IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 17, 2020, accepted July 4, 2020, date of publication July 7, 2020, date of current version July 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007750

HHA: An Attentive Prediction Model for Academic Abnormality

YAWEN ZENG[1], YONG OUYANG[1], RONG GAO[1], YE QIU[1], YONGHONG YU[2], (Member, IEEE), AND CHUNZHI WANG[1]

[1]School of Computer Science, Hubei University of Technology, Wuhan 430068, China
[2]College of Tongda, Nanjing University of Posts and Telecommunications, Nanjing 211121, China

Corresponding author: Yong Ouyang (oyywuhan@163.com)

This work was supported in part by the General Program of the National Natural Science Foundation of China under Grant 61772180, in part by the funds for the Department of Science and Technology of Hubei Province under Grant 2013CFB021, in part by the Teaching Research Project of Hubei Province under Grant 2018316, and in part by the Doctoral Scientific Research Foundation of the Hubei University of Technology under Grant BSQD2019026.

ABSTRACT Warning students with poor performance in advance based on historical academic data, namely, the academic abnormality prediction is important task in education. The majority of existing methods focus on digging out abnormal complex clues from historical data, while ignoring two basic considerations:(1)these works fail to handle unrecorded/missing data when this part is sparse; (2)these works ignore the complex relationship between courses. The different courses are used as the attention weight vector for abnormality prediction, but they do not notice the mutual influence between courses. To this end, we contribute a ***H**ybrid Neural Network Model based on **H**igh-Order **A**ttention Mechanism*, called **HHA**, to address the academic abnormality prediction problem. Specifically, we first exploit Generative Adversarial Network(GAN) to mine hidden factors in the unrecorded/missing data reasonably by simulating student behavior. Thereafter, a high-order attention mechanism is proposed to measure the importance of course and course combination. Lastly, a multi-layer projection abstracts feature and classifies whether the student is abnormal. By experimenting on real-world dataset, we demonstrate the effectiveness and rationality of our proposed model.

INDEX TERMS Academic abnormality prediction, high-order attention mechanism, hybrid neural network, generative adversarial network.

I. INTRODUCTION

How to understand the various phenomena, problems or behaviors of students, and taking measures to improve their ability is not only important for students but also a widely studied topic in the whole education field. In the center of the spotlight, the early abnormal academic performance warning has attracted great attention from educators and the public for its forward-looking and improving education quality [1]. For this tasks, it is generally treated as a prediction issue, that is, to learn a mapping function from historical academic data to whether students are abnormal. Therefore, the main challenge of this task is how to mine useful information from a large number of teaching systems' data for anomaly prediction.

In fact, great progress has been made on the academic abnormality prediction in the research community [2], [3], such as cognitive diagnosis [4], matrix factorization [5], deep learning [6]. Generally, most methods devote efforts to modeling student grades for prediction and regard the students with poor grades as abnormal: cognitive modeling to predict students' test grades based on student achievement [7] to identify students with poor grades; cluster analysis to identify students' hidden patterns in feedback learning [8]. Meanwhile, there are some work extension data to courses and homework [9]. Reference [11] propose building a bayesian network model [10] based on homework to classify whether students are abnormal. Yu and Liu [12] used the LSTM neural network to learn the abnormal feature representation of after-class assignments.

Unfortunately, they focus on digging out abnormal complex clues from historical data, while ignoring two basic considerations. Despite its value and significance, the academic abnormality prediction has not been well addressed due to the following challenges:

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

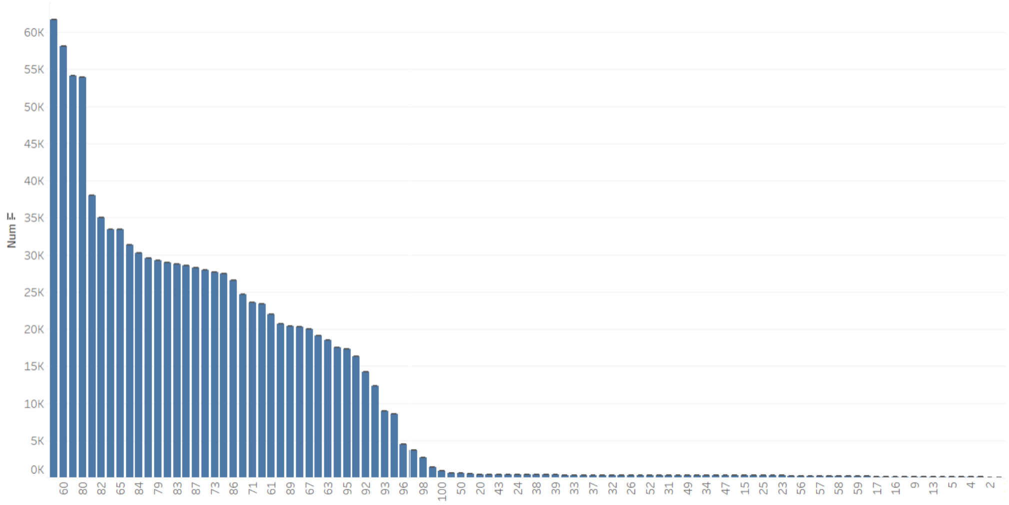


FIGURE 1. Grade distribution. The x-axis as the specific course grades and the y-axis as the number of records of the corresponding grade.

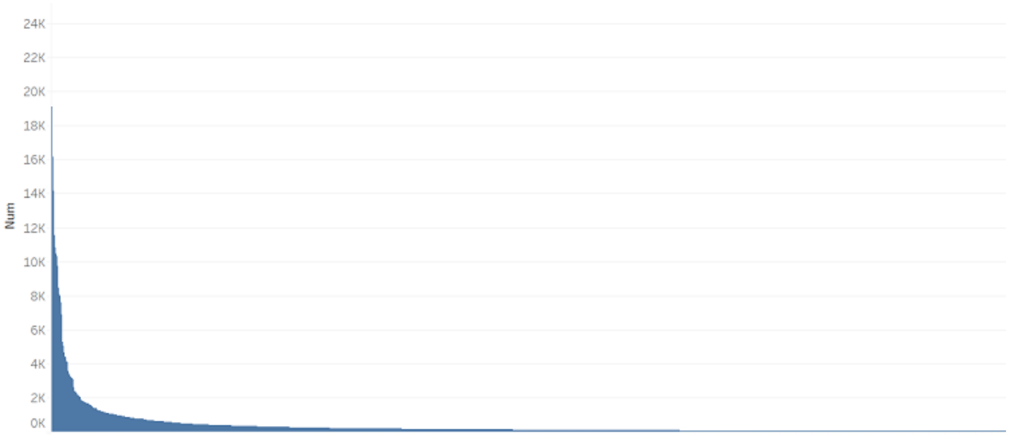


FIGURE 2. Course distribution. The x-axis represent different courses, and the y-axis shows the number of records that appear in the corresponding course. (Source: a real university's 2011-2017 student performance data).

1)Existing models fail to handle unrecorded/missing data when this part is sparse. In particular, the discarded damage is more serious in the massive data. According to statistics, actually, 63% of the grade data have less than 1,000 records in Figure 1, and only 8.3% of courses have records due to the existence of elective courses and incomplete records in Figure 2. For such massive and highly sparse situation, previous researchers [8]–[12] often discard, fill it through zero/grade averages, or other preprocessing methods. But this is obviously lack of exploration, because the course that the student does not have grades may be what he is good at, and the unrecorded/missing data does not indicate abnormality, which may imply unexplored knowledge.

2)Course relationships are not independent. Grades exist only in the presence of courses, but previous researchers [8], [9] often regarded the courses as equal and independent, which are unreasonable. The course design often have similar, complementary, advanced and other complex relationships, and they often appear in the form of combination, which proves that the courses are not independent, and there is a complicated relationship. Some basic courses have a greater impact on students than elective courses, furthermore, the course combination is a great complement to course information.

In recent years, deep learning methods have made breakthroughs in many fields such as image and speech [15], [16], which has become a boom in machine learning and has produced many meaningful applications [17]–[20]. Especially, Generative Adversarial Network(GAN) finds the hidden pattern of data through adversarial training to generate samples, which can generate realistic data. This adversarial paradigm has been proven to be effective in filling missing data and expanding richness [39]. And attention serves as a tool to bias the allocation of available resources towards the most informative parts in input, and has been successfully applied [21]–[23]. This also provides new ideas for academic abnormality prediction.

To address the aforementioned problems, we develop a *hybrid neural network based on high-order attention mechanism(HHA)* application in the academic abnormality prediction of college students. The general framework of HHA is illustrated in Figure 4. First of all, we employ GAN to fill in the unrecorded/missing data reasonably by simulating student behavior and mine the hidden factors. Thereafter, a high-order attention mechanism is proposed to measure the importance of course and course combination. Ultimately, a multi-layer projection is used to abstract feature and classifies whether the student is abnormal. Along this manner, the student behavior and course combination in the student's academic achievement data can be further discovered. By conducting experiments on real-world dataset, we validate that our proposed HHA is superior to other competitors on both overall performance comparison and specific case test.

Corresponding to the above questions, the main contributions of this work are three-fold:

1 Introduce GAN to simulate student behavior. To the best of our knowledge, this is the first work that attempts to solve filling data in academic abnormality prediction problem by adversarial learning paradigm. Specifically, a generator simulation is used to fill in the student's grade and a discriminator to distinguish whether it is true or not. This adversarial training method can find hidden factors of unrecorded/missing data.

2 Develop attention mechanism to measure the importance of the course, and further refined under a high-order method, which helped to find the correlation between students with academic abnormalities and the course. In particular, high-order attention can also find the importance of course combination. Meanwhile, we employ multi-layer projection to further abstract features and predict students' abnormal state.

3 In four datasets for different years, a large number of experiments have proved that our proposed method can be more discriminatory for students with abnormal academic performance.

The rest of this paper is organized as follows: Section 2 will introduce the related work. The HHA model will be described in detail in Section 3. Section 4 is data cleaning and integration processing. Experimental settings and results are reported in Section 5, followed by the conclusion and future work in Section 6.

II. RELATED WORK

We briefly review academic abnormality prediction, high-order attention mechanism and GAN in this section.

A. ACADEMIC ABNORMALITY PREDICTION

The development of machine learning and big data technology has opened the way for academic abnormality prediction and has yielded many methods [24]. At present, the methods adopted by mainstream research work can be mainly divided into the following two categories: traditional data mining methods for data analysis, and the other to combine deep learning.

Researchers mainly look for the correlation between courses and anomalies by combining various attributes. Such as analyzing students' behavior performance [25], association rules between courses [26], multi-factor combination of courses [28], and then making predictions through decision trees [9] and Bayesian methods [27]. Deep learning is a family of state-of-the-art techniques, which has achieved great success in many applications, such as speech recognition [29], image classification [30], and natural language processing [31]. Academic abnormality prediction in deep learning, which is mainly handled as a prediction problem [32], [33]. Similar to other prediction tasks, its purpose is to learn a mapping function from historical academic data to whether students are abnormal: [34] uses enhanced feedforward BP network, and [35], [36] uses RNN to analyze the courses and scores to predict whether the students are abnormal.

B. HYBRID NEURAL NETWORK

Hybrid neural network refers to a structure composed of multiple network models, which complement each other and enhance each other. There are many successful applications, such as wind power [18], cloud data centers [20], etc., which combines the autoencoder and wavelet decomposition to enhance the prediction task. Since [37] proposed a generative adversarial network, GAN has become very popular in the deep learning community. It employs two neural networks, pitting one against the other (thus the adversarial), to generate newly synthesized instances of data. By utilizing the merit of adversarial learning, GAN has been widely applied to multiple applications, [38] use it to process sparse data with excellent results, and [39], [46] use an adversarial training to learn optimal negative feedback representations. These all reflect its ability to mine hidden factors of data, which is very suitable to fill in some missing records.

C. ATTENTION MECHANISM

Attention serves as a tool to bias the allocation of available resources towards the most informative parts of an input [40]. However, in some complex or specific scenarios, the abstraction ability of vanilla attention is insufficient, so the research on the extension of multiple attention mechanisms is also endless. For example, the dual-flow attention mechanism [21] is used to perform feature refinement modeling on both dynamic and static aspects. The Transformer can even replace the RNN for feature capture [41]. High-order attention is the new attention-grabbing mechanism proposed by Chen *et al.* [42], which is also a new method for high-order expression in fine-grained image features.

All the above approaches fail to solve data sparseness and only consider the attention weight of dependent courses. Different from the above researchers' work, we proposes a hybrid neural network model based on high-order attention mechanism with GAN. On the one hand, GAN can find hidden factors in sparse data by simulating student behavior, on the other hand the high-order attention mechanism to mine detailed importance of the course and course combination.

III. PROPOSED HHA FRAMEWORK

In this section, we first formally introduce student abnormality prediction problems and give a solution overview in section 3.1. Then we describe the details of HHA model in section 3.2.

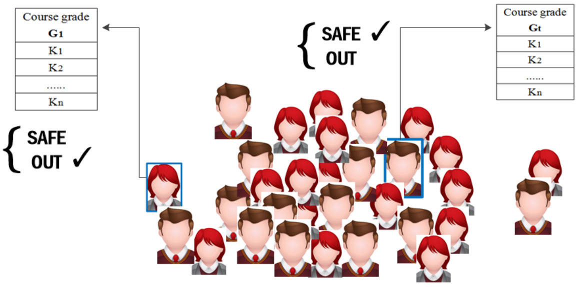


FIGURE 3. The HHA model predicts whether students' academic performance is normal based on historical academic performance.

A. FORMALIZATION

Overview as shown in Figure 3, the academic abnormality prediction task described in this paper refers to train a model that predicts whether a student is abnormal or not given the historical academic performance.

Definition 1 [Historical Course Grades]: According to the general processing method of academic abnormal prediction [7]–[12], suppose we have a sample of N college students' historical grades. Each college student's grades are composed of n courses. And $G_t = \{K_1, K_2, K_3, \ldots, K_{n-1}, K_n\}$ is used to indicate the course grade of a student t, where $K_1 \sim K_n$ are the grades of different courses, including compulsory courses and elective courses. For those without corresponding courses, the grade is marked with NaN. The dataset of college students' historical academic performance can be expressed as G and its format is described in detail in Section 4.1.

$$G = [G_1, \ldots, G_{N-1}, G_N]^T \tag{1}$$

Definition 2 [Abnormal status]: A student with poor performance and uneven performance may be considered abnormal, so we choose whether the student can graduate normally as one of the evaluation criteria. Then the advice is provided by human experts to determine if the student is abnormal, which is represented by $goal = \{SAFE, OUT\}$. OUT indicates that the student is in an abnormal state and is likely to fail to graduate normally [43]. The goal of our task is to distinguish OUT students as much as possible.

It is worth noting that all used notations in this paper are shown in Table 1.

TABLE 1. Notations.

Notation	Description
G	Historical Academic Performance
K	Different Courses
N	Total Number of Students
n	Total Number of Courses
k	Embedding size
goal	Abnormal State
SAFE	Normal State
OUT	Abnormal State
T	Generator of GAN
G	Discriminator of GAN
z	Adversarial Noise
ϵ, μ	Parameters of T and D
P_G, P_z	Probability Distribution of G and z
a,b,c	First, Second, and Third Order Attention Weights
bb,ccc	Attention Weight Under the Corresponding Order
i	Number of calculations required for each order
W^a, W^b, W^c	Weight Matrix
$b^{a_i}, b^{b_i}, b^{c_i}$	Weight Bias
m	Feature after Attention Weighting
ϕ	Activation Function
Y_{loss}	Loss Function
L_2	Regularizer
Θ	Network Parameters
α	Learning Rate

B. HHA MODEL

HHA aims to find a function for academic status from students' academic data. The framework of HHA is presented in Figure 4. In general, it consists of three components: behavior simulation layer, high-order convolution attention layer, and multi-layer projection layer.

1) BEHAVIOR SIMULATION LAYER

How to mine hidden factors from sparse data and understand student behavior is very important for predicting academic abnormalities. Furthermore, we observe that although some grades are missing, it is possible to infer the data from other observable. The intuitive idea of infering the data is utilizing random, zero filling. However, these methods have a very shallow understanding of student behavior, and also hurt the prediction results. To this end, we borrow the idea of the recent advance of GAN. By utilizing the merit of adversarial learning, GAN has been widely applied to multiple applications [46]. Especially, the adversarial paradigm has been proven to be effective in filling missing/noisy data and expanding richness [38].

Inspired by these pioneering efforts, the GAN layer is used to simulate student under adversarial paradigm. It includes following two components:

Generator T : $T_\epsilon(G'|G, z)$, parameterized by ϵ, it tries to generate student course grades G' by imitating student behavior. We add adversarial noise z [46] to the real-behavior data G to help the generator learn a better hidden factor, where the probability distribution of z is P_z.

Discriminator D : $D_\mu(G, G')$, parameterized by μ, it tries to discriminate real-behavior grade and generated sample pair $< G, G' >$. The ability of the discriminator can help the

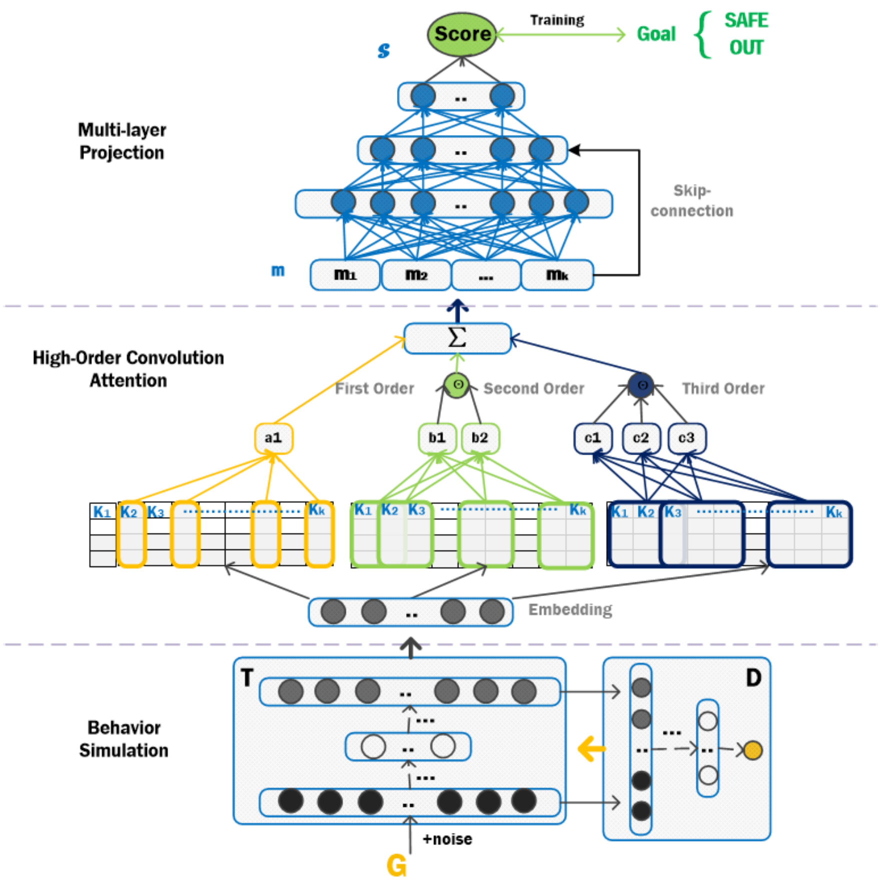


FIGURE 4. The graphical representation of our proposed model HHA. It is built upon three components: behavior simulation layer, high-order convolution attention layer, and multi-layer projection layer. The behavior simulation layer is exploited to fill in missing data, the high-order convolution attention layer measures the importance of courses and course combinations, and the multi-layer projection layer obtains anomaly prediction results.

generator to generate samples that are more like the distribution of student behavior P_G.

Specifically, the GAN generator and discriminator run the minimax game during the training process:

$$\min_{\epsilon} \max_{\mu} V(D, T) = \mathop{\mathbb{E}}_{G \sim P_G} [logD(G)] + \mathop{\mathbb{E}}_{z \sim P_z} [log(1 - D(T(G, z)))] \quad (2)$$

The purpose of the generator is to generate data that approximates the distribution to confuse the discriminator, and the discriminator strives to distinguish these samples. In the same way as the optimization method used by [39], we optimize D with stochastic gradient descent and G with policy gradient based reinforcement learning algorithm. Because the performance of the generator and the discriminator depend on each other, this filling method based on adversarial learning can force the model to learn other complete data to be close to the real distribution. Therefore when the game converges, the generator can mine hidden factors and generate data that simulates student behavior.

In this way, all the unrecorded/missing data are filled by GAN simulation. And we discuss the advantages of this adversarial method in detail in section 4.4. It is worth noting that GAN will be trained in advance as a pretrained module, which makes the model more efficient, then those new data become input to the next layer.

2) HIGH-ORDER CONVOLUTION ATTENTION LAYER

High-order convolution attention layer is used to distinguish the importance of different courses. The vanilla attention mimics the visual focus of the human brain, increasing the fine-grain by increasing the ''attention'' of the partial area. Based on the attention mechanism, we can compare the differences between courses feature.

However, vanilla attention can only compare the relationship between the two courses, and cannot explore the impact of course combination on academic abnormalities. For example, the attention mechanism can find that K_1 may be more important than K_2, but it cannot be deduced that (K_1, K_2, K_3) is more important than (K_4, K_5, K_6). However, in practice, courses often appear as a combination of choices.

Based on these observations, we try to transform attention into higher order, which can simultaneously capture the complex relationship between courses and course combinations. First of all, we design three convolutional modes $conv_{1,2,3}$ with $N \times 1, N \times 2, N \times 3$ kernel to find course combinations. This multi-scale convolution kernel can provide more information of receptive field on the course combination. Then we calculate the attention weight of the course combination under the same convolution kernel, and turn it into a high-dimensional representation. In particular, the attention mechanism is highly dimensioned by high-order polynomials to further enhance the discrimination [42] and richness of course combination in an explicit manner. For brevity, we call these three convolution modes first-order, second-order, and third-order, where the order refers to the polynomial level.

$$a_i = \sum_{r=1}^{k} \phi(W^{a_i} \times K_r + b^{a_i}) \tag{3}$$

It is worth noting that the combined weight obtained by each course will be obtained by the average of the weights of the multiple groups it belongs to. For example, the second-order attention weight bb_i, under the $N \times 2$ convolution kernel, belongs to the combination of two courses, so its combination weight is the average of the two groups. Therefore, second-order and third-order attention are shown in Eq.(4), Eq.(5).

$$\begin{cases} b_i = \dfrac{bb_{i-1} + bb_i}{2} \\ bb_i = \displaystyle\sum_{r=1}^{k-1} \phi(W^{b_i} \times [K_r, K_{r+1}] + b^{b_i}) \end{cases} \tag{4}$$

$$\begin{cases} c_i = \dfrac{ccc_{i-2} + ccc_{i-1} + ccc_i}{3} \\ ccc_i = \displaystyle\sum_{r=1}^{k-2} \phi(W^{c_i} \times [K_r, K_{r+1}, K_{r+2}] + b^{c_i}) \end{cases} \tag{5}$$

where k is embedding factor and ϕ is softmax fuction to assign weights a_i, b_i, c_i. K_r is the course feature, W^a, W^b, W^c is the weight matrix that converts the embedded vector into the attention network, and a_i, b_i, c_i is the intermediate result of the first, second, and third order attention weights, respectively. "i" is the number of calculations required for each order. bb_i calculates the attention weight for the combined features of the two courses, and finally the two courses share the weight equally in Eq.(5). And ccc assigns weights to the combined convolution features of the three courses.

Finally, we integrate the attention in the three convolutional modes into high-order polynomials. For example, the third order needs to be calculated 3 times, that is, c_1, c_2, c_3 and dot product. Then, the high-order weights are obtained through the combination of polynomials, and then multiplied by the original features to complete the redistribution of the weights. The model uses ReLU as the activation function of the hidden layer to project the features of the high-order combination into m. Computation of attention from different convolutional perspectives can complement the combination of courses to enhance the prediction of abnormal academics, we can infer some useful knowledge of course combination, which will be discussed high-order in detail in experiments.

$$m^{H=3} = K^T ReLU(a_1 + b_1 \cdot b_2 + c_1 \cdot c_2 \cdot c_3) \tag{6}$$

Algorithm 1 Learning of HHA Model

```
Require:
    Dataset $G$ that has been processed;
    Embedding $k$;
    Learning rate $\alpha$;
    Maximum epoch;
    Abnormal label goal.
Ensure: Trained network parameters $\Theta$.
    Initialize the parameters $\Theta$ randomly;
    Pretraining GAN in Eq.(2);
    repeat
        for each student do
            Calulate the layers of HHA in turn;
            Get the loss in Eq.(3)-(8);
            Update $\Theta$ in Eq.(9).
        end for
    until converges
```

3) MULTI-LAYER PROJECTION LAYER

The multi-layer projection layer abstract features and projection the prediction results *goal* through multi-layer full connection(FC) with skip-connection, which keep the invariance of features. Lastly, the model calculates the score in each category separately, and then obtains the highest category as the prediction result *goal′* by softmax.

$$goal' = \phi(FC_3([FC_2(FC_1(m)), m])) \tag{7}$$

For network training, the network's loss function such as Eq. (9), which L_2 is the regularizer for limiting parameters. The optimization algorithm used is stochastic gradient descent (SGD), and the detail of learning steps are as Algorithm 1. Note that, since GAN will be used as a pre-training module with a complexity of $O(N^2)$, the computational consumption is mainly in the convolution layer and the projection layer. The time complexity is $O(k^2N + N^2)$ and the space complexity is $O(N^2)$, where k is the embedding dimension and N is the number of samples.

$$\begin{aligned} Y_{loss} = &-\frac{1}{N} \sum_{G_i \in G} (goal_i lngoal'_i + (1 - goal_i) ln(1 - goal'_i)) \\ &+ L_2 \end{aligned} \tag{8}$$

$$\Theta_j := \Theta_j - \alpha(\frac{\partial Y_{loss}(i)}{\Theta_j}) \tag{9}$$

IV. EXPERIMENTS

In this section, we conduct extensive experiments to demonstrate the effectiveness of HHA model from various aspects.

RQ(1) How does our proposed HHA perform as compared to other competitors?

RQ(2) How do different components contribute to the performance of HHA? Especially attention and behavior filling.

RQ(3) How do different hyperparameter affect our model?

RQ(4) Can HHA be effective in real scenes?

A. EXPERIMENTAL SETTINGS

1) DATESETS

We have collected the real-world experimental datasets that nearly 28,429 students' records from 22 departments over spanning almost six years. These academic records include elective courses and compulsory courses. The attributes of the academic achievement data are (TASK-NO, CURNAME, CURTYPE, CURDEP, CURCREDITH, STUID, STUNAME, STUSEX, STUCLASS, STUDEP, GRADE) represent (course number, course name, course type, course affiliated college, credits, student number, student name, gender, class, student's college, grade) respectively. The entry of grades is based on the course and the attributes distribution is shown in Table 1.

TABLE 2. Attributes of the academic achievement data.

	TASK-NO	CURTYPE	CREDITH	STUID	STUDEP	GRADE
Type	String	String	String	String	String	Int
2014	18505	88	30	20781	25	102
2015	19834	91	30	20818	27	102
2016	23221	119	29	20870	14	102
2017	17809	93	28	21845	19	103

At the same time, we have collected the abnormal students for four years(2014-2017) with different distribution to guide this task. Advice is provided by human experts to determine if the student is abnormal, which is represented by $goal = \{SAFE, OUT\}$. OUT indicates that the student is in an abnormal state and is likely to fail to graduate normally. The number of expert marks for abnormal students per year is shown in Table 2. One-hot processing is used for all course scores including each student. The "NaN" is missing mark, and the "goal" is an abnormal mark.

TABLE 3. Abnormal students per year.

	SAFE	OUT
2014	5182	318
2015	4621	293
2016	4631	298
2017	4832	217

The dataset composition as input G is shown in Table 3. And 70% of the academic achievement dataset G was randomly selected as the training set G_{train} (including 20% as the validation set of the training process), and another 10% of the data G_{test} was used to evaluate the results.

2) EVALUATION PROTOCOLS

To make a quantitative evaluation of model's validity, the evaluation indicators use popular precision and recall. The precision indicates the proportion of students who correctly predicted the model in the test set. The recall indicates the prediction accuracy of different goal categories.

TABLE 4. Matrix G as HHA input.

	K_1	K_2	K_3	...	K_{n-2}	K_{n-1}	K_n	goal
G_1	92	71	86		75	NaN	NaN	SAFE
G_2	75	85	79		NaN	80	NaN	SAFE
G_3	65	80	74		NaN	NaN	65	OUT
G_4	75	67	NaN		85	NaN	87	SAFE
...								

$$Precision = \frac{\sum_{goal'_i=goal_i} G_i}{G_{test}} \tag{10}$$

$$Recall(x) = \frac{\sum_{goal'_i=x} G_i}{\sum_{goal_j=x} G_j} \tag{11}$$

3) BASELINES

To demonstrate the effectiveness of HHA, this group of experiments compared the following four advanced methods of academic abnormality prediction.

1. **AkN [44]**, AkN finds k nearest neighbors for each query object, which can be used to deal with the problem of abnormal points. The similarities and differences between students are judged. Students who are off-center are considered abnormal. We set the optimal k to 4.
2. **GBDT [45]**, GBDT is a method of Ensemble learning, which constructs multiple regression trees as various features to find the best decision point. Using the training data to build the best GBDT, the combined results of multiple trees is to get a student's two classifications between SAFE and OUT.
3. **EBP [34]**, which is still essentially a neural network. The network results are continuously adjusted to obtain a network model suitable for a specific category, and the output probability of neural network as the prediction result. We set the optimal number of layers is 8.
4. **AFM [33]** is a strong baseline which adds an attention mechanism to the factor decomposition machine and can find hidden factors well. The model settings are the same as the author, and we use it for training and prediction.
5. **APR [46]**, this method improves the quality of generated samples by adding adversarial noises and uses BPR for ranking. We realized this adversarial learning scheme demonstrate its positive effect on academic abnormality. The adversarial coefficient is set to 0.2.
6. **DNN-MRT [18]**, which exploits the learning ability of deep neural network based ensemble technique and the concept of transfer learning. We use this paper's excellent deep auto-encoders and deep belief network ideas to handle this prediction tasks.

To further evaluate the effectiveness of our designed high-order attention, we have designed various variants of our method.

7 **HHA-**, to further evaluate the effectiveness of our designed attention, the HHA- without the attention mechanism is compared.

8 **HHA*{1, 2, 3, 4}**, the model with the attention order*{1, 2, 3, 4}–HHA*{1, 2, 3, 4} is used under the same conditions, where HHA*1 is the basic attention model, then continue to increase the order for component testing.

4) IMPLEMENTATION DETAILS

We implemented HHA and all baselines based on the PyTorch framework on a server equipped with a NVIDIA 2080TI-11G GPU. To initialize the embedding layer and hidden layers of neural networks, we randomly set their parameters with a Gaussian distribution (a mean of 0 and a standard deviation of 0.1), SGD optimizer is employed for all gradient-based methods where the mini-batch size and learning rate were set as 128 and 0.01, respectively. The dimensions of the multi-layer projection are 128-64-32, shown in Table 5. For specific hyper-parameters in our framework, the number of order is set as 3, the embedding size is set as 30, which are discussed in detail in experiments 4.3 and 4.5 respectively.

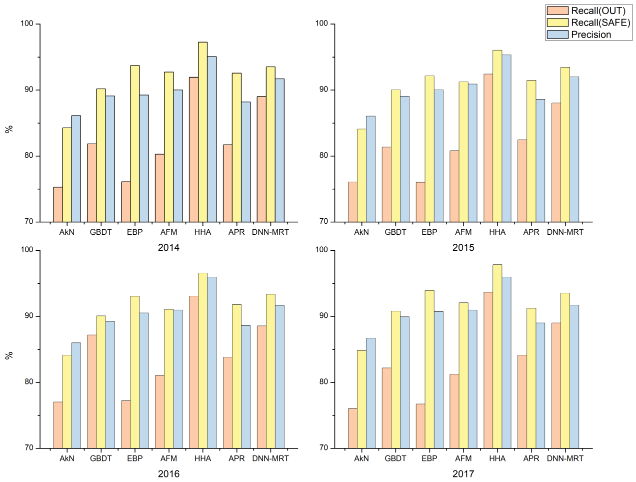


FIGURE 5. Performance comparison of various baselines.

TABLE 5. General parameter setting for all the dataset.

Number of Epochs	800
Batch Size	128
Learning Rate	0.01
Number of Order	3
Embedding Size	30
Multi-layer Projection	128-64-32

B. OVERALL PERFORMANCE COMPARISON (RQ1)

To demonstrate the effectiveness of HHA, we compared it with several approaches: AkN, GBDT, EBP, and AFM. Experimental results are shown in Figure 5. We have the following observations:

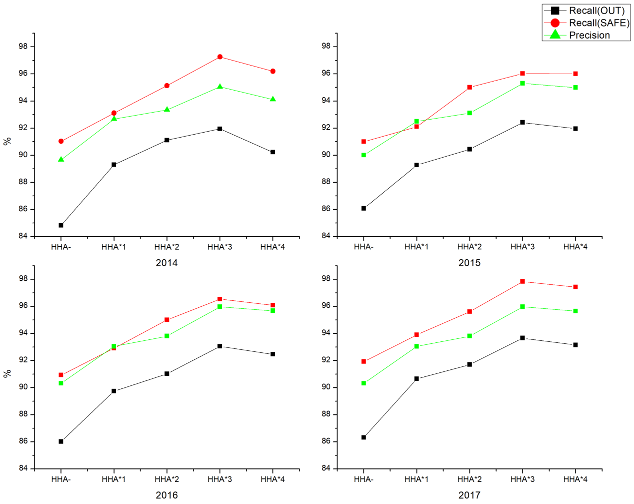


FIGURE 6. Performance comparison of various attention variants.

- AkN: AkN has a general effect on precision and recall(SAFE), and its simple operation makes it widely used in practice. However, the clustering method obviously cannot respond to students with abnormal trends in the aggregated data, and only recognizes the obvious abnormal points deviating from the cluster, which also causes Recall(OUT) to be the worst in all models.
- GBDT: GBDT makes better use of course features than AkN, and has improved in various indicators. However, in the process of building the decision tree, the difference between the characteristics of students is not obvious, which leads to the unclear branching rules of the tree.
- EBP: Although the use of neural networks has improved abstraction capabilities of the model and gained a better representation of the curriculum characteristics, both Precision and Recall(SAFE) have increased. However, EBP did not have a deep understanding of the data, and the effective improvement was not obvious enough. Also, the ability to capture students' abnormal tendencies is insufficient, and Recall(OUT) is not as good as GBDT, only slightly better than AkN.
- AFM: AFM has achieved good performance, which is related to its mining of hidden factors and the use of attention mechanisms. And it is very stable on four datasets, which also shows its excellent ability, but it is still limited by the incompleteness of the data, which is why we should introduce adversarial training.
- APR: APR improves the performance by adding the adversarial learning strategy. It well demonstrates the effectiveness of adversarial learning in solving the academic abnormality problem.
- DNN-MRT: The performance of DNN-MRT is the best among all baselines. We think this is because its deep auto-encoders handles the simulation and filling as GAN, and the deep belief network also captures good features.

- HHA: HHA ranks first in all indicators. This may be because HHA uses the high-order attention mechanism to balance the importance of the course performed well, so it can better understand historical data. Among them, the performance of attention on the third order(HHA*3) is the best, which detailed discussion is in section 4.3.

In addition, in the model training time, AkN and GBDT training speed is faster, and other networks based on neural networks are slower, this is because the neural network needs to train a large number of parameters. Training our model can be completed in about 1.5 hours.

But in the testing phase, AkN needs to calculate a large amount of similarity, which is very slow, and the parameters of other models are saved after training, so the results can be obtained very quickly. Among them, GBDT has the fastest speed because the time complexity is $O(NlogN * k * l)$ (N is the number of samples, k is the number of features, and l is the depth of the tree). Other methods are all related to deep learning, the complexity is at least $O(N^2)$, but the speed difference between each other is small. To predict the abnormal state of a single student, the results can be obtained within 1s.

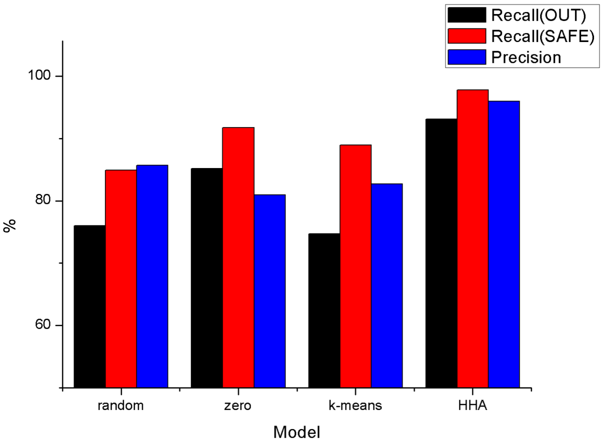


FIGURE 7. Comparison of each filling method.

C. VARIOUS ATTENTION VARIANTS COMPARISON (RQ2)

To further evaluate the effectiveness of our designed attention mechanism that measure the courses and course combinations' importance for academic abnormality prediction, we have designed various variants of our method: HHA-, HHA*{1, 2, 3, 4}. HHA- is a variant with no attention, and HHA*{1, 2, 3, 4} is a variant with increasing attention order from first-order to fourth-order to get a suitable higher order. From Figure 6, we have the following observations:

HHA-: Compared with HHA*1, HHA*2 and HHA*3, HHA- has a different degree of decline in all aspects of performance, which proves that the addition of attention mechanism is effective. It can capture the complex relationships between courses, such as similarity, complementarity, advancement, etc., to better distribute weights.

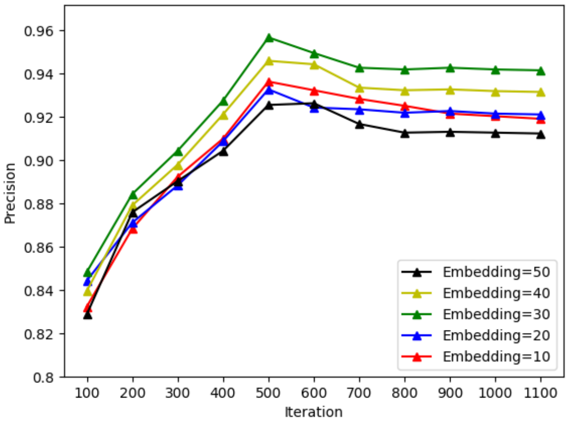


FIGURE 8. Process of training performance.

HHA*{1, 2, 3, 4}: In the four datasets, the results all rise first, reach the highest when the order is 3, and then fall. Such a curve proves that when the order is 3, HHA works best. We believe that the high-order attention of the appropriate order can enhance the attention mechanism's ability to mine the nuances between the features and supplement combination information, and third-order attention in academic data has already met the task needs.

D. PRE-TRAINING METHOD COMPARISON (RQ2)

To investigate the effectiveness of our GAN pre-training scheme, we performed the study on various classic filling initialization methods, such as random filling method(random) based on user's random setting, zero filling method(zero) set directly to 0, and K-means filling method(K-means) based on K-means for initial setting.

Figure 7 show that the effect of zero filling is better than random and k-means, which shows that the prediction of normal students does not need to rely on missing data to have good results. This proves that setting to zero is more reasonable than taking other methods, which may be due to the sparseness of the data. The filling method will greatly disturb the data distribution, thus making the prediction result inaccurate.

And GAN is superior to other initialization methods. Compared with other simple filling methods, the main reason for this result is that GAN can find some hidden relationships in the data through adversarial training, which is much better than using ordinary filling. This proves that HHA using adversarial training can find the hidden factors of students and fill in the unrecorded/missing data more appropriately.

E. MODEL PERFORMANCE AND PARAMETER TUNING (RQ3)

The experiment evaluates the training performance of model under different epoch and the influence of embedding parameters. Since the change of hidden factor dimension k of embedding will have no small impact on the robustness of the proposed model, in general, the larger the dimension, the more time and memory the algorithm will run, and it will affect the iterative process. Therefore, it is necessary to explore the influence of dimensions on the performance of this model.

The experimental results are shown in Figure 8, in which the lines of different shapes represent the model training cases with different embedding numbers k, and the number of embeddings is set from the range of 10-50. As can be seen from the trend in figure, the setting of dimension has an effect on the model, and the more the number of embedding, the more parameters will be caused, the convergence will be slower, but precision will rise first and then fall. The reasons for this result are that we believe that there are not so many courses that deserve special attention due to the complex relationships such as similarity, complementarity, and upgrades that may exist between courses and the sparse form of elective courses. HHA can sensitively find suitable feature dimensions and achieve effective results.

F. PRACTICAL APPLICATION PERFORMANCE ANALYSIS (RQ4)

To verify the performance of HHA in practical applications, we first compare the model's performance on dataset of different sizes, and then conduct case tests on real students to prove practicability.

1) APPLICABILITY ON DIFFERENT SCALE DATASETS

To study the impact of HHA as the scale of data increases, we add two additional datasets on the basis of the 2014-2017 datasets. ALL consists of all students available for training, and its data size is four times that of a single year. In addition, we reduce the data sparsity on the basis of ALL, and delete all records with less than 50% integrity to get $ALL50+$. The results are shown in Table 6.

TABLE 6. Applicability of HHA on different scale datasets.

dadaset	Number of records	Precision	Recall(SAFE)	Recall(OUT)
2014	352658	95.05%	97.26%	91.94%
2015	365983	95.30%	96.03%	92.42%
2016	381236	95.08%	96.51%	93.05%
2017	351241	95.97%	97.83%	93.66%
ALL	1416875	95.11%	96.32%	91.77%
$ALL50+$	1013569	94.83%	96.57%	91.59%

From the table, we can find that for larger data sizes, the performance of HHA decreased slightly, but it is still relatively stable. The performance of ALL with a larger amount of data is even better than $ALL50+$, which proves the effectiveness of our model.

2) PRACTICABILITY ON SPECIFIC STUDENTS

To demonstrate the effectiveness of our model in real-world scenarios, this experiment tests the abnormality prediction of each student in each year and tracks their academic status. We randomly take the predictions of the 12 juniors in the freshman, sophomore, and junior stages, and there is a trend chart as shown in Figure 9. The y-axis represents the probability that the goal is predicted to be OUT, and the student who is likely to be abnormal is selected.

As shown in Figure 9, our model can not only identify students with extreme anomalies, but also sensitively detect students with strong fluctuations, which implies their current anomalies. Specifically, the academic performance of students U_8 is extremely extreme, and will attract the attention of educators in practical teaching. However, in addition to the need to pay high attention to U_8 students, it may be more important for students such as U_4 and U_9, whose employment difficulties are increasing, and students with large fluctuations such as U_3 and U_{12}.

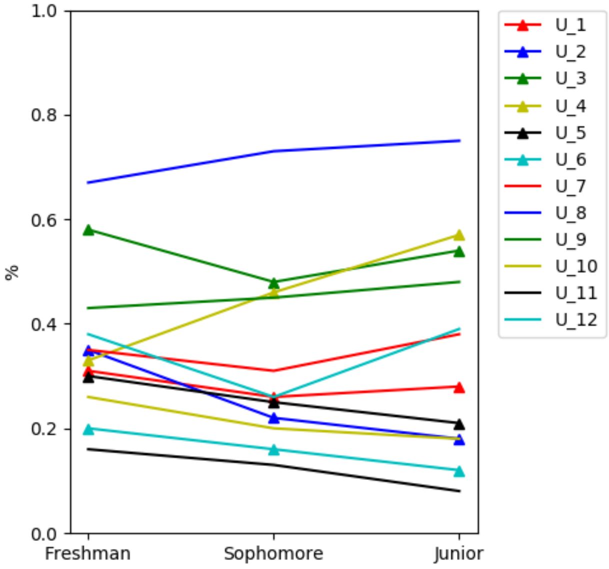


FIGURE 9. Academic performance status trend on specific students.

HHA has well completed the tracking of students' timely learning status, and can complete the academic early warning task more carefully and accurately. The possible reasons are: 1) The model treats the academic performance as a probabilistic expression and can visually observe the change. 2) The sparseness of the training data itself enables testing in the case of freshman and sophomore who have incomplete courses, which also proves the importance of GAN for student behavior imitation. 3) Experiments from Section 4.2-4.6 have demonstrated the effectiveness of the HHA architecture, which will allow HHA to be successfully applied in practice.

V. CONCLUSION

In this paper, we present a novel hybrid neural network based on high-order attention mechanism with GAN to the academic abnormality prediction. Specifically, we first exploit GAN for pretraining to find hidden factor of unrecorded/missing data by simulating student behavior, then the high-order attention mechanism to balance the importance of the course and course combination. The experimental results show that HHA achieves better performance for the academic abnormality prediction task; further analyses demonstrate how different components in HHA contribute to the performance of HHA, how HAA is sensitive to attention and filling method, and how HAA achieve good performance via specific case test.

In the future, we will continue to understand the meaning of dataset and consider how to dig more expressive features. Technically, such as deep auto-encoders and wavelet decomposition [20] widely used in prediction tasks and showed their powerful ability. Furthermore, complex course relationships can be abstracted as the graph [47], [48], so combining graph may further enhance the understanding of the course.

REFERENCES

[1] Y.-H. Hu, C.-L. Lo, and S.-P. Shih, "Developing early warning systems to predict students' online learning performance," *Comput. Hum. Behav.*, vol. 36, pp. 469–478, Jul. 2014.

[2] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, Mar. 2014.

[3] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Mining*, vol. 1, no. 1, pp. 8–16, 2009.

[4] L. DiBello, V. Roussos, and W. Stout, "31A review of cognitively diagnostic assessment and a summary of psychometric models," *Handbook Statist.*, vol. 26, pp. 979–1030, Dec. 2006.

[5] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Comput. Sci.*, vol. 1, no. 2, pp. 2811–2819, 2010.

[6] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 505–513.

[7] R. Wu, Q. Liu, Y. Liu, E. Chen, Y. Su, Z. Chen, and G. Hu, "Cognitive modelling for predicting examinee performance," in *Proc. AAAI*, 2015, pp. 1017–1024.

[8] A. M. Abaidullah, N. Ahmed, and E. Ali, "Identifying hidden patterns in students' feedback through cluster analysis," *Int. J. Comput. Theory Eng.*, vol. 7, no. 1, pp. 16–20, Feb. 2014.

[9] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. V. Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. Bus. Res.*, vol. 94, pp. 335–343, Jan. 2019.

[10] J. Huang, "Bayesian network in the application of student performance prediction," *Comput. Sci.*, vol. 39, no. S3, pp. 280–282, 2012.

[11] J. Zhang, H. Zhang, and X. Zhao, "The probability of rules semi-automatic learning soft logic reasoning model," *Comput. Appl.*, vol. 38, no. 11, pp. 3144–3149, 2018.

[12] S. Yu and Q. Liu, "Exercise-enhanced sequential modeling for student performance prediction," in *Proc. AAAI*, 2018, pp. 2435–2443.

[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, 2009, Art. no. 15.

[14] C. Sammut and G. Webb, *Encyclopedia of Machine Learning and Data Mining II*. Springer, 2017.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] T. Liu, S. Yu, B. Xu, and H. Yin, "Recurrent networks with attention and convolutional networks for sentence representation and classification," *Appl. Intell.*, vol. 48, no. 8, pp. 1–10, 2018.

[17] J. Bi, H. Yuan, and M. Zhou, "Temporal prediction of multiapplication consolidated workloads in distributed clouds," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 4, pp. 1763–1773, Oct. 2019.

[18] A. S. Qureshi, A. Khan, A. Zameer, and A. Usman, "Wind power prediction using deep neural network based meta regression and transfer learning," *Appl. Soft Comput.*, vol. 58, pp. 742–755, Sep. 2017.

[19] G. Wang, J. Qiao, J. Bi, Q.-S. Jia, and M. Zhou, "An adaptive deep belief network with sparse restricted Boltzmann machines," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 24, 2019, doi: 10.1109/TNNLS.2019.2952864.

[20] J. Bi, H. Yuan, L. Zhang, and J. Zhang, "SGW-SCN: An integrated machine learning approach for workload forecasting in geo-distributed cloud data centers," *Inf. Sci.*, vol. 481, pp. 57–68, May 2019.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and Ł. Kaiser, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 1–11.

[22] D. Cao, Z. Yu, H. Zhang, J. Fang, L. Nie, and Q. Tian, "Video-based cross-modal recipe retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1685–1693.

[23] Z. Jinliang, L. Wang, Z. Liu, Q. Zhang, G. Hua, and N. Zheng, "Attention-based temporal weighted convolutional neural network for action recognition," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.*, 2018, pp. 97–108.

[24] E. Nasiri, A. L. I. Pour-Safar, M. Taheri, A. S. Pashaky, and A. A. Louyeh, "Presenting the students' academic achievement causal model based on goal orientation," *J. Adv. Med. Educ. Professionalism*, vol. 5, no. 4, pp. 195–202, 2017.

[25] B. Emond and S. Buffett, "Analyzing student inquiry data using process discovery and sequence classification," *Int. Educ. Data Mining Soc.*, Jun. 2015, pp. 412–415.

[26] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," vol. 22, no. 2, pp. 207–216, 2011.

[27] Q. Zhou and Y. Xiao, "Based on data mining technology analysis of the college students' academic warning," *J. Edu. Tech. Equip. China*, vol. 6, pp. 36–39, 2018.

[28] N. M. Rusli, Z. Ibrahim, and R. M. Janor, "Predicting students' academic achievement: Comparison between logistic regression, artificial neural network, and neuro-fuzzy," in *Proc. Int. Symp. Inf. Technol.*, Aug. 2008, pp. 1–6.

[29] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[30] A. Krizhevsky, I. Sutskever, and Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[32] A. Toscher and M. Jahrer, "Collaborative filtering applied to educational data mining," *KDD Cup*, Jan. 2010.

[33] N. Thai-Nghe, L. Drumond, T. Horváth, A. Krohn-Grimberghe, A. Nanopoulos, and L. Schmidt-Thieme, "Factorization techniques for predicting student performance," *Educ. Recommender Syst. Technol., Practices Challenges*, pp. 129–153, Jan. 2011.

[34] A. R. Ajiboye, R. Abdullah-Arshah, and H. Qin, "Using an enhanced feed-forward BP network for predictive model building from students' data," *Intell. Automat. Soft Comput.*, vol. 22, no. 2, pp. 169–175, 2015.

[35] M. Li *et al.*, "Research and implementation of college students' academic warning algorithm," *J. Edu. Inf. Forum*, vol. 4, pp. 75–76, 2018.

[36] M. Paliwal and U. Kumar, "A study of academic performance of business school graduates using neural network and statistical techniques," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7865–7872, 2009.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[38] J. Liu, Z.-J. Zha, R. Hong, M. Wang, and Y. Zhang, "Deep adversarial graph attention convolution network for text-based person search," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 665–673.

[39] R. Gao, H. Xia, J. Li, D. Liu, S. Chen, and G. Chun, "DRCGR: Deep reinforcement learning framework incorporating CNN and GAN-based for interactive recommendation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 1048–1053.

[40] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[41] P. Sun, L. Wu, and M. Wang, "Attentive recurrent social recommendation," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.* Arlington, VA, USA: Sigir, Jun. 2018.

[42] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," 2019, *arXiv:1908.05819*. [Online]. Available: https://arxiv.org/abs/1908.05819

[43] C. Canals, E. Goles, A. Mascareño, S. Rica, and G. A. Ruz, "School choice in a market environment: Individual versus social expectations," *Complexity*, vol. 2018, pp. 1–11, Dec. 2018.

[44] H.-D. Zhang, Z.-H. Xing, L. Chen, and Y.-J. Gao, "Efficient metric all-k-nearest-neighbor search on datasets without any index," *J. Comput. Sci. Technol.*, vol. 31, no. 6, pp. 1194–1211, Nov. 2016.

[45] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Dropout prediction in Edx MOOCs," in *Proc. IEEE 2nd Int. Conf. Multimedia Big Data (BigMM)*, Apr. 2016, pp. 440–443.

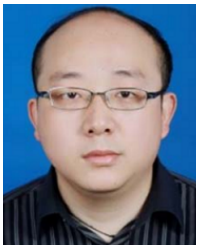
[46] X. He, Z. He, X. Du, and T.-S. Chua, "Adversarial personalized ranking for recommendation," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 355–364.

[47] Y. Ouyang, Y. Zeng, R. Gao, Y. Yu, and C. Wang, "Elective future: The influence factor mining of students' graduation development based on hierarchical attention neural network model with graph," *Int. J. Speech Technol.*, Apr. 2020, doi: 10.1007/s10489-020-01692-6.

[48] H. Wang, G. Xiao, N. Han, and H. Chen, "Session-based graph convolutional ARMA filter recommendation model," *IEEE Access*, vol. 8, pp. 62053–62064, 2020, doi: 10.1109/access.2020.2984039.

YAWEN ZENG received the B.S. degree in computer science from the Hubei University of Technology, Wuhan, China, in 2019. Her research interests include data mining and information retrieval.

YONG OUYANG received the M.S. degree from the Hubei University of Technology, Wuhan, China, in 2007. He is currently an Associate Professor with the School of Computer Science, Hubei University of Technology, where he is also the Director of the Department of Computer Science and Technology. His research interests include data mining and intelligent education.



RONG GAO received the Ph.D. degree from Wuhan University, Wuhan, China, in 2018. He is currently an Assistant Professor with the School of Computer Science, Hubei University of Technology, Wuhan. His research interests include data mining and intelligent recommendation.

YE QIU received the B.S. degree in computer science from the Hubei University of Technology, Wuhan, China, in 2019. Her research interests include data mining and intelligent recommendation.

YONGHONG YU (Member, IEEE) received the Ph.D. degree from Nanjing University, Nanjing, China, in 2016. He is currently an Associate Professor with the College of Tongda, Nanjing University of Posts and Telecommunications, Nanjing. His current research interests include data mining and machine Learning.

CHUNZHI WANG received the Ph.D. degree from the Wuhan University of Technology, Wuhan, China, in 2013. She is currently a Professor with the School of Computer Science, Hubei University of Technology, Wuhan. Her research interests include computer networks and data processing.