to analyze why students choose one optional course. However, it is limited to only the case of teratology. Kardan *et al.* [5] conducted a study on the factors influencing online course selection of college students in the context of e-learning using neural network. However, it is limited to the e-learning which may significantly differ from the face-to-face learning in the traditional college education [17], [18]. Additionally, there is still a lack of study on how the course selection would impact the students' college education, e.g. the course achievements.

B. COURSE ACHIEVEMENT

In the literature, course achievements have been widely studied from various perspectives [6], [7], [19]–[27]. Dodge *et al.* [6] proposed to identify students at-risk of failing in two high enrollment courses by means of learning analytics methods. Shell *et al.* [22] investigated the relationship between the students' entering motivation and their subsequent course achievement and retention in college CS1 courses. In [23], a study was conducted to investigate the relationship between students' perceptions of the flipped course model and their self-regulated learning behaviors, and then analyze how the course achievement can be affected accordingly. Conijn *et al.* [24] conducted a study on predicting student performance from data collected in Learning Management Systems (LMSs), where 17 blended courses with 4,989 students in a single institution using Moodle LMS are analyzed. In [26], a study was conducted which uses Classification and Regression Trees (CART) for analyzing student course activity data and predicting student course achievement. A predictive model is designed to classify students into pass/fail categories for further early identifying struggling students. In [19], a study was conducted to analyze the impact of dynamic web technologies on student academic achievement in a problem-based collaborative learning environment. Moore and Rutledge [20] investigated how the usage of guided learning can affect students on online course achievement and retention. Despite success, relatively less effort has been made in analyzing the relationship between course achievements of different chronological courses. In [25], logistic regression was used to analyze the effectiveness of the combination of scientific reasoning scores and ACT mathematics scores to predict students' future achievement. Ren *et al.* [21] developed a factorization-based approach for score prediction called Matrix Factorization with Temporal Course-wise Influence that incorporates course-wise influence effects and temporal effects for grade prediction. However, there is still a lack of work on optional course recommendation based on the predicted course achievement.

C. RECOMMENDATION BASED PERSONALIZED LEARNING

In the literature, many efforts have been made in applying recommendation in various personalized learning

TABLE 1. Statistics of the datasets. The third column lists the student number of each major in class 2013 and class 2014. The fourth and fifth columns list the total number of compulsory courses and optional courses of each major in class 2013 and class 2014 respectively. The last four columns list the mean value and the standard deviation of the number of compulsory courses and optional courses enrolled by each student.

Majors	Classes	#Student	#Course		#Course per student			
			Compulsory	Optional	Compulsory		Optional	
					Mean	Std. Dev.	Mean	Std. Dev.
Major1	2013	119	29	64	25.98	0.18	24.57	2.00
	2014	88	29	52	23.86	1.51	24.39	3.30
Major2	2013	49	27	50	25.87	0.63	22.91	2.80
	2014	59	24	53	24.00	0.00	23.74	0.78
Major3	2013	13	27	40	25.30	1.49	21.76	6.72
	2014	40	26	42	24.00	0.32	22.85	3.85
Major4	2013	30	29	39	29.00	0.00	18.63	1.54
	2014	30	24	34	23.23	4.20	17.30	3.45
Major5	2013	34	26	49	25.94	0.24	19.08	3.02
	2014	28	25	39	24.96	0.19	17.60	2.50
Major6	2013	76	26	56	25.75	0.96	16.94	3.18
	2014	101	25	46	24.95	0.41	17.20	1.90
Major7	2013	144	34	70	25.75	1.11	17.34	3.24
	2014	119	26	47	24.98	0.18	17.78	1.12
Major8	2013	116	26	60	25.94	0.56	17.82	1.72
	2014	120	26	41	24.96	0.37	17.59	1.54

scenarios [2], [8]–[10], [28]–[37]. Ibrahim *et al.* [2] developed a framework of an ontology-based hybrid-filtering system, which integrates information from multiple sources based on hierarchical ontology similarity so as to personalize course recommendations that will match the individual needs of students. In [9], an e-learning system with a recommendation module was developed, which recognizes varying patterns of learning style and learning habits of learners by learning style identification and server log mining. Therefore, it is able to adapt to learners' interests and knowledge levels. Mangina and Kilbride [10] developed a document recommendation system in an online e-learning environment based on user modeling, information retrieval/extraction and collaborative filtering. Chatti *et al.* [28] studied a number of different tag-based collaborative filtering recommendation algorithms for dealing with the information overload problem in personal learning environment. To address the similar issue in personal learning environment, Salehi *et al.* [30] developed a recommendation framework by using learner preference tree and genetic algorithm. In [31], a knowledge-based strategy was designed for recommending educational resources in open educational repositories. The strategy includes the following two key aspects, namely the description of the educational resources and the contextual information about the resource users. In [33], a set of practical guidelines for designing and evaluating educationally oriented recommendation has been studied from the perspectives of both educational goals and learners' preferences.

Different from the aforementioned work, Hoic-Bozic *et al.* [8] studied the usage of recommendation and web 2.0 tools for the blended learning model consisting of face-to-face learning and e-learning. Zheng *et al.* [32] developed a new hybrid, trust-based recommender system to mitigate the bounded rationality and metacognition issues for online communities of practice. In [34], an object-object similarity measure was developed by considering the usage context for boosting recommendation in technology enhanced learning. For supporting teachers in selecting learning objects from existing learning object repositories in a unified manner, Sergis and Sampson [35] developed a recommender systems. Yao [36] proposed to integrate a context-aware technique into the personalized recommendation learning so as to recommend the suitable learning materials for a learner at a specific location. Despite significant success of recommendation in various personalized learning applications, the efficacy of recommendation in college students' optional course selection remains largely unknown.

III. DATASETS

In this section, we will describe the datasets that are used in this paper, i.e. a database consisting of 52311 course-score records from School of Data and Computer Science, Sun Yat-sen University. The database consists of 1166 students belonging to 8 different majors in class 2013 and class 2014. Each major is regarded as a separate testing dataset. For class 2013, the course-score records in each semester from 2013 Fall to 2017 Spring are used, covering all the 8 semesters; While for class 2014, the course-score records in each semester from 2013 Fall to 2017 Fall are used, covering 7 semesters (due to the reason that the course-score records in 2018 Spring are not yet available). Table 1 summarizes the statistics of the 8 testing datasets. From the fifth column, we can see that the total number of optional courses enrolled by one class in a major varies from 34 to 70. By comparing the last four columns, we can see that, in general, the number of optional courses enrolled by each student has much larger diversity than the number of compulsory courses enrolled by each student at the same major in the same class.

That is, the standard deviations of the number of optional courses enrolled by each student are larger than those of compulsory courses except at Major 4 in class 2014, which is a special major in China, namely national defense — the educational policy may change from year to year.

Moreover, Figure 3 plots the distribution of course number enrolled by each student. This figure further verifies the fact that the number of optional courses enrolled by each student has much larger diversity than the number of compulsory courses enrolled by each student at the same major in the same class. For instance, in Figure 3(a), most of the students at Major 1 in class 2013 have enrolled on 26 compulsory courses, while there are many students enrolling on 24, 25 and 26 optional courses. Notice that, from the figure, there are some students enrolling on a very small number of compulsory or optional courses (i.e. the bar corresponding to the numbers smaller than 10), who are "academically abnormal" students omitted in the current study.

From the above basic statistics of the 8 testing datasets, one safe conclusion can be drawn that the college students have many choices to select optional courses. And for the students at one major in one class, the number of optional courses they have enrolled on is much larger than that of compulsory courses.

IV. THE PROPOSED METHOD

In this section, we will describe in detail the proposed method, which consists of two main phases, namely score prediction based on cross-user-domain collaborative filtering and optional course recommendation based on the predicted scores and the curriculum schedule.

A. CROSS-USER-DOMAIN COLLABORATIVE FILTERING

A novel cross-user-domain collaborative filtering (CUDCF) algorithm is designed to accurately predict the score of the optional course for one student by using the course score distribution of the most similar senior students.

Let $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ denote the set of n students, who are required to select several optional courses from a set of m optional courses $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$. For predicting the score of each optional course $c \in \mathcal{C}$ for each student $s \in \mathcal{S}$, a small set of senior students who have already enrolled on course c and have the most similar previous score distribution to student s will be discovered by means of Pearson correlation coefficient. Let $\bar{\mathcal{S}}_c$ denote the set of senior students who have already enrolled on course c. For any senior student $\bar{s} \in \bar{\mathcal{S}}_c$, the following Pearson correlation coefficient is used to measure the course score similarity between student s and the senior student $\bar{s}$,

$$Sim(s, \bar{s}) = \frac{\sum_{i\in\mathcal{C}_{s\bar{s}}} \left(r_{si} - \text{mean}(r_s)\right)\left(r_{\bar{s}i} - \text{mean}(r_{\bar{s}})\right)}{\sqrt{\sum_{i\in\mathcal{C}_{s\bar{s}}} \left(r_{si} - \text{mean}(r_s)\right)^2}\sqrt{\sum_{i\in\mathcal{C}_{s\bar{s}}} \left(r_{\bar{s}i} - \text{mean}(r_{\bar{s}})\right)^2} \tag{1}$$

where $\mathcal{C}_{s\bar{s}}$ denotes the courses (including compulsory courses and optional courses) that are enrolled by both students s and $\bar{s}$, r_{si} and $r_{\bar{s}i}$ denote the score of course i by students s and $\bar{s}$ respectively, mean(r_s) and mean$(r_{\bar{s}})$ denote the average scores of courses enrolled by students s and $\bar{s}$ respectively.

The top k most similar senior students from $\bar{\mathcal{S}}_c$ will be selected according to the course score similarity $Sim(s, \bar{s}), \forall \bar{s} \in \bar{\mathcal{S}}_c$, denoted as $\mathcal{N}_{s,c} \subset \bar{\mathcal{S}}_c$. Accordingly, the score of the optional course c by student s can be predicted as follows,

$$r_{sc} = \frac{\sum_{\bar{s}\in\mathcal{N}_{s,c}} Sim(s, \bar{s}) r_{\bar{s}c}}{\sum_{\bar{s}\in\mathcal{N}_{s,c}} Sim(s, \bar{s})} \tag{2}$$

That is, the predicted score of course c for student s is obtained by the weighted average score of course c enrolled by the top k most similar senior students.

The underlying rationale is that, students with similar scores in the previous courses will generally obtain similar scores in the subsequent courses. To verify the above fact, the courses enrolled by class 2013 and class 2014 are separated into two parts: the previous courses and the subsequent courses.

1) For class 2013, the courses in the first two years, i.e. in semesters of 2013 Fall, 2014 Spring, 2014 Fall and 2015 Spring are regarded as the previous courses, denoted as Class2013PreCourses; While the courses in the last two years, i.e. in semesters of 2015 Fall, 2016 Spring, 2016 Fall and 2017 Spring are regarded as the subsequent courses, denoted as Class2013SubCourses.
2) For class 2014, the courses in the first two years, i.e. in semesters of 2014 Fall, 2015 Spring, 2015 Fall and 2016 Spring are regarded as the previous courses, denoted as Class2014PreCourses; While the courses in the last two years, i.e. in semesters of 2016 Fall, 2017 Spring and 2017 Fall are regarded as the subsequent courses, denoted as Class2014SubCourses.

For each student in class 2014, $s \in \mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ (i.e. the junior student), the top k most similar senior students (in class 2013) $\mathcal{N}_s$ who have enrolled on at least one identical previous courses and at least one identical subsequent courses to student s are selected according to the Pearson correlation coefficient similarity. Then we will calculate the score similarity between Class2014SubCourses and Class2013SubCourses, as well as the score similarity between Class2014PreCourses and Class2013PreCourses. In particular, we will calculate the following values.

1) $Sim_s^{Pre} = \text{mean}_{\bar{s}\in\mathcal{N}_s}(Sim^{Pre}(s, \bar{s}))$ where $Sim^{Pre}(s, \bar{s})$ denotes the Pearson correlation coefficient between the scores of the Class2014PreCourses obtained by student s and the scores of the Class2013PreCourses obtained by $\bar{s} \in \mathcal{N}_s$.
2) $Sim_s^{Sub} = \text{mean}_{\bar{s}\in\mathcal{N}_s}(Sim^{Sub}(s, \bar{s}))$ where $Sim^{Sub}(s, \bar{s})$ denotes the Pearson correlation coefficient between the scores of the Class2014SubCourses obtained by

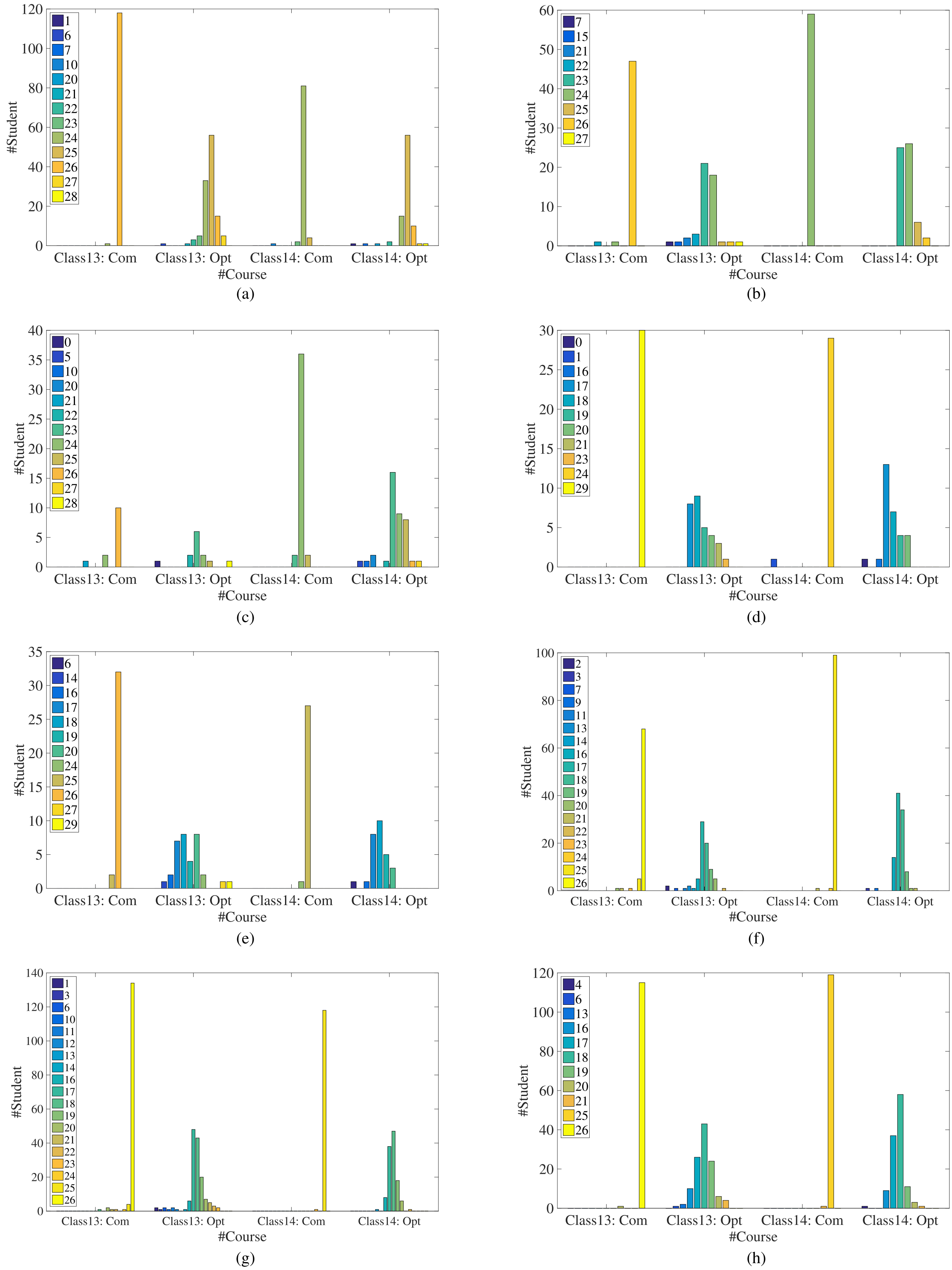

FIGURE 3. Distribution of course number enrolled by each student, where "Com" and "Opt" stand for "Compulsory" and "Optional" respectively.

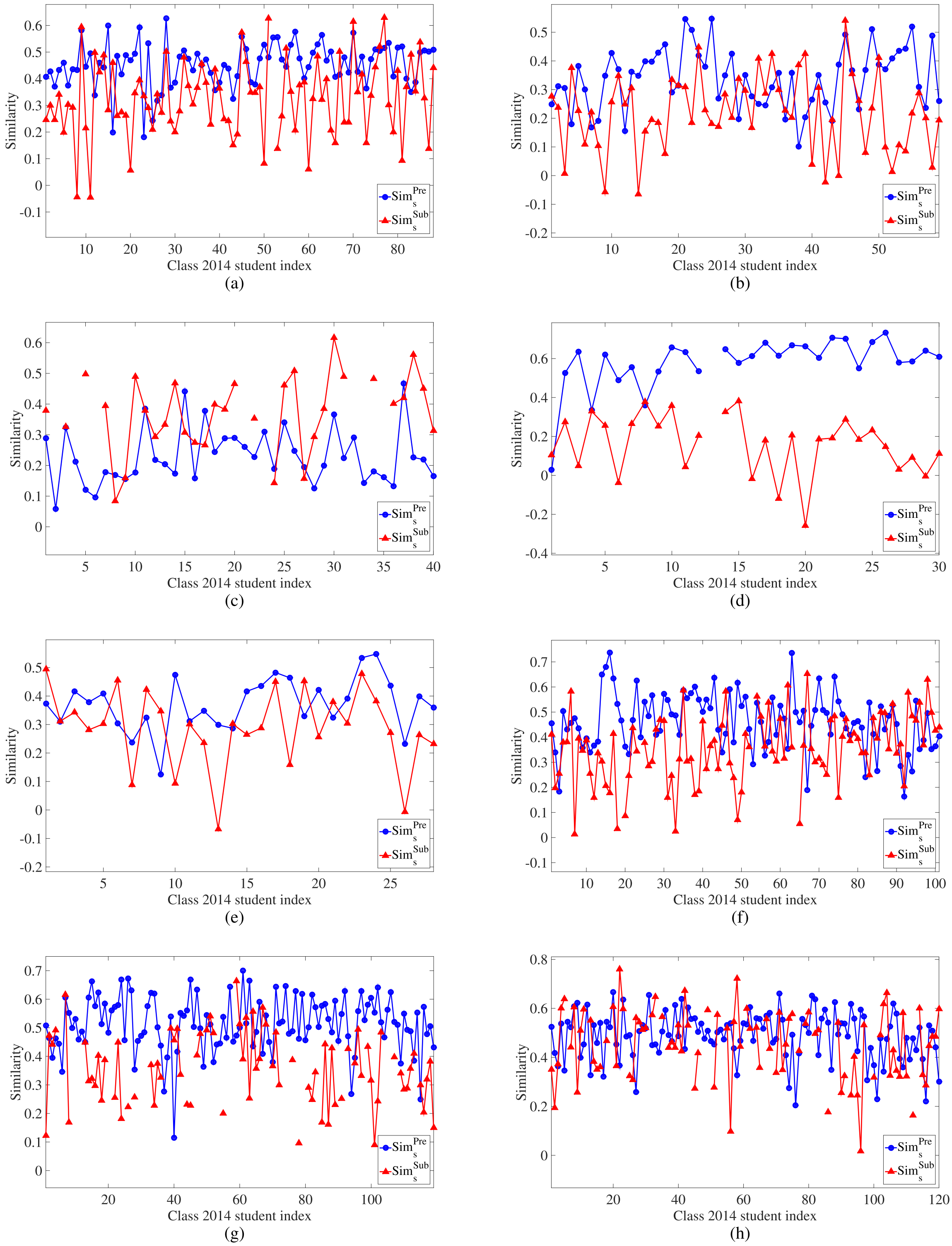

FIGURE 4. **Comparison between the values of Sim_s^{Pre} and Sim_s^{Sub} for each student $s \in \mathcal{S}$ on the 8 testing datasets. (a) Major1. (b) Major2. (c) Major3. (d) Major4. (e) Major5. (f) Major6. (g) Major7. (h) Major8.**

student s and the scores of the Class2013SubCourses obtained by $\bar{s} \in \mathcal{N}_s$.

Then we will compare the values of Sim_s^{Pre} and Sim_s^{Sub} for each student $s \in \mathcal{S}$. The values for the 8 majors (8 testing datasets) are plotted in each subfigure in Figure 4. From the figure, we can see that in most cases except Major4, the values of Sim_s^{Pre} and Sim_s^{Sub} are very close to each other for most of the students, which verifies that students with similar scores in the previous courses will generally obtain similar scores in the subsequent courses. The exception of Major4 is mainly due to that it is a special major, namely national defense, in which students have very large diversity on course selection, training and evaluation according to their specific skills. Additionally, it should be noticed that, at each major, there exist a small amount of students (less than 5%) in class 2014 who have relatively low Sim_s^{Sub} that are very close to or even below 0. This is because they are the ''academically abnormal'' students who cannot achieve expected scores in the subsequent courses and hence exhibit poor correlation with the senior students having similar previous course score distributions. Discovery of ''academically abnormal'' students is another interesting topic in our future work.

It should be noticed that, different students may have varying score deviations, i.e. the score deviation compared with the average score among all students. That is, some students may be likely to obtain relatively higher scores while some others may be likely to obtain relatively lower scores. Similarly, different courses may have varying score deviations, i.e. the score deviation compared with the average score among all courses. That is, some courses may be likely to have overall higher scores while some others may be likely to have overall lower scores (e.g. due to the varying difficulties).

The above prediction in Eq. 2 has taken into account neither the score deviation of student s nor the score deviation of course c. To this end, a new prediction formula will be used, which predicts the score by the deviation-aware weighted average score of course c by the top k most similar senior students,

$$r_{sc} = b_{sc} + \frac{\sum_{\bar{s} \in \mathcal{N}_{s,c}} Sim(s, \bar{s})(r_{\bar{s}c} - b_{\bar{s}c})}{\sum_{\bar{s} \in \mathcal{N}_{s,c}} Sim(s, \bar{s})} \tag{3}$$

where $b_{sc} = \mu + b_s + b_c$ denotes the baseline estimate for r_{sc} with μ being the overall mean score of all courses enrolled by all students (here since we take each major as a separate dataset, the ''all students'' indicate all students from the same major of all years), $b_s = \text{mean}(r_s) - \mu$ being the score deviation of student s and $b_c = \text{mean}(r_c) - \mu$ being the score deviation of course c. That is, in the above prediction formula, the predicted score takes into account not only the score information from the top k most similar senior students but also the score deviations of student s and course c.

B. OPTIONAL COURSE RECOMMENDATION

For each student $s \in \mathcal{S}$, after generating the predicted scores of all optional courses, i.e. r_{sc}, the top t optional courses with the highest predicted scores will be recommended to student s. However, in practical applications, it is also necessary to consider the curriculum schedule in order to avoid time conflict between recommended optional courses and compulsory courses, or time conflict among recommended optional courses.

For clarity, Algorithm 1 summarizes the main procedure of the proposed optional course recommendation approach based on the course score prediction.

Algorithm 1 The Proposed Approach

Input: The set of n junior students $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$
The set of m optional courses $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$
The course score records $\{r_{s,:}, \forall s \in \mathcal{S}\}$ of the junior students
The course score records $\{r_{\bar{s},:}, \forall \bar{s} \in \bar{\mathcal{S}}\}$ of the senior students
`Curriculum schedule`.

```
for Each junior student s ∈ S do
    Phase I: Course score prediction
    for Each optional course c ∈ C do
        Obtain the set of senior students S̄_c ⊂ S̄ who have already enrolled on course c.
        for Each senior student s̄ ∈ S̄_c do
            Compute the similarity Sim(s, s̄) between s and s̄ via Eq. (1).
        end for
        Obtain the top k most similar senior students N_{s,c} ⊂ S̄_c according to {Sim(s, s̄), ∀s̄ ∈ S̄_c}.
        Predict the score r_sc of course c by student s via Eq. (3).
    end for
    Phase II: Optional course recommendation
    Recommend the top t optional courses to student s with the highest predicted scores without time conflict according to {r_sc, ∀c ∈ C} and Curriculum schedule.
end for
```

Output: The optional course recommendation results.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. RESULTS

In this subsection, we will report the results obtained by the proposed method, namely optional course recommendation by score prediction. First of all, we will describe in detail the experimental setting used in our work. Then the experimental results will be reported and analyzed.

1) EXPERIMENTAL SETTING

In this study, to confirm the effectiveness of the proposed method in recommending optional courses, the students of class 2014 on each of the 8 datasets listed in Table 1 will be used as the testing students, where the ground-truth optional courses he/she has enrolled on are regarded as the baseline. Each major is regarded as a separate dataset. On each of the 8 datasets, for each student, the number of optional courses recommended to him/her, i.e. t is adaptively set as the ground-truth number of optional courses he/she has enrolled on. Since most of the optional courses are only provided in the last two years for undergraduate students, for the class 2014, the subsequent semesters of 2016 Fall, 2017 Spring and 2017 Fall are regarded as the testing semesters, each of which is tested separately.

The results are evaluated by comparing the predicted optional courses and the ground-truth optional courses he/she has enrolled on. In particular, on each dataset, the results are

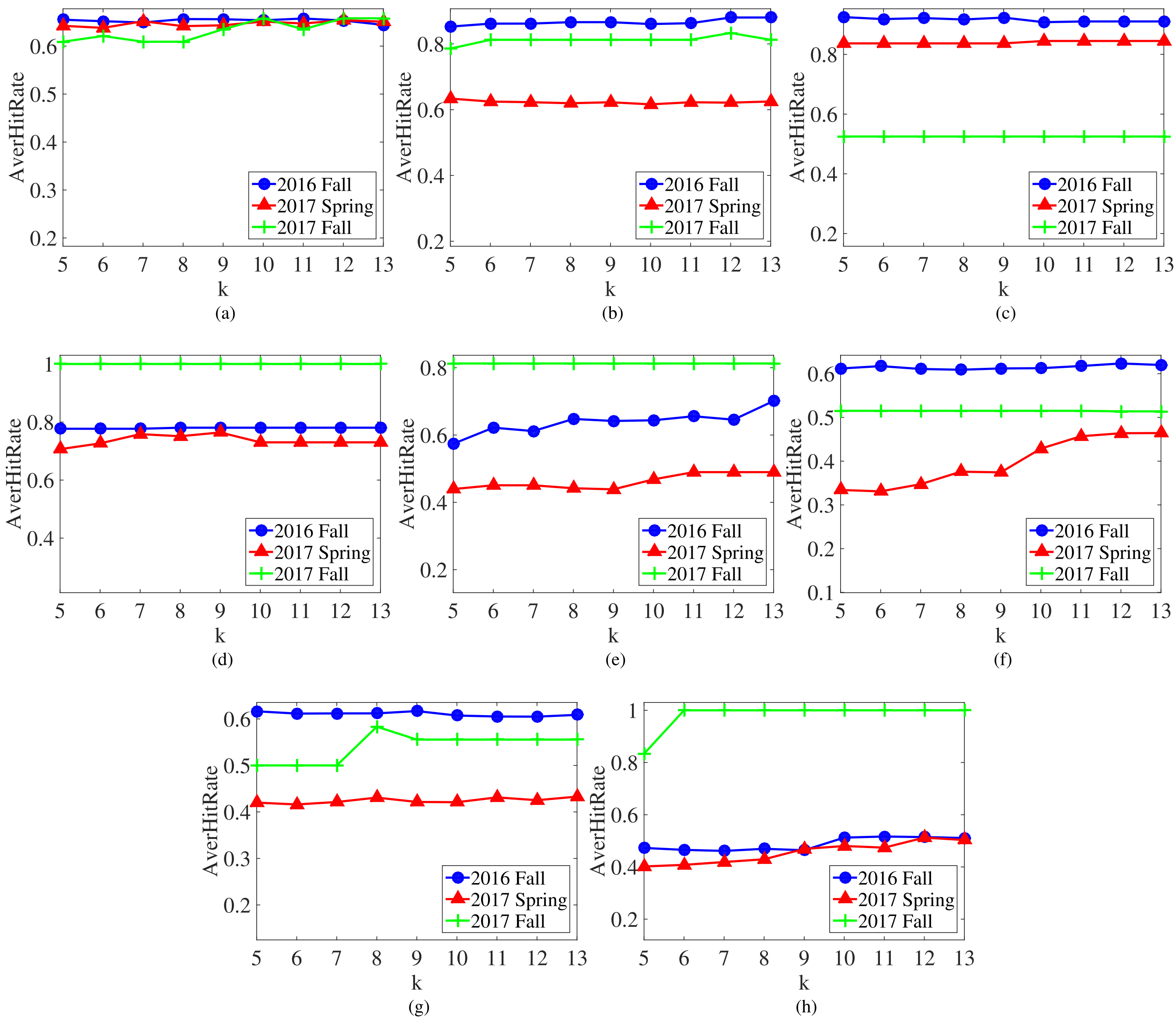

FIGURE 5. **Results in terms of average hit rate (*AverHitRate*) as a function of parameter *k*. (a) Major1. (b) Major2. (c) Major3. (d) Major4. (e) Major5. (f) Major6. (g) Major7. (h) Major8.**

evaluated by means of the *average hit rate* and the *average accuracy*. The average hit rate is defined as follows,

$$AverHitRate = \text{mean}_{s \in \mathcal{S}}(HitRate_s) \tag{4}$$

where $HitRate_s$ denotes the ratio of the number of the recommended optional courses that are enrolled by student s to the number of the recommended optional courses, which is also the ground-truth number of optional courses enrolled by student s since we set the number of recommended optional courses to be the number of ground-truth optional courses. That is, it reveals the percentage of the "accepted" optional courses recommended by the proposed system. However, notice that since the proposed system has not yet been deployed to application, i.e. the testing students have not yet used the proposed system for selecting their optional courses, a higher hit rate only indicates that the proposed system would be able to recommend more interesting optional courses to students if deployed for real-world application.

The average accuracy is defined as follows,

$$AverACC = \text{mean}_{s \in \mathcal{S}}(ACC_s) \tag{5}$$

where ACC_s is the recommendation accuracy of the "accepted" optional courses. Specifically, it is calculated as follows,

$$ACC_s = \frac{|\mathcal{C}_s^{correct}|}{|\mathcal{C}_s^{accepted}|} \tag{6}$$

where $|\mathcal{C}_s^{correct}|$ denotes the number of correctly recommended optional courses for student s, i.e. the number of "accepted" optional courses that student s has eventually

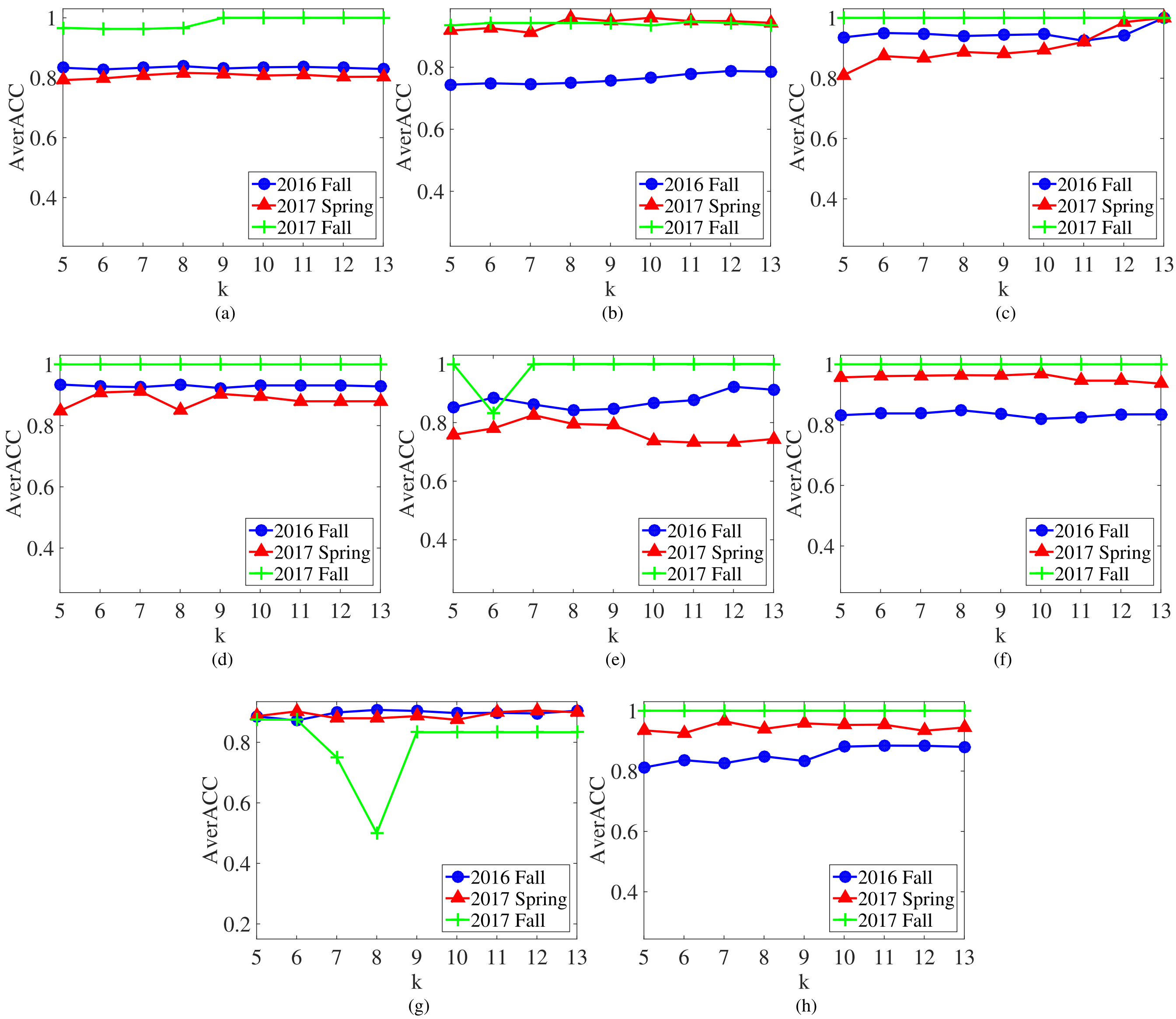

FIGURE 6. Results in terms of average accuracy (*AverACC*) as a function of parameter k. (a) Major1. (b) Major2. (c) Major3. (d) Major4. (e) Major5. (f) Major6. (g) Major7. (h) Major8.

obtained scores higher than or equal to expected (i.e. predicted). And $|C_s^{accepted}|$ denotes the number of "accepted" optional courses for student s. A higher average accuracy indicates that the proposed system would be able to correctly recommend optional courses that students can obtain relatively higher scores.

Since different numbers of top k most similar senior students would affect the recommendation results, we will report the values of *AverHitRate* and *AverACC* as a function of k on each dataset for each testing semester.

2) RESULT ANALYSIS

Figure 5 and Figure 6 report the values of average hit rate (*AverHitRate*) and average accuracy (*AverACC*) respectively as a function of parameter k on each of the 8 datasets for each of the 3 testing semesters. From Figure 5, we can see that on a relatively wide range of k, the proposed method is able to recommend most suitable optional courses for students. In particular, on the testing datasets of Major 4 and Major 8, the value of average hit rate reaches as high as 1 in semester of 2017 Fall, meaning that the recommended optional courses are exactly the same as those enrolled by students. Most importantly, on these two testing datasets of Major 4 and Major 8, the average accuracy (*AverACC*) as a function of parameter k also reaches as high as 1, meaning that 100% students at these two majors in semester of 2017 Fall can obtain scores as expected on all the recommended optional courses they have enrolled on. Combining the cases of average hit rate (*AverHitRate*) and average accuracy (*AverACC*) together, safe conclusion can be drawn that in semester of 2017 Fall,

TABLE 2. Comparison results generated by CUDCF and its variants in terms of average hit rate (*AverHitRate*).

Majors	Semesters	Variants of CUDCF			CUDCF
		CUDCF-NoScoreDev	CUDCF-NoCourseScoreDev	CUDCF-NoStudentScoreDev	
Major1	2016 Fall	0.5721	0.6009	0.6017	0.6538
	2017 Spring	0.5710	0.5994	0.6003	0.6513
	2017 Fall	0.5732	0.5918	0.6008	0.6580
Major2	2016 Fall	0.7301	0.7935	0.8012	0.8604
	2017 Spring	0.5290	0.5773	0.5844	0.6163
	2017 Fall	0.7119	0.7942	0.7952	0.8125
Major3	2016 Fall	0.7164	0.8812	0.8800	0.9078
	2017 Spring	0.7727	0.8218	0.8233	0.8449
	2017 Fall	0.4368	0.4965	0.4900	0.5250
Major4	2016 Fall	0.7348	0.7668	0.7650	0.7805
	2017 Spring	0.6526	0.6838	0.6857	0.7301
	2017 Fall	0.8616	0.9247	0.9125	1.0000
Major5	2016 Fall	0.6019	0.6380	0.6211	0.6438
	2017 Spring	0.4182	0.4368	0.4250	0.4682
	2017 Fall	0.7218	0.7506	0.7525	0.8125
Major6	2016 Fall	0.5173	0.5310	0.5352	0.6126
	2017 Spring	0.4000	0.4061	0.4015	0.4287
	2017 Fall	0.4654	0.4978	0.4900	0.5152
Major7	2016 Fall	0.5892	0.6008	0.6002	0.6073
	2017 Spring	0.4018	0.4028	0.4025	0.4209
	2017 Fall	0.5014	0.5361	0.5255	0.5556
Major8	2016 Fall	0.4768	0.5083	0.5090	0.5123
	2017 Spring	0.4117	0.4506	0.4125	0.4802
	2017 Fall	0.8808	0.9474	0.9151	1.0000

on the testing datasets of Major 4 and Major 8, a very satisfactory optional course recommendation has been made.

Another interesting phenomenon is that on most testing datasets except Major 7, the average accuracy ($AverACC$) obtained in semester of 2017 Fall is higher than those obtained in 2016 Fall and 2017 Spring. The main reason may be that for 2017 Fall, more course-score data are available for achieving a much more accurate prediction of optional course scores.

Notice that in each semester, for students at each major, they have to select about 5 optional courses from about 25 major-level optional courses and about 400 university-level optional courses. Therefore, reaching the average hit rate (*AverHitRate*) as high as 0.8 is a challenging task. However, as revealed by the above result analysis, when selecting the optional courses from the recommended optional course list, most of students would be able to obtain relatively higher scores. That is, overall, the values of average accuracy ($AverACC$) reach as high as 0.8, even 0.9 or 1 in most cases.

In order to verify the effectiveness of considering the score deviations of students and courses in CUDCF, we compare the results generated by CUDCF and its variants, namely CUDCF without considering score deviations of students (CUDCF-NoStudentScoreDev for short), CUDCF without considering score deviations of courses (CUDCF-NoCourseScoreDev for short), and CUDCF without considering score deviations of students and courses (CUDCF-NoScoreDev for short). In this comparison experiment, the parameter k is set as 10. The results in terms of average hit rate (*AverHitRate*) and average accuracy ($AverACC$) are listed in Table 2 and Table 3 respectively. From the two tables, we can see that, the results generated by the variants without considering the score deviations are not as good as those generated by CUDCF considering the score deviations. In particular, the results generated by CUDCF-NoScoreDev, i.e. the variant considering neither the score deviations of students nor the score deviations of courses, are the worst in most cases. After considering either the score deviations of students or the score deviations of courses, some improvements can be made. And the best results have been obtained when both the score deviations of students and the score deviations of courses are considered. The above comparison results have confirmed that considering the score deviations would improve the score prediction, leading to better optional course recommendation results.

B. DISCUSSION

From the results reported in the previous subsection, it can be expected that dramatic decrease of dropout rate could be obtained by the proposed method, although not yet confirmed by the experiments due to the lack of ground-truth dropout information of the testing students on the testing courses. According to the survey on the dropout of optional courses shown in Figure 2, the perceived difficulty of courses is the main reason for students to drop out optional courses. However, due to the requirement of the minimum number of optional courses that each student should enroll on during the undergraduate study, recommending the optional courses that students would expectedly obtain higher scores would surely lead to much lower dropout rate — after listening several weeks of the recommended courses on trial, the students would feel easier to obtain relatively higher achievement than

TABLE 3. Comparison results generated by CUDCF and its variants in terms of average accuracy (*AverACC*).

Majors	Semesters	Variants of CUDCF			CUDCF
		CUDCF-NoScoreDev	CUDCF-NoCourseScoreDev	CUDCF-NoStudentScoreDev	
Major1	2016 Fall	0.7332	0.7620	0.7667	0.8347
	2017 Spring	0.6438	0.7180	0.7345	0.8072
	2017 Fall	0.8353	0.9097	0.9300	1.0000
Major2	2016 Fall	0.7244	0.7485	0.7305	0.7654
	2017 Spring	0.8511	0.8783	0.8781	0.9594
	2017 Fall	0.8320	0.9067	0.9051	0.9345
Major3	2016 Fall	0.8155	0.8863	0.8750	0.9458
	2017 Spring	0.8098	0.8732	0.8730	0.8925
	2017 Fall	0.9082	0.9783	1.0000	1.0000
Major4	2016 Fall	0.7854	0.8803	0.8933	0.9314
	2017 Spring	0.8378	0.8609	0.8600	0.8947
	2017 Fall	0.8724	0.9123	0.9250	1.0000
Major5	2016 Fall	0.7405	0.7963	0.7963	0.8670
	2017 Spring	0.6331	0.7083	0.7030	0.7372
	2017 Fall	0.9512	0.9841	0.9841	1.0000
Major6	2016 Fall	0.7132	0.7650	0.7749	0.8199
	2017 Spring	0.8453	0.9312	0.9427	0.9690
	2017 Fall	0.9344	0.9891	1.0000	1.0000
Major7	2016 Fall	0.8036	0.8366	0.8333	0.8965
	2017 Spring	0.8474	0.8661	0.8600	0.8743
	2017 Fall	0.8018	0.8258	0.8207	0.8333
Major8	2016 Fall	0.7559	0.8484	0.8422	0.8808
	2017 Spring	0.8382	0.9160	0.9167	0.9528
	2017 Fall	0.8493	0.9594	0.9421	1.0000

those of unrecommended courses and therefore would not drop out the courses.

VI. CONCLUSIONS AND FUTURE WORK

By conducting a survey, it has been revealed that optional course selection is a critical but challenging issue for college students. To this end, in this paper, we have developed an optional course recommendation system based on score prediction. In particular, a novel cross-user-domain collaborative filtering (CUDCF) algorithm is designed to accurately predict the score of the optional course for each student by using the course score distribution of the most similar senior students. The underlying rationale is that, students with similar scores in the previous courses will generally obtain similar scores in the subsequent courses, which has been verified by some basic data analysis. After generating the predicted scores of all optional courses, the top t optional courses with the highest predicted scores without time conflict will be recommended to the student. Extensive experiments have been conducted to evaluate the effectiveness of the proposed method, the results of which show that the proposed method is able to achieve relatively high *average hit rate* and *average accuracy*. According to our discussion, it can also be expected that dramatic decrease of dropout rate could be obtained by the proposed method.

As the future work, in our approach, a novel cross-user-domain collaborative filtering (CUDCF) algorithm is designed, which is quite different from the existing cross-domain collaborative filtering (CDCF) algorithms [12], [15]. In particular, the existing CDCF algorithms usually share the common user domain but adopt different item domains (e.g. Book, CD, Music and Movie in the e-commerce platforms). Different from the existing CDCF algorithms, CUDCF shares the common item domain (i.e. the course domain) but adopts different user domains (i.e. the junior student domain and the senior student domain). In the future work, this may inspire the development of more cross-user-domain collaborative filtering algorithms in different recommendation tasks.

ACKNOWLEDGMENTS

The authors would like to thank Sun Yat-sen University for providing course-score data for academic research purpose.

REFERENCES

[1] P. Kinnunen and L. Malmi, "Why students drop out CS1 course?" in *Proc. 2nd Int. Workshop Comput. Educ. Res.*, 2006, pp. 97–108.
[2] M. E. Ibrahim, Y. Yang, D. L. Ndzi, G. Yang, and M. Al-Maliki, "Ontology-based personalized course recommendation framework," *IEEE Access*, vol. 7, pp. 5180–5199, 2018.
[3] M. Paechter, B. Maier, and D. Macher, "Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction," *Comput. Educ.*, vol. 54, no. 1, pp. 222–229, 2010.
[4] S. Kausar, X. Huahu, I. Hussain, W. Zhu, and M. Zahid, "Integration of data mining clustering approach in the personalized E-learning system," *IEEE Access*, vol. 6, pp. 72724–72734, 2018.
[5] A. A. Kardan, H. Sadeghi, S. S. Ghidary, and M. R. F. Sani, "Prediction of student course selection in online higher education institutes using neural network," *Comput. Educ.*, vol. 65, pp. 1–11, Jul. 2013.
[6] B. Dodge, J. Whitmer, and J. P. Frazee, "Improving undergraduate student achievement in large blended courses through data-driven interventions," in *Proc. 5th Int. Conf. Learn. Anal. Knowl.*, 2015, pp. 412–413.
[7] X. Ke, Y. Zeng, and H. Luo, "Autoscoring essays based on complex networks," *J. Educ. Meas.*, vol. 53, no. 4, pp. 478–497, 2016.
[8] N. Hoic-Bozic, M. H. Dlab, and V. Mornar, "Recommender system and Web 2.0 tools to enhance a blended learning model," *IEEE Trans. Educ.*, vol. 59, no. 1, pp. 39–44, Feb. 2016.

[9] A. Klašnja-Milićević, B. Vesin, M. Ivanović, and Z. Budimac, "E-learning personalization based on hybrid recommendation strategy and learning style identification," *Comput. Educ.*, vol. 56, no. 3, pp. 885–899, 2011.

[10] E. Mangina and J. Kilbride, "Evaluation of keyphrase extraction algorithm and tiling process for a document/resource recommender within e-learning environments," *Comput. Educ.*, vol. 50, no. 3, pp. 807–820, 2008.

[11] C.-D. Wang, Z.-H. Deng, J.-H. Lai, and P. S. Yu, "Serendipitous recommendation in E-commerce using innovator-based collaborative filtering," *IEEE Trans. Cybern.*, to be published.

[12] Z.-L. Zhao, L. Huang, C.-D. Wang, and D. Huang, "Low-rank and sparse cross-domain recommendation algorithm," in *Proc. 23rd Int. Conf. Database Syst. Adv. Appl.*, 2018, pp. 150–157.

[13] Z.-L. Zhao, C.-D. Wang, and J.-H. Lai, "AUI&GIV: Recommendation with asymmetric user influence and global importance value," *PLoS ONE*, vol. 11, no. 2, 2016, Art. no. e0147944.

[14] Z.-L. Zhao, C.-D. Wang, Y.-Y. Wan, and J.-H. Lai, "Recommendation in feature space sphere," *Electron. Commerce Res. Appl.*, vol. 26, pp. 109–118, Nov./Dec. 2017.

[15] X. Xin, Z. Liu, C.-Y. Lin, H. Huang, X. Wei, and P. Guo, "Cross-domain collaborative filtering with review text," in *Proc. IJCAI*, 2015, pp. 1827–1834.

[16] V. Tallón, P. Carvalho, H. Vinyals, and M. C. Manzanares, "Why do they choose an optional course? The rare case of teratology," *Eur. J. Dental Educ.*, vol. 18, no. 2, p. e14, 2014.

[17] B. J. LaMeres and C. Plumb, "Comparing online to face-to-face delivery of undergraduate digital circuits content," *IEEE Trans. Educ.*, vol. 57, no. 2, pp. 99–106, May 2014.

[18] J. L. M. Núñez, E. T. Caro, and J. R. H. González, "From higher education to open education: Challenges in the transformation of an online traditional course," *IEEE Trans. Edu.*, vol. 60, no. 2, pp. 134–142, May 2017.

[19] A. T. Korucu and H. Cakir, "The effect of dynamic Web technologies on student academic achievement in problem-based collaborative learning environment," *Malaysian Online J. Educ. Technol.*, vol. 6, no. 1, pp. 92–108, 2018.

[20] S. Moore and A. Rutledge, "The impact of guided learning on student achievement and retention in an online course environment: A reflection on introductory MIS courses," in *Proc. 13th Midwest Assoc. Inf. Syst. Conf.*, St. Louis, MO, USA, May 2018, pp. 1–5.

[21] Z. Ren, X. Ning, and H. Rangwala, "Grade prediction with temporal course-wise influence," in *Proc. 10th Int. Conf. Educ. Data Mining EDM*, Wuhan, China, Jun. 2017, pp. 48–55.

[22] D. F. Shell, L.-K. Soh, A. E. Flanigan, and M. S. Peteranetz, "Students' initial course motivation and their achievement and retention in college CS1 courses," in *Proc. 47th ACM Tech. Symp. Comput. Sci. Edu.*, 2016, pp. 639–644.

[23] S. R. Sletten, "Investigating flipped learning: Student self-regulated learning, perceptions, and achievement in an introductory biology course," *J. Sci. Educ. Technol.*, vol. 26, no. 3, pp. 347–358, 2017.

[24] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17–29, Jan./Mar. 2017.

[25] E. D. Thompson, B. V. Bowling, and R. E. Markle, "Predicting student success in a major's introductory biology course via logistic regression analysis of scientific reasoning ability and mathematics scores," *Res. Sci. Educ.*, vol. 48, no. 1, pp. 151–163, 2018.

[26] N. Z. Zacharis, "Classification and regression trees (CART) for predictive modeling in blended learning," *Int. J. Intell. Syst. Appl.*, vol. 10, no. 3, p. 1, 2018.

[27] E. Popescu and F. Leon, "Predicting academic performance based on learner traces in a social learning environment," *IEEE Access*, vol. 6, pp. 72774–72785, 2018.

[28] M. A. Chatti, S. Dakova, H. Thüs, and U. Schroeder, "Tag-based collaborative filtering recommendation in personal learning environments," *IEEE Trans. Learn. Technol.*, vol. 6, no. 4, pp. 337–349, Oct./Dec. 2013.

[29] A. Aztiria, J. C. Augusto, R. Basagoiti, A. Izaguirre, and D. J. Cook, "Learning frequent behaviors of the users in intelligent environments," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 6, pp. 1265–1278, Nov. 2013.

[30] M. Salehi, I. N. Kamalabadi, and M. B. G. Ghoushchi, "An effective recommendation framework for personal learning environments using a learner preference tree and a GA," *IEEE Trans. Learn. Technol.*, vol. 6, no. 4, pp. 350–363, Oct./Dec. 2013.

[31] A. Ruiz-Iniesta, G. Jiménez-Díaz, and M. Gómez-Albarrán, "A semantically enriched context-aware OER recommendation strategy and its application to a computer science OER repository," *IEEE Trans. Educ.*, vol. 57, no. 4, pp. 255–260, Nov. 2014.

[32] X. L. Zheng, C.-C. Chen, J.-L. Hung, W. He, F.-X. Hong, and Z. Lin, "A hybrid trust-based recommender system for online communities of practice," *IEEE Trans. Learn. Technol.*, vol. 8, no. 4, pp. 345–356, Apr. 2015.

[33] O. C. Santos and J. G. Boticario, "Practical guidelines for designing and evaluating educationally oriented recommendations," *Comput. Educ.*, vol. 81, pp. 354–374, Feb. 2015.

[34] K. Niemann and M. Wolpers, "Creating usage context-based object similarities to boost recommender systems in technology enhanced learning," *IEEE Trans. Learn. Technol.*, vol. 8, no. 3, pp. 274–285, Jul. 2015.

[35] S. Sergis and D. G. Sampson, "Learning object recommendations for teachers based on elicited ICT competence profiles," *IEEE Trans. Learn. Technol.*, vol. 9, no. 1, pp. 67–80, Jan./Mar. 2016.

[36] C.-B. Yao, "Constructing a user-friendly and smart ubiquitous personalized learning environment by using a context-aware mechanism," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 104–114, Jan./Mar. 2017.

[37] M. M. Rahman and N. A. Abdullah, "A personalized group-based recommendation approach for Web search in E-learning," *IEEE Access*, vol. 6, pp. 34166–34178, 2018.

LING HUANG received the bachelor's and master's degrees from the South China University of Technology, in 2009 and 2013, respectively. She is currently pursuing the Ph.D. degree with Sun Yat-sen University. She has published several papers in international journals and at conferences, such as *Pattern Recognition*, *Information Sciences*, *Knowledge-Based Systems*, AAAI, the IEEE ICDM, the IEEE BIBM, and DASFAA. Her research interest includes data mining.

CHANG-DONG WANG received the Ph.D. degree in computer science from Sun Yat-sen University, Guangzhou, China, in 2013, where he is currently an Associate Professor with the School of Data and Computer Science. He was a Visiting Student with the University of Illinois at Chicago, from 2012 to 2012. In 2013, he joined the School of Mobile Information Engineering, Sun Yat-sen University, as an Assistant Professor. He has published over 100 scientific papers in international journals and at conferences, such as the IEEE TPAMI, the IEEE TKDE, the IEEE TCYB, the IEEE TSMC-C, *Pattern Recognition*, the IEEE ACCESS, KAIS, AAAI, ICDM, CIKM, SDM, DASFAA, and BIBM. His current research interests include machine learning and data mining. His ICDM 2010 paper received the Honorable Mention for Best Research Paper Awards. He received the 2012 Microsoft Research Fellowship Nomination Award and the Outstanding Dissertation Award from the Chinese Association for Artificial Intelligence, in 2015.

HONG-YANG CHAO received the B.S. and Ph.D. degrees in computational mathematics from Sun Yet-sen University, Guangzhou, China, where she is currently a Full Professor with the School of Data and Computer Science. In 1988, she joined the Department of Computer Science, Sun Yet-sen University, where she was an Assistant Professor and became an Associate Professor. She has published extensively in the area of image/video processing. She holds three U.S. patents and four Chinese patents in the related area. Her current research interests include image and video processing, image and video compression, massive multimedia data analysis, and content-based image (video) retrieval.

JIAN-HUANG LAI received the M.Sc. degree in applied mathematics and the Ph.D. degree in mathematics from Sun Yat-sen University, China, in 1989 and 1999, respectively. In 1989, he joined Sun Yat-sen University as an Assistant Professor, where he is currently a Professor with the School of Data and Computer Science. He has published over 200 scientific papers in the international journals and at conferences on image processing and pattern recognition, including the IEEE TPAMI, the IEEE TCYB, the IEEE TKDE, the IEEE TNN, the IEEE TIP, *Pattern Recognition*, ICCV, CVPR, and ICDM. His current research interests include pattern recognition, data mining, and computer vision.

PHILIP S. YU received the B.S. degree in electrical engineering from National Taiwan University, the M.S. and Ph.D. degrees in electrical engineering from Stanford University, and the M.B.A. degree from New York University.

He was with IBM, where he was the Manager of the Watson Research Center, Software Tools and Techniques Group. He is currently a Distinguished Professor in computer science with the University of Illinois at Chicago, where he is also the Wexler Chair in information technology. He has published more than 1000 papers in refereed journals and at conferences. He holds or has applied for more than 300 U.S. patents. His research interests include big data, including data mining, data stream, database, and privacy. He is a Fellow of the ACM and the IEEE. He serves on the Steering Committee of the ACM Conference on Information and Knowledge Management. He was a member of the Steering Committee of the IEEE Data Engineering and the IEEE Conference on Data Mining. He was a recipient of the ACM SIGKDD 2016 Innovation Award for his influential research and scientific contributions on mining, fusion, and anonymization of big data, the IEEE Computer Society's 2013 Technical Achievement Award for pioneering and fundamentally innovative contributions to the scalable indexing, querying, searching, mining, and anonymization of big data, and the Research Contributions Award from the IEEE International Conference on Data Mining (ICDM), in 2003, for his pioneering contributions to the field of data mining. He received the ICDM 2013 ten-year Highest-Impact Paper Award and the EDBT Test of Time Award, in 2014. He received several IBM honors, including two IBM Outstanding Innovation Awards, the Outstanding Technical Achievement Award, two Research Division Awards, and the 94th Plateau of Invention Achievement Awards. He was an IBM Master Inventor. He was the Editor-in-Chief of the IEEE Transactions on Knowledge and Data Engineering, from 2001 to 2004, and the *ACM Transactions on Knowledge Discovery from Data*, from 2011 to 2017.