Predicting students’ performance is one of the most concerned issues in education data mining (EDM), which has received more and more attentions. Feature selection is the key step to build prediction model of students’ performance, which can improve the accuracy of prediction and help to identify factors that have significant impact on students’ performance. In this paper, a hybrid feature selection method named rank and heuristic (RnkHEU) was proposed. This novel feature selection method generates the set of candidate features by scoring and ranking firstly and then uses heuristic method to generate the final results. The experimental results show that the four major evaluation criteria have similar performance in predicting students’ performance, and the heuristic search strategy can significantly improve the accuracy of prediction compared with forward search method. Because the proposed RnkHEU integrates ranking-based forward and heuristic search, it can further improve the accuracy of predicting students’ performance with commonly used classifiers about 10% and improve the precision of predicting students’ academic failure by up to 45%.

1. Introduction

Intelligent and personalized education has developed rapidly with the development of big data, artificial intelligence, Internet of things (IoT), and other new generations of information technologies in recent year [1], which is one of the most important directions of sustainable development of education. Instructors use more and more digital systems to support and assist in educational activities, such as Student Information System (SIS), Course Learning Management System (LMS), Massive Open Online Course (MOOC), and Virtual Learning System. These systems record and generate huge amount of data daily, which is called educational big data. How to analyze educational big data thoroughly to find new knowledge to improve educational performance is a tremendous challenge faced by educational institutions and researchers. Traditional analysis of educational data uses sampling, hypothesis testing, linear regression, and other methods from the field of statistics, which cannot meet the requirement of analyzing educational big data completely.

At present, Educational Data Mining (EDM) is one of the most important ways to analyze educational data thoroughly. It is widely used in learning path planning, resource recommendation of personalized learning, data support decision-making, and so on. As an interdisciplinary research field, EDM applies machine learning, statistics, data mining, educational psychology, cognitive psychology, and other theories and technologies to analyze educational data, helping to solve various problems in education effectively [2]. The disciplines involved in EDM are shown in Figure 1. EDM method is a combination of methods coming from statistics, machine learning, data mining, and other fields. It
can be roughly divided into six categories: data extraction, prediction, association mining, structure discovery, model-based recognition, and hybrid methods [3].

Predicting students’ performance is one of the most important issues in the field of EDM. By predicting students’ performance, we can identify the risk of students’ academic failure in the learning process as early as possible, so as to intervene and guide in advance. It can also provide recommendation basis for personalized learning and provide support for educational administrators to make decisions by analyzing the factors affecting students’ performance [5, 6]. In the field of EDM, predicting students’ performance is a classification problem in machine learning. Researchers train the prediction model based on supervised learning classification algorithm using labeled students’ historical academic data. The trained prediction model outputs the class of students’ performance according to students’ demographic or historical academic performance and other features. However, students’ historical academic dataset collected from SIS, LMS, and other digital systems often contains a large number of features, some of which have no influence on students’ performance. It is necessary to select the features that have a significant impact on the prediction output through feature selection.

The focus of feature selection is to select a subset of features from the input, which can efficiently describe the input data, by reducing noise or irrelevant features [7]. In the past decades, researchers have proposed a huge number of feature selection methods, which can be divided into three categories: filtering method, wrapper method, and embedded method [7]. Filtering methods use evaluation criteria based on correlation or information entropy to rank all features and select best features to form a feature subset as the result of feature selection. ReliefF [8], F-statistical [9], and information gain [10] are frequently used filtering methods. Wrapper methods use machine learning algorithm to evaluate the quality of feature subsets, which lead to the high quality results. Embedded methods are some feature selection methods included in machine learning algorithms, such as feature selection method based on information gain used in decision tree classification algorithm. Feature selection has achieved good results in many fields, such as gene array analysis [11], intrusion detection [12], and text mining [13].

In order to improve the accuracy of predicting students’ performance, researchers have used many feature selection methods in previous studies. Many researchers rely on the experience of education experts to manually select the features that have a significant impact on students’ performance from the originals [14–20]. These features mainly include demographic information such as gender, age, and nationality, previous curriculum and quiz scores, and other previous academic performance features. Many researchers also use feature selection methods in machine learning for automatic selection. Some researchers use dependency-based feature selection methods to generate better subset of original features to improve accuracy of prediction [20–23], and Gene Algorithm is another feature selection method that researchers often use to predict students’ performance [9, 24]. Xing et al. comprehensively used five filtering methods based on dependency, information gain, information gain rate, relief, and symmetric uncertainty and took the intersection of the results of the five methods as the result of feature selection [25]. The results of these researches show that feature selection methods can effectively improve the accuracy of predicting students’ performance, but there is no comparative analysis of different feature selection methods used in this field.

In this study, we proposed a hybrid feature selection method named RnkHEU to improve the accuracy of predicting students’ performance. This novel feature selection method generates the set of candidate features by ranking and Naive Bayes (NB) classifier firstly and then uses heuristic method to generate the final results of feature selection.

The main contributions of this paper are as follows.

1. We compared the performance of the four major evaluation criteria which are dependence, distance, information metric, and consistency in predicting students’ performance through experiments.

2. We compared the performance of the two major search strategies which are sequential selection and heuristic search strategies in predicting students’ performance through experiments.

3. We propose a hybrid feature selection method named RnkHEU, which can further improve the accuracy of predicting students’ performance. This hybrid method uses rank search and Naive Bayes classifier to generate sets of candidate features and uses heuristic search to generate final results of feature selection.

The remainder of this paper is organized as follows: The preliminaries of this research are introduced in Section 2. In Section 3, we discuss the feature selection methods and their effects used by researchers to establish students’ performance prediction models in previous studies. The novel hybrid feature selection method named RnkHEU is proposed in section 4. The experimental results and discussion of the performance of different evaluation criteria, search
strategies, and RnkHEU are in Section 5. Our conclusion and some suggestions for future work are provided in Section 6.

2. Preliminaries and Related Work

2.1. Predicting Students’ Performance. Predicting students’ performance is one of the most important issues in the field of EDM. By predicting students’ performance, we can identify the risk of students’ academic failure in the learning process as early as possible, so as to intervene and guide in advance. It can also provide recommendation basis for personalized learning and provide support for educational administrators to make decisions by analyzing the factors affecting students’ performance. The purpose of predicting students’ performance is to obtain academic performance according to features of students. These features can be demographic characteristics of students (such as gender, age, and family background), behavior of e-learning (such as number of quiz or visiting learning resources), or previous academic performance (such as curriculum scores).

The problem of predicting students’ performance can be described as follows. There is a set of students denoted as \( S = \{ s_1, s_2, \ldots, s_n \} \); each \( s_i \) in \( S \) has \( m \) features represented by a vector \( d_i = [d_{i1}, d_{i2}, \ldots, d_{im}] \). Students’ performance is described by a set of categories denoted by \( P = \{ p_1, p_2, \ldots, p_m \} \). Each \( p_i \) in \( P \) can be one of a dichotomy such as pass or fail, or one of multiclass such as low, medium, good, and excellent; \( p_i \) also can be a number such as GPA. Building prediction model for students’ performance is to establish the function \( f \) according to statistical regularity such as entropy or prior probability contained in students’ historical academic performance records. The process of establishing students’ performance prediction model is briefly described in Figure 2. Predicting students’ performance is using function \( f \) and a new feature vector \( s_i \) to get the performance category \( p_{i, y} \), that is, \( p_{i, y} = f(d_i) \).

2.2. Feature Selection. A feature or attribute refers to an aspect of the dataset; features can be numeric or nominal. Feature selection is a process of selecting the best features among all the features based on specific evaluation criteria. Through feature selection, irrelevant and redundant features that have little influence on the output can be removed to improve the performance of machine learning tasks. We assume that \( X \) is a dataset with \( n \) instances \( X = \{ x_1, x_2, \ldots, x_n \} \); all features of \( X \) are set \( F = \{ f_1, f_2, \ldots, f_d \} \); that is, each \( x_i \) has \( d \) features. \( F’ \) is a subset of \( F \), and \( R \) is a function used to evaluate a subset of features. Let \( FS \) be the result of feature selection for \( F \), \( FS = \{ F’ \} \), where \( \max (R(F’)) \leq F \). Feature selection removes the redundant, irrelevant, or noisy data, reduces the dimensionality of the feature space, and increases speed and accuracy of the machine learning algorithm [26]. The framework of all feature selection method is shown in Figure 3.

As can be seen from Figure 3, feature selection method includes four factors: search starting-point, search strategy, evaluation criterion, and stopping criteria. Researchers have proposed a huge number of feature selection methods [7, 27], which can be summarized in Table 1.

Among the four factors of feature selection, search strategy and evaluation criteria are the most important. \( N \) features will generate \( 2^N \) subsets, so feature selection is an NP-hard problem, and the exhaustive search strategy is not available. The forward or backward search strategies are very simple. They add features one by one from the starting-point of empty or remove individual features from the starting-point of all. Heuristic search method includes Genetic Algorithm (GA) [28], Particle Swarm Optimization (PSO) [29], Ant Colony Optimization (ACO) [30], etc. This kind of search strategy adopts the idea of gradual approximation which is an effective way to find the global optimal solution. With the increasing of features, heuristic search methods will suffer from problems of huge search space, local optimization, and poor performance. Xue et al. made a series of improvements to PSO and improved the performance of PSO to solve large-scale feature selection problems by introducing an adaptive mechanism [31, 32]. Song et al. proposed that hybrid feature selection method consists of three phases, which used the filter-based and clustering methods to reduce the search space of PSO [33]. Another optimized PSO algorithm is called VS-CCPSO. It uses the idea of “divide and conquer” to partition whole search space [34] and can classify relevant features into the same subspace with a low computational cost. Zhang et al. proposed an effective feature selection method based on firefly algorithm (FFA), which is called return-cost-based binary FFA (RcBBFA). The proposed method has the capability of preventing premature convergence and is particularly efficient [35].

In addition to the accuracy of machine learning algorithm, most of the other evaluation criteria are based on the
influence of an individual feature on the output. F-test [36], Pearson Correlation Coefficient [37], and so on are commonly used to measure the dependence between features and outputs and are suitable for numeric features with orthonormal distribution. Researchers use Euclidean distance [38], Mahalanobis distance [39], and other distance evaluation criteria to calculate the distance between features and output to measure the impact of features on output. While evaluation criterion of dependence and distance can only reveal linear relationship between feature and output, evaluation criterion based on information theory overcomes this defect, including information gain [40], information gain rate [40], and mutual information [41]. In addition, there are chi-square [42], significant, and other criteria, most of which are used to evaluate an individual feature and rank the features according to the scores generated by evaluation.

According to whether the evaluation criteria rely on a machine learning algorithm, filter method and wrapper method are the two most important kinds of feature selection methods. The filter method generally uses information gain, dependence, distance, and other evaluation criteria to evaluate a single feature and adopts forward or backward search strategy to add or remove the features with the highest or lowest score from the subset in turn such as CFS [43], Relief [8], mRMR [44], etc. Wrapper method uses accuracy of machine learning algorithm as the criteria to evaluate the feature subset. There are some feature methods embedded in specific machine learning algorithms. For example, C4.5 [45] and CART [46] include feature selection methods based on information gain rate. Generally, filter methods are not associated with specific machine learning algorithm, which run fast, but their accuracy is not easy to control. Wrapper methods are associated with machine learning algorithm such as clustering or classification, which have high time complexity but better accuracy.

Table 1: Factors of feature selection methods.

<table>
<thead>
<tr>
<th>Search starting-point</th>
<th>Search strategy</th>
<th>Evaluation criterion</th>
<th>Stopping criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empty</td>
<td>Exhaustive</td>
<td>Dependence</td>
<td>Accuracy</td>
</tr>
<tr>
<td>All</td>
<td>Forward</td>
<td>Distance</td>
<td>Searching time</td>
</tr>
<tr>
<td>Random</td>
<td>Backward</td>
<td>Information metric</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Random (heuristics)</td>
<td>Consistency</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy of specific algorithm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Iteration times</td>
<td></td>
</tr>
</tbody>
</table>

2.3. Related Work. In previous studies, many researchers rely on the experience of educational experts to select features manually. Saa et al. reviewed 36 studies from 2009 to 2018 [47] and identified nine types of factors influencing students' performance, among which the most commonly used four types are students' previous grades and class performance, students' e-learning activity, students' demographics, and students' social information, and the most frequently used single factor is Cumulative Grade Point Average (CGPA). Francis et al. divided the features that affect students' performance into four categories: demographic features, academic features, behavioral features, and extra features [48]. Demographic features include student number, name, gender, age, etc.; academic features include school grade, class level, semester, etc.; behavioral features include discussion, access to resources, comments, and other learning behaviors in online learning environment.

In addition to manual selection, researchers also use filter-based or wrapper-based feature selection methods in the process of establishing students' performance prediction model. Most researchers use correlation-based filter method [18, 20, 22, 23, 49–56], and the correlation between features and students' performance is measured by Pearson correlation coefficient (1). In particular, Bertolini et al. integrated the correlation-based filter method with cross validation and performed correlation-based feature selection in each fold of data [57].

The information gain-based filtering method is also one of the commonly used methods by researcher [58, 59]. Similar to the correlation-based filtering method, this method calculates the information gain of each feature and students' performance and selects the best feature from the data.

Since Gene-based algorithms are commonly used to solve global optimization problems, many wrapper methods based on Gene algorithms have been used. Wutzl et al. used genetic-based algorithm (GA) [60] to imitate the natural selection process of biological evolution for feature selection and continuously adjusted the feature selection method through the evaluation of prediction results, so as to achieve the best feature selection in the process of approaching the best prediction results [9, 24]. Turabieh et al. used a special method called binary GA, which uses artificial neural network as the feedback model of prediction results, and continuously optimized feature selection by performing three operations similar to natural evolution of selection, crossover, and mutation on Feature Selection Scheme [9, 24]. The method of feature selection based on Genetic Algorithm proposed by Farissi et al. is called GAFS, which is similar to the two methods mentioned above. The difference is that the results of KNN, decision tree, random forest, and Naive Bayes are used to optimize feature selection [61]. Shahiri et al. proposed a hybrid feature selection method integrating filtering and wrapper. This method uses fast correlation-based feature selection (FCBF) firstly to generate subsets of features and then uses Wrapper Sequential Forward Selection (WSFS) to generate the final results. However, there is no detailed description of this hybrid method in that literature [62].

Zaffar et al. present an analysis of the performance of filter feature selection algorithms and classification algorithms on two different student datasets. The results, obtained from different feature selection algorithms and
classifiers on two student datasets with different number of features, help researchers to find the best combinations of filter feature selection algorithms and classifiers [63]. But that research only involves the performance of filter methods, and there is no comprehensive analysis between different factors of feature selection methods. In order to improve the prediction accuracy and identify the optimal features for building productive strategies for the improvement in students’ academic performance, Zaffar et al. proposed hybrid feature selection framework with the utilization of cosine-based fusion [64].

3. Proposed RnkHEU

3.1. Framework of RnkHEU. From the experimental results in 4.2 and 4.3, it can be seen that the subset composed of the best features is not the best subset of features, and global heuristic search strategy can be used to find the subset of features with higher prediction accuracy. However, there are some obvious faults in heuristic search. Firstly, this kind of search strategy uses the accuracy of the classifier as the evaluation criteria of the feature subset, and their time complexity is dependent on the selected classifier. Secondly, the starting-point of heuristic search is feature subsets generated randomly, which leads to the unstable result of feature selection. Thirdly, when the number of features is large, the time complexity of feature selection increases exponentially with the number of features, and the quality of result is poor.

In order to further improve the speed of feature selection method using heuristic search strategy and the accuracy of predicting students’ performance, we need to provide a better candidate set of features for heuristic search, which has smaller size and higher quality than the original set of features. For the above objectives, we design a hybrid feature selection method named RnkHEU, which further improves the quality of feature selection results by providing better candidate set of features for heuristic methods. This method uses forward search to generate the subset of features from ranking features (FL) and selects the subset with the highest accuracy of classifier as the candidate set (CandSubset) and then performs heuristic search on the candidate set to obtain the final result of feature selection (FS). The flowchart of RnkHEU is shown in Figure 4.

The hybrid feature selection method proposed by us uses both rank and heuristic search strategies. Therefore, we call it rank and heuristic (RnkHEU). RnkHEU can also be regarded as a framework of concrete algorithm, in which the evaluation criteria of features, the classifier used by the evaluation subset, and the specific heuristic algorithm can be flexibly specified. The pseudocode of RnkHEU is shown in Algorithm 1.

As described in Algorithm 1, RnkHEU uses the specified evaluation criteria E to score all features in D (Line 1) and arranges all features in descending order according to their scores to generate FL (line 2). The GenerateSubSetInForward function is used to generate feature subsets; if FL = \{f_1, f_2, \ldots, f_n\}, then SL = \{\{f_1\}, \{f_1, f_2\}, \ldots, \{f_1, f_2, \ldots, f_n\}\} (line 3). For each subset in SL, the specified classifier is used to evaluate its performance, and the subset with the highest accuracy is selected as the candidate set of features CandSubset (line 5–11). Finally, RnkHEU uses the specified heuristic search method H to search in the candidate set of features to produce the final result of feature selection FS (line 13). According to the description in Figure 4, when we specifically implement RnkHEU, the specified evaluation criterion E is Pearson correlation coefficient, the classifier C is NB, and the heuristic search algorithm H is GA.

3.2. Scoring and Ranking Features to Get FL. From the description of Algorithm 1, we can see that RnkHEU can be a framework for designing concrete feature selection algorithms. In addition to the input dataset, it mainly has three parameters, evaluation criteria, classifiers for evaluating feature subsets, and heuristic search methods. Researchers can flexibly adjust according to the characteristics of different datasets. From the characteristics of students’ historical academic datasets, the features that have a significant impact on the output, such as the number of visits to digital learning resources, grade scores, and CGPA, have obvious characteristics of normal distribution, and there is a direct linear correlation with academic performance. According to the experimental results in V. B, we also find that different types of evaluation criteria have very similar performance, and the performance based on correlation Pearson coefficient is more stable. Therefore, in the process of implementing RnkHEU, we use Pearson correlation coefficient (1) as the evaluation criterion:

\[
R(f_i, c) = \frac{\text{Cov}(f_i, c)}{\sqrt{\text{Var}(f_i)\text{Var}(c)}},
\]

where \( R(f_i, c) \) denotes correlation between feature \( f_i \) and class \( c \), \( \text{Cov}(f_i, c) \) denotes covariance of features and class, and \( \text{Var}(f_i, c) \) denotes variance of features and class.

The score of each feature \( s_i \) can be obtained by evaluating each feature \( f_i \) in the dataset using Pearson correlation coefficient, and FL can be obtained by ranking all features in descending order according to their scores. If there are \( n \) features in dataset, then \( \text{FL} = \{s_1, s_2, \ldots, s_n\} \), while \( s_i \geq s_j \) and \( 1 \leq i < j \leq n \).

3.3. Generating Feature Subsets Using Forward Method to Get SL. After scoring and ranking, we can easily use the forward method to select the features in FL one by one to generate feature subsets; that is, if \( \text{FL} = \{s_1, f_1, s_2, \ldots, f_n\} \), all subsets generated are \( \text{SL} = \{\{f_1\}, \{f_1, f_2\}, \ldots, \{f_1, f_2, \ldots, f_n\}\} \). If \( \text{FL} \) contains \( n \) features, \( n \) feature subsets will be generated.

3.4. Using Classifier to Evaluate Subsets in SL. In previous studies about predicting students’ performance, most researchers used supervised learning classification algorithm to build the prediction model. Saa et al. summarized the classification algorithms used in 36 studies from 2009 to 2018 [65]. A total of 74 different classification algorithms were used in these studies, among which the most used eight algorithms...
were Naive Bayes, SVM, logistic regression, KNN, decision tree (ID3, C4.5, C5.0), and ANN. Since SVM and logistic regression are more suitable for binary classification, and students’ performance is usually identified by multiclass, Naive Bayes, KNN, C4.5, and ANN are the most commonly used classifiers to predict students’ performance. Naive Bayes is based on the Bayes theorem proposed by Bayes, which can get the posterior probability through the known prior probability. Let two independent random variables \( X \) and \( Y \), \( P(X, Y) \) denote their joint probability, and let \( P(Y) \) and conditional probability \( P(X|Y) \) be prior probability; then the posterior probability \( P(Y|X) \) can be calculated by Bayesian (2). For Naive Bayes algorithm, if a feature vector described by \( n \) attributes \( X = \{x_1, x_2, \ldots, x_n\} \), then the prior probability \( P(X|Y) \) can be decomposed into the product of multiple vectors (3), and then the posterior probability \( P(Y|X) \) can also be calculated through the prior probability (4). If \( Y = \{y_1, y_2, \ldots, y_m\} \) is a set of students’ performance with \( m \) values, then the

\[
P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}
\]

Figure 4: Flowchart of RnkHEU.

Algorithm 1: Pseudocode of RnkHEU.
final classification result is the category represented by \( y_i \) in \( P(y_i | X) \) with the largest probability (5). It has good interpretability and high training efficiency for a large number of training data and is not very sensitive to noise [66]. It is especially suitable for evaluating the performance of feature subsets. Therefore, we use Naive Bayes as a classifier to evaluate the performance of feature subsets.

\[
P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)} \tag{2}
\]
\[
P(X | Y) = \prod_{i=1}^{n} P(x_i | Y), \tag{3}
\]
\[
P(Y | X) = \frac{P(Y) \prod_{i=1}^{n} P(x_i | Y)}{P(X)} \tag{4}
\]
\[
P(y_i | X) = \frac{P(Y = y_i) \prod_{i=1}^{n} P(x_i | Y = y_i)}{P(X)} \tag{5}
\]

3.5. Heuristic Search in CandSubset to Get FS. Feature selection is an HP-hard problem; that is, unless exhaustive search is used, the best feature subset cannot be obtained. However, the size of search space is \( 2^n \), and exhaustive search is not acceptable in time and efficiency, so heuristic search is an inevitable choice to obtain approximate global optimal results. The experimental results in Section 5 show that the classical GA has better performance in predicting students’ performance. Therefore, we use GA [28] as a heuristic search method in RnkHEU. The flowchart of generating the final result in CandSubset using GA and NB is shown in Figure 5.

According to the description of Figure 5, we first use binary coding to generate the initial population (IP) of CandSubset containing \( n \) feature subsets, and each feature subset is represented by a binary sequence (Figure 6). If CandSubset contains \( m \) features, every binary sequence contains \( m \) bits, and each bit represents whether the corresponding feature exists. NB is used to evaluate each subset in IP, top-k subsets are selected for mutation (Figure 7) and crossover (Figure 8) in proportion to generate new populations (NP), and the above process is repeated until the end of the iteration.

4. Experimental Results and Discussion

4.1. Datasets and Experimental Setting. In previous studies, most of the datasets used to build students’ performance prediction model by researchers are private. These datasets are collected from the information management system used by educational institutions or collected manually through questionnaires. Because these datasets are private, we cannot compare the feature selection methods used in these studies. Therefore, we collected the students’ historical academic dataset from UCI machine learning repository [67] and Kaggle [68] which are two world-famous open machine learning data warehouses as the experimental datasets. These datasets are real students’ historical academic data collected by different researchers in various educational institutions and learning support system (identifiers in brackets):

(1) xAPI-Edu-Data [69] (D1): this is an educational dataset which is collected from learning management system (LMS) called Kalboard 360.

(2) Students Performance [70] (D2): this dataset consists of the marks secured by the students in various subjects.

(3) Student Academics Performance [71] (D3): the dataset tries to find the end semester percentage prediction based on different social, economic, and academic attributes.

(4) Student Performance on an Entrance Examination [72] (D4): this dataset contains data of the candidates who qualified the medical entrance examination for admission to medical colleges of Assam of a particular year and collected by Prof. Jiten Hazarika.

(5) Student-por [73] (D5): this dataset approaches student achievement in secondary education of two Portuguese schools provided regarding the performance in Portuguese language (por).

(6) Student-mat [73] (D6): this dataset approaches student achievement in secondary education of two Portuguese schools provided regarding the performance in Math (mat).

(7) Turkiye Student Evaluation [74] (D7): this dataset contains evaluation scores provided by students from Gazi University in Ankara (Turkey). There is a total of 28 course specific questions and additional 5 attributes.

(8) Student Performance [75] (D8): this dataset contains SGPA and KT of four semesters, as well as the final grade of students’ performance.

The important properties of experimental datasets are shown in Table 2.

We use the accuracy of classifier to evaluate the performance of features subset in following experiments. Accuracy is the ratio of the number of samples’ correct prediction to the number of total samples (6). When using the classifier for verification, we use the tenfold cross validation [76], and the final accuracy is the average of the accuracy of ten rounds.

\[
\text{Accuracy} = \frac{\text{the number of samples correct prediction}}{\text{the number of total samples}} \times 100\%. \tag{6}
\]

The significant objective of predicting students’ performance is to identify students who are at the risk of academic failure as soon as possible and remind educators to intervene and guide timely. Therefore, in order to further verify the performance of RnkHEU, we use the precision of predicting students’ academic failure (7) to test the impact of different feature selection methods on different classifiers.
Binary coding features in CandSubset

generate n feature subsets of CandSubset Randomly (IP)

Using NB to evaluate the performance of each subset

Select subsets for mutation and crossover to generate new feature subsets (NP)

iteration over N

Select the subset with the best performance

FS={f_1,...,f_m}

**Figure 5:** Flowchart of generating the final result using GA and NB.

<table>
<thead>
<tr>
<th></th>
<th>f_1</th>
<th>f_2</th>
<th>f_3</th>
<th>...</th>
<th>f_m</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>s_2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s_n</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

**Figure 6:** Generate initialization population using binary coding.

<table>
<thead>
<tr>
<th></th>
<th>f_1</th>
<th>f_2</th>
<th>f_3</th>
<th>...</th>
<th>f_m</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_i</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

**Figure 7:** Mutation: A subset is negated at random positions.

<table>
<thead>
<tr>
<th></th>
<th>f_1</th>
<th>f_2</th>
<th>f_3</th>
<th>...</th>
<th>f_m</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_i</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>s_j</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

**Figure 8:** Crossover: Two subsets are spliced at random positions.
We use tenfold cross validation to obtain the accuracy and precision of classifiers used in these experiments.

4.2. Performance of Different Evaluation Criteria. The purpose of this group of experiments is to investigate the performance of different evaluation criteria in predicting students’ performance. First, we use an evaluation criterion to score each feature in the experimental dataset and rank all the features in descending order according to their scores. The sorting feature sequence \( \{f_1, f_2, \ldots, f_n\} \) can generate \( n \) feature subsets \( \{f_1\}, \{f_1, f_2\}, \ldots, \{f_1, f_2, \ldots, f_n\} \). Then, we use the accuracy of Naive Bayes to evaluate the performance of feature subset, select the subset with the highest accuracy as the result, and regard the highest accuracy as the performance of this criterion. Finally, we compare the performance of the four evaluation criteria listed in Table 3 on eight experimental datasets. The flowchart of this group of experiments is shown in Figure 9.

We selected four types of representative evaluation criteria for experimental evaluation, which are Pearson correlation coefficient, Mahalanobis distance (8), information gain rate (9), and chi-square (10). In order to avoid the influence of programming style on performance of algorithm, we selected algorithms provided by open source machine learning tool named Weka (v3.8.5) [77].

Pearson correlation coefficient is one of the most widely used correlation measurement methods in the field of machine learning. It is widely used to measure the linear correlation between two random variables. ReliefF uses distance to measure the similarity between different samples, and the most commonly used distance measure is Cartesian distance. Information gain rate is based on information theory, which uses the entropy of features to measure the importance of features. ChiSquare is a nonparametric test method, which compares the correlation between two random variables by comparing the coincidence between the theoretical frequency and the actual frequency.

The results of this group of experiments are shown in Tables 4 and 5. To make results easy to read and analyze, we use ID instead of name to represent a feature in the dataset. From the experimental results in Tables 4 and 5, we can get the inspirations as follows.

Firstly, although every four orders of features generated by different evaluation criteria are different on a dataset, the highest prediction accuracies of different subsets of features are very similar, especially D1, D2, and D7. It is demonstrated that the features of high dependence and consistency with the attributes representing students’ performance have a significant impact on students’ academic performance. From the perspective of information theory, these features have a lot of information entropy and can be used to distinguish different types of students effectively.

Secondly, prediction accuracies based on feature subsets generated by feature selection methods using different evaluation criteria are very close, indicating that these four types of evaluation criteria have the similar performance in predicting students’ performance.

Thirdly, the above results also reflect that features of students’ historical academic dataset which have significant impact on output have linear relationship with academic performance. Besides, the number of instances and features in our experimental datasets are small and these datasets have been preprocessed to eliminate the noise, which are important limitations of the experimental results.

4.3. Performance of Different Search Strategy. The purpose of this group of experiments is to compare the performance of different search strategies in predicting students’ performance. Because exhaustive search is not available, forward and backward search strategies are similarly essential, so we choose forward and heuristic search strategies for comparison. The feature selection using forward search strategy is to select the best features one by one in “Best First” way after ranking features. In this group of experiments, we use correlation-based evaluation criteria to rank all features. We choose the most commonly used Genetic Algorithm (GA) [28], Particle Swarm Optimization (PSO) [29], and Multiobjective Evolutionary Algorithm [78] proposed in recent years as the experimental algorithms of heuristic search strategy; their parameters are shown in Table 6. We also apply the accuracy of Naive Bayes as the performance of different search strategies.

\[
\text{Precision}_{\text{fail}} = \frac{\text{the number of samples correct prediction failure}}{\text{the number of samples with failure}} \times 100\%.
\]
The results of different search strategies in this group of experiments are shown in Table 7. The performance comparison of these search strategies is shown in Figure 10.

From Table 7, we can see that, firstly, the results of feature selection methods using forward search strategy and heuristic search strategy are very different, which reflects the essential differences between the two search strategies. Secondly, the results of the three heuristic search algorithms are significantly different on experimental datasets except D4 and D8, which reflects that the results of the heuristic algorithm are unstable. Thirdly, the result size of the feature selection method using heuristic search strategy is smaller than that using forward search strategy except D3. A smaller result can speed up the training of the prediction model significantly and improve the prediction accuracy.
As can be seen from Figure 10, heuristic search algorithm has better performance than forward method on experimental datasets except for D5. Three commonly used heuristic search algorithms have their own advantages in different datasets, but they are all better than forward method. It is further demonstrated that feature subset composed of good features may be not a good subset, and feature selection method using heuristic search strategy can produce feature subset with smaller size and higher quality.

### 4.4 Performance of RnkHEU

In order to evaluate the performance of proposed method, we compare RnkHEU with the ranking-based forward search and GA. The three parameters to be specified in RnkHEU are E-Pearson correlation, C-NB, and H-GA. In this group of experiments, we used the accuracy of four classificational algorithms NB, C4.5, MLP, and KNN, which are the most commonly used to predict students’ performance, to compare the performance of three feature selection methods; their significant parameters are listed in Table 8.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Evaluation criteria</th>
<th>Result</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D5</td>
<td>Correlation</td>
<td>32, 31, 15, 16, 30, 21, 9, 7, 20, 23, 28, 29</td>
<td>59.75</td>
</tr>
<tr>
<td></td>
<td>GainRatio</td>
<td>32, 31, 15, 21, 16, 30, 9, 27, 7, 23, 28, 3, 22, 8, 20, 29</td>
<td>58.48</td>
</tr>
<tr>
<td></td>
<td>ReliefF</td>
<td>32, 31, 15, 9, 30, 16, 23, 18, 8, 7, 28, 29, 6, 17, 22, 21, 20</td>
<td>59.49</td>
</tr>
<tr>
<td></td>
<td>ChiSquared</td>
<td>32, 31, 15, 30, 9, 3, 7, 28, 29</td>
<td>57.97</td>
</tr>
<tr>
<td>D6</td>
<td>Correlation</td>
<td>32, 31, 15, 21, 1, 4, 16, 27, 13, 22, 30</td>
<td>68.88</td>
</tr>
<tr>
<td></td>
<td>GainRatio</td>
<td>32</td>
<td>68.72</td>
</tr>
<tr>
<td></td>
<td>ReliefF</td>
<td>32</td>
<td>68.72</td>
</tr>
<tr>
<td></td>
<td>ChiSquared</td>
<td>32</td>
<td>68.72</td>
</tr>
<tr>
<td>D7</td>
<td>Correlation</td>
<td>32, 20, 1, 3</td>
<td>44.57</td>
</tr>
<tr>
<td></td>
<td>GainRatio</td>
<td>32, 3</td>
<td>45.29</td>
</tr>
<tr>
<td></td>
<td>ReliefF</td>
<td>32, 20</td>
<td>43.72</td>
</tr>
<tr>
<td></td>
<td>ChiSquared</td>
<td>32, 20</td>
<td>43.72</td>
</tr>
<tr>
<td>D8</td>
<td>Correlation</td>
<td>8, 2, 4, 6, 7, 5, 1, 3</td>
<td>57.94</td>
</tr>
<tr>
<td></td>
<td>GainRatio</td>
<td>8, 2, 2, 6, 3, 4, 1, 5</td>
<td>57.94</td>
</tr>
<tr>
<td></td>
<td>ReliefF</td>
<td>1, 5, 7, 3, 8, 4, 6</td>
<td>58.20</td>
</tr>
<tr>
<td></td>
<td>ChiSquared</td>
<td>7, 3, 5, 1, 8, 2, 6</td>
<td>58.34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic Algorithm (GA)</td>
<td>populationSize-20, crossoverProb-0.6, mutationProb-0.033, maxGenerations-20</td>
</tr>
<tr>
<td>Particle Swarm Optimization (PSO)</td>
<td>populationSize-20, iterations-20, socialWeight-0.33, inertiaWeight-0.34, inertiaWeight-0.33</td>
</tr>
<tr>
<td>Multiobjective Evolutionary Algorithm</td>
<td>populationSize-100, Generations-20</td>
</tr>
</tbody>
</table>

### Table 5: Result and performance of different evaluation criteria (D5–D8).

### Table 6: Names and parameters of heuristic algorithms used in this group of experiments.

### Table 7: The results of different search strategy.
We selected four datasets D3, D5, D6, and D7 with more features as the experimental datasets of this group. The final results of feature selection $FS$ generated by RnkHEU on four experimental datasets are shown in Table 9. As can be seen from Table 9, RnkHEU further eliminates the features that affect the accuracy of predicting in $CandSubset$, reducing the size about 30%. Fewer features can not only improve the accuracy of prediction, but also improve the training speed of classifier, especially for MLP.

The accuracy of different classifiers caused by these three methods is shown in Table 10.

As can be seen from Table 10, the results of RnkHEU make different classifiers achieve higher accuracy of prediction; the accuracy of different classifiers is improved by nearly 10%. It also demonstrated that providing smaller and better candidate sets of features for heuristic search algorithm can further improve their performance, and forward method is one of the effective methods to generate better candidate sets of features.

The precision of different classifiers caused by these three methods for predicting students' academic failure is shown in Table 11.

From Table 11, we can see that the results generated by RnkHEU can significantly improve the precision of predicting academic failure with commonly used classifiers, except in D3. It is demonstrated that the result of feature selection generated by RnkHEU is the most significant factors of students' academic failure and can help instructor

| Table 8: Names and parameters of classifier used in this experiment. |
|-----------------|-----------------|
| Name            | Parameters      |
| Naive Bayes (NB)| None            |
| K-Nearest Neighbours (KNN)| KNN-1, distanceWeighting-NO nearestNeighbourSearchAlgorithm-LinearNNSearch distance Function-EuclideanDistance |
| C4.5            | confidenceFactor-0.25, minNumObj-2 numFolds-3 |
| MultiLayerPerceptron (MLP)| hiddenLayers-a, learningRate-0.3 momentum-0.2, trainingTime-500 |

| Table 9: Final result comparison of forward and RnkHEU. |
|-----------------|-----------------|-----------------|
| Dataset         | Forward ($CandSubset$) | RnkHEU           |
| D3              | 10, 17, 1, 3, 6   | 1, 6, 17         |
| D5              | 32, 31, 15, 16, 30, 21, 9, 7, 20, 23, 28, 29 | 9, 15, 16, 20, 28, 29, 30, 31, 32 |
| D6              | 32, 31, 15, 21, 1, 4, 16, 27, 13, 22, 30 | 13, 16, 27, 30, 31, 32 |
| D7              | 32, 20, 1, 3     | 3, 32            |

Figure 10: The performance comparison of different features selection methods.
provide more accurate intervention and guidance to learners.

5. Conclusion and Future Work

Feature selection is one of the most important steps to predict students’ performance. We have conducted an empirical study on the performance of the feature selection methods using different evaluation criteria and search strategies in the performance prediction of students and proposed a hybrid feature selection method named RnkHEU which improves the accuracy of prediction. Through experiments, firstly, we find that evaluation criteria based on dependence, distance, information metric, and consistence all work well in predicting students’ performance. It also reflects that the features which have a significant influence on output used for training prediction model of students’ performance have a significant linear correlation with the output. Secondly, we find that the feature selection method using heuristic search strategy can obtain higher accuracy of predicting students’ performance than the sequential selection strategy. The main reason is that the sequential selection of the best features does not necessarily get the best feature subset, and the heuristic search strategy is more likely to get better subsets from all features. Thirdly, we find that the hybrid feature selection method that we proposed named RnkHEU can achieve higher accuracy in the most commonly used classification algorithm of students’ performance prediction. This shows that providing better candidate feature set for heuristic algorithm can further improve the performance of feature selection method, and using forward strategy to obtain the best feature subset is one of the effective methods to generate candidate feature set.

We will continue to study the problem of feature selection in predicting students’ performance in the following three aspects. Firstly, we will gather datasets of students’ academic performance with larger size and more features and conduct more experiments to further investigate the performance of different feature selection methods. Secondly, we will design more effective methods which are used to generate the candidate set of features needed for heuristic search to get better accuracy of predicting students’ performance and investigate and design unsupervised feature selection methods used for predicting students’ performance. Thirdly, we will investigate the performance of the integration of different feature selection methods and classification algorithms.

Data Availability

The experimental datasets used to support the results of this study are cited from Kaggle and UCI machine learning library and references are also added in the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by a funding project for top discipline talents of the Anhui Provincial Department of Education in 2020 under Grant gxbjzd2020101.

References


an overview,” in Proceedings of the SESYR Sustainable Education through European Studies for Young Researchers Jean Monnet Module, Iasi, Romania, October 2020.


datasets using data mining tools: a case study,” in *Cybern
tics and Algorithms in Intelligent Systems, CSOC2018, Advances in
Intelligent Systems and Computing*, R. Silhavy, Ed., vol. 765,
Springer, Cham, Switzerland, 2018.
[73] P. Cortez and A. Silva, “Using data mining to predict sec-
ondary school student performance,” in *Proceedings of 5th
Future Business Technology Conference (FUBUTEC 2008)*,
University of California, School of Information and Com-
puter Science, Irvine, CA, USA, 2013.
[76] K. Battula, “Research OF machine learning algorithms using
K-fold cross validation,” *International Journal of Engineering
[77] F. Eibe, I. Witten, T. Len, M. Hall, G. Holmes, and J. C. Sally,
Weka: practical machine learning tools and techniques with
java implementations,” *Acm Sigmod Record*, vol. 31, no. 1,
[78] F. Jiménez, G. Sánchez, J. M. García, G. Sciavicco, and
L. Miralles, “Multi-objective evolutionary feature selection for
online sales forecasting,” *Neurocomputing*, vol. 234, pp. 75–
92, 2017.