Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Construction of Big Data-Driven Students' Career Planning and Innovation and Entrepreneurship Education Path by Integrating Support Vector Machine Algorithm

Gang Wang[1,✉], Yongling Qian[2], Jing Zhao[3], Yifan Xue[4]

[1] *The Office of Student Affairs, College Student Employment Guidance Center, School of Innovation and Entrepreneurship, Shaanxi Technical College of Finance and Economics, Xianyang, Shaanxi, 712000, China*

[2] *The School of Humanities and Arts, Shaanxi Technical College of Finance and Economics, Xianyang, Shaanxi, 712000, China*

[3] *The School of Accounting, Shaanxi Technical College of Finance and Economics, Xianyang, Shaanxi, 712000, China*

[4] *The Academic Affairs Office, Shaanxi Technical College of Finance and Economics, Xianyang, Shaanxi, 712000, China*

ABSTRACT

This study aims to construct an effective pathway for students' career planning and innovative industry education by integrating support vector machine algorithm with big data analysis technology. By effectively integrating multi-source data and combining the improved genetic algorithm for feature selection and extraction of student data, the support vector machine algorithm is used to conduct in-depth analysis of the data related to students' career planning and innovation and entrepreneurship education, to provide students with accurate and personalized career and entrepreneurship guidance, and based on which, the career planning and innovation and entrepreneurship education path is constructed. Experimental analysis of the classification prediction performance of the support vector machine algorithm and comparison with other classification prediction algorithms show that the support vector machine algorithm used in this paper has the highest classification accuracy in the assessment of students' career planning and innovation and entrepreneurship ability, and the model performance is the most stable. The results of the educational experiment show that after using the educational path proposed in this paper, the students' satisfaction with career planning and the mean value of the assessment score of innovation and entrepreneurship ability increase by 70.89% and 170.73%, respectively. The above results fully demonstrate the effectiveness of the educational path constructed in this paper, which provides a useful reference for efficient education and teaching reform.

Keywords: support vector machine, improved genetic algorithm, feature selection, classification prediction, educational pathway

✉ Corresponding author.
E-mail address: wgangqq@sina.com (G. Wang).
Received 30 October 2024; Revised 15 December 2024; Accepted 10 March 2025; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-065

1. Introduction

Career planning is a comprehensive educational work, for college students to carry out learning planning and career planning, so that college students can be clear about their own development direction, so as to promote their own learning behavior. Whether the career planning is effective or not will affect the quality of study and life of college students, and even affect the career performance of college students after they get out of school [1-4]. Therefore, in order to improve the effect of higher education, so that college students can be more smoothly in the career path after graduation, it is necessary to do a good job of career planning education, so that college students can form a certain understanding of their own career [5-8]. Innovation and entrepreneurship education, also known as "double creation" education, is a new trend in higher education in recent years. In reality, the main purpose of innovation and entrepreneurship education is to cultivate students' innovation ability and entrepreneurial literacy, so that college students can flexibly cope with the workplace or start their own business after leaving school [9-12]. Innovation and entrepreneurship education has been paid attention to by many institutions of higher learning, and schools are actively opening specialized courses or simulation bases for entrepreneurial practice to provide students with opportunities and environments for practice. Exercise to cultivate students' innovation and entrepreneurship ability, to ensure that students can cope with the employment problems outside the campus [13-16].

With the rapid development of network technology, information has become one of the basic conditions for success, and "big data" technology has been widely used in many fields [17-18]. Driven by big data, college students' career planning and innovation and entrepreneurship education has ushered in a new development opportunity, big data can make accurate analysis of the current and future situation, and reasonably grasp the development of students' career and innovation and entrepreneurship situation [19-21].

In today's highly competitive job market, college students are faced with severe pressure to choose and develop their careers. In order to better plan their careers, it is crucial to understand the needs, trends and opportunities of the job market. Career big data and analytics is a powerful tool to help college students to plan their careers. Literature [22] reveals that there are many problems such as unclear goals in the career planning of college students nowadays. Based on this, with the help of big data analysis method, the analysis of many colleges and universities has been carried out, and proposed measures to optimize the career planning of college students. Literature [23] emphasizes the importance of career choice for students. And proposed a method based on XGBOOST model clustering center model to predict students' career choices. Based on evaluating previous student behavioral data, it was shown that the proposed method is more effective than other prediction techniques. Literature [24] used machine learning technique XGBoost based on data from specific universities in order to predict the career choices of university students and the results pointed out that XGBoost is able to predict the career choices of students and facilitate educational institutions to conduct educational activities accordingly. Literature [25] examined the application of deep learning in career planning and entrepreneurship for college students. Models affecting college students' career choices and entrepreneurial intentions were developed and validated, and relevant recommendations were made. The results pointed out that deep learning can better help students plan their careers and improve their entrepreneurial abilities. Literature [26] introduces a novel career guidance and career planning system that incorporates machine learning algorithms. By constructing a career matching framework and a career planning recommendation system, and implementing a database security design. The results emphasized the ability of the system to meet the basic requirements of a career planning system. Literature [27] describes that postgraduate precision employment in the context of big data faces advantages such as diversified data capture and disadvantages such as imperfect working mechanisms, and emphasizes that postgraduate precision employment education requires effective strategies such as full integration of advantages and disadvantages, opportunities and

challenges. Literature [28] visualized the career planning path of college students in order to help them adapt to the flexible employment environment. By proposing the LSTM-Canopy algorithm and introducing it into the CP path visualization system for college students, in order to effectively improve the analysis and judgment of career. The validity and reliability of the system is verified through experiments.

With the rapid development of social economy and the continuous progress of science and technology, innovation and entrepreneurship education has become an important part of university education. In order to better promote the development of innovation and entrepreneurship education in colleges and universities, the application of big data analysis and evaluation is gradually becoming a trend. Literature [29] aims to investigate the current situation of innovation and entrepreneurship education system construction in colleges and universities. And it discusses the relationship between big data and the establishment of innovation and entrepreneurship education environment in colleges and universities. It helps to promote the deep integration of big data technology and innovation and entrepreneurship education, and provides reference for "innovation and entrepreneurship education ecosystem". Literature [30] explored the application of adaptive identification weighting algorithm in the grid-based analysis of innovation and entrepreneurship education in colleges and universities. A grid analysis ai teaching model is proposed. The findings reveal that the innovative entrepreneurship education model in colleges and universities based on grid analysis network teaching and adaptive recognition weighting algorithm can diagnose students' teaching data, thus realizing the innovation of big data analysis technology in colleges and universities. Literature [31] constructs a framework model of dual innovation education system in colleges and universities from the dimensions of internal system module and operation mode, and focuses on the internal system module of dual innovation education and the operation mode of three-shift system, which is conducive to improving the level of dual innovation education in colleges and universities. Literature [32] proposes a new approach to college students' dual-creation education in the context of big data. By utilizing the k-mean clustering method, the college students' dual-creation data are deeply mined. The experimental results mention that the proposed method has excellent entrepreneurial data mining accuracy and more accurate risk assessment results. Literature [33] realized the quantitative evaluation of teaching effect based on the analysis of teaching effect of dual-creation courses, and proposed the evaluation method of teaching effect of dual-creation courses based on big data analysis. The simulation results proved that the evaluation results of the method were accurate and reliable. Literature [34] constructed a dual-creation skill training model for college students based on data mining technology. And the local DTRS model and multi-granular decision theory rough set model are proposed. The data mining and data analysis modules of the college students' education model were implemented. The results show that the model realizes the expected requirements and is improved. Literature [35] constructed a framework for the application of big data bi-inventive education based on the problems of bi-inventive education. By determining the content and process of the application of big data in dual-creation education, the strategy of improving dual-creation education with the help of big data is proposed.

In this paper, the support vector machine algorithm is introduced in detail from the perspectives of statistical learning theory, support vector machine model linear separable and indivisible cases, and the selection of kernel function, which provides technical support for the construction of student career planning and innovative entrepreneurship education model based on the support vector machine algorithm in the latter part of the paper. The improved genetic algorithm is utilized to select and propose features for the collected high-dimensional student career and entrepreneurship data, and the superiority of the feature selection algorithm in this paper is demonstrated through comparative experiments. The classification prediction performance of this paper's support vector machine model and the effectiveness of the proposed educational paths are analyzed in depth, highlighting the reasonableness of the educational paths of student career planning and innovation and entrepreneurship constructed based on the support vector machine algorithm in this paper.

2. Support vector machine algorithm

2.1. Statistical learning

2.1.1. Empirical risk minimization. Suppose that data set $(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_t)$ is l independent and identically distributed samples.

It is usually believed that there is a certain relationship between inputs and outputs in a dataset $F(x, y)$, through which new input samples can be accurately predicted when they appear, which is the purpose of machine learning. The problem to be considered in this process is how to find the parameter that minimizes the expectation based on these samples and the set of functions in which they are located $\{f(x, \omega)\}$. ω_0 The expected risk is defined as:

$$R(\omega) = \int L\left(y, f\left(x, \omega\right)\right) dF\left(x, y\right) \tag{1}$$

where ω is the parameter and $L\left(y, f\left(x, \omega\right)\right)$ is the loss function.

In categorization problems, the output variables often correspond to different categories, assuming that the output variable is $y = \{+1, -1\}$, the loss function is:

$$L\left(y, f\left(x, \omega\right)\right) = \begin{cases} 0, if y = f(x, \omega) \\ 1, if y \neq f(x, \omega) \end{cases} \tag{2}$$

In practice, it is difficult to find this unknown relationship, so this paper adopts an alternative strategy of replacing the expected risk with the mean square error of the sample, which is the empirical risk. The empirical risk is:

$$R_{emp}\left(\omega\right) = \frac{1}{l} \sum_{i=1}^{1} L\left(y_i, f\left(x_i, \omega\right)\right) \tag{3}$$

This approach transforms minimizing the expected risk to minimizing the empirical risk and is called the empirical risk minimization criterion (ERM) [36].

2.1.2. Minimization of structural risk. The VC dimension of a function set can be intuitively defined as follows: for h a sample that can be separated by a function set in all 2^h forms, the maximum number of samples h that can be broken up by this function set is its VC dimension.

Statistical learning theory defines the relationship between function sets, empirical risk $R_{cmp}\left(\omega\right)$ and expected risk $R\left(\omega\right)$ as follows:

$$R\left(\omega\right) \leq R_{emp}\left(\omega\right) + \sqrt{\frac{h\left(\ln\left(2l / h\right) + 1\right) - \ln\left(\eta / 4\right)}{l}} \tag{4}$$

where h is the VC dimension and l is the number of samples.

Equation (4) can be simplified as:

$$R\left(\omega\right) \leq R_{emp}\left(\omega\right) + \phi\left(l / h\right) \tag{5}$$

From equation (5), it can be seen that the minimization of the expected risk in the machine learning method, it is necessary that both the empirical risk and the confidence range are as small as possible, and the confidence range is related to the VCV and training samples of the learning method. And when the training samples are certain, $\phi(l / h)$ increases with the increase of h, and the confidence range increases, leading to a decrease in the generalization performance of the learning machine.

When solving classification problems, the VC dimension is often determined according to the form of the classifier, such as a linear classifier, which has a low VC dimension, so when it is used for classification, it tends to get better results. However, in real life, it is often not easy to get the VC

dimension of a complex machine learning method to minimize the expected risk through equation (5). Thus statistical learning theory proposes the principle of structural risk minimization, i.e., the function set $S=\{f(x,\omega),\omega\in\Lambda\}$ is decomposed into an ordered sequence of ordered according to the size of the VC dimension:

$$S_1\subset S_2\subset\cdots\subset S_k\subset\ldots\subset S, S=\bigcup_i S_i \tag{6}$$

Find the function that minimizes the empirical risk in each subset according to Eq. (6).SVM is a better implementation of the SRM principle.

2.2. Support vector machine model

In statistical theory, SVM is one of the youngest and fastest growing machine learning techniques and it is also in a state of continuous development [37]. It adopts the idea of SRM, and compared to the traditional methods based on the ERM idea, SVM has outstanding performance ability in nonlinear and high-dimensional pattern recognition, as well as regression estimation and small-sample learning problems, and the solutions obtained are globally optimal. So it is greatly favored in the whole field of machine learning. The entire schematic diagram of SVM is shown in Fig. 1.

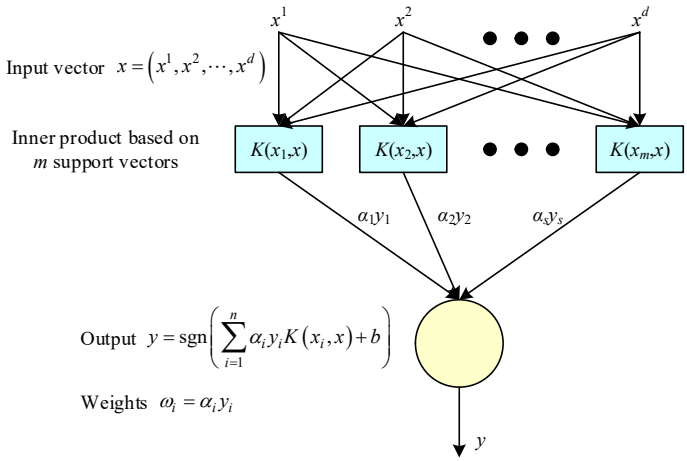


Fig. 1. **Support vector machine mode**

2.2.1. Linear separability. SVMs have been proposed based on the linearly differentiable case where an optimal classification plane exists. The optimal classification plane is defined to satisfy the need for maximizing the classification interval while ensuring accurate classification of the two classes of samples. Classifying the two classes of samples correctly achieves the ERM. While maximizing the interval ensures that the confidence range is as small as possible, the combination of these two satisfies the SRM principle. The purpose of linear separability is to find a hyperplane that will correctly classify the samples of these two classes respectively, and the structure of the whole process is schematically shown in Fig. 2. The two classes of training samples in the figure are represented by circular and square images, P is the optimal classification line, δ is the classification maximum interval, and P_1 and P_2 are the straight lines where the support vectors are located. If generalized to higher dimensions, P represents the optimal hyperplane, and P_1 and P_2 represent the planes where the support vectors are located. From the statistical knowledge, we can get that the optimal hyperplane satisfies the SRM principle, so the generalization ability of the optimal hyperplane is higher than that of other classification hyperplanes.

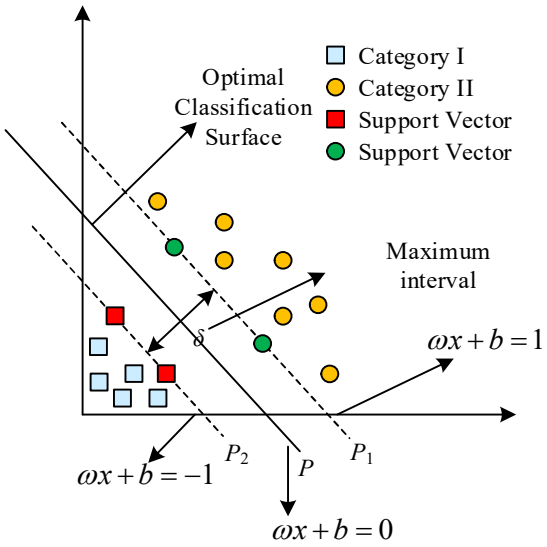


Fig. 2. The optimal hyperplane of a linear separable case

Suppose the set of training samples is $\{x_i, y_i\}$, $i=1,2,\ldots,n$, $x \in R^d$, $y \in \{1,-1\}$ for classification labels. Then the general form of the linear discriminant function in the d-dimensional space can be expressed as: $g(x)=\omega^* x+b$ and the classification surface equation is:

$$\omega^* x+b=0 \tag{7}$$

The decision function is:

$$f(x)=\operatorname{sgn}\left(\omega^* x+b\right) \tag{8}$$

where $\operatorname{sgn}(\cdot)$ is the sign function.

Then the function $g(x)$ is normalized so that $|g(x)|=1$ of the samples closest to the classification surface, at which point the classification interval is $2/\omega$. The problem of maximizing the classification interval can be replaced by the problem of minimizing ω (or ω^2). Then the problem of optimizing the solution of linearly divisible SVM can be summarized as follows:

$$\begin{gathered}\min \frac{\|\omega\|^2}{2} \\ s.t. y_i\left(\omega^* x_i+b\right) \geq 1, i=1,2,\ldots,n\end{gathered} \tag{9}$$

Next, the following function is defined using Lagrange's method for finding extreme values:

$$L(\omega,b,\alpha)=\|\omega\|^2/2-\sum_{i=1}^{n}\alpha_i\left[y_i\left(\omega^* x_i+b\right)-1\right] \tag{10}$$

where $\alpha_i>0$ is called the Lagrange multiplier, and our problem is to find the optimal ω^* and b^* to make the function reach a minimal value. By taking the partial derivatives of ω and b in Eq. (10) and then making them equal to 0, we can convert the original convex quadratic optimization problem into its dual, i.e., in the constraints:

$$\begin{gathered}\sum_{i=1}^{n} y_i\alpha_i=0 \\ \alpha_i \geq 0 i=1,2,\ldots,n\end{gathered} \tag{11}$$

Solve for the maximum value of the following function under α_i:

$$\max Q(\alpha)=\sum_{i=1}^{n}\alpha_i-1/2\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i * x_j\right) \tag{12}$$

if α_i^* is the optimal solution:

$$\begin{cases}\omega^*=\sum_{i=1}^{n}\alpha_i^* x_i y_i\\ b^*=-\frac{1}{2}\omega^*\left(x_r+x_s\right)\end{cases} \tag{13}$$

That is, the weight coefficients of the optimal classification plane can be represented by the training sample vectors, where x_r and x_r are any pair of support vectors in the two categories.

In the training samples, there will be a lot of samples α_i^* will be zero, and only a small part of the samples α_i^* is not zero; the optimal classification plane is determined by the support vectors, i.e., the samples that α_i^* is not zero.

Then the optimal classification function is:

$$f(x)=\operatorname{sgn}\left(\omega * x+b^*\right)=\operatorname{sgn}\sum_{i=1}^{n}\left[\alpha_i^* y_i\left(x_i\cdot x\right)+b^*\right] \tag{14}$$

b^* is the threshold for classification, which can be derived from any of the support vectors.

2.2.2. Linear indivisibility. Except for the linearly separable case described above, most problems in real situations are linearly indivisible. As shown in Fig. 3, this situation often produces a large error using the optimal classification surface in the graph. When the linearly indivisible case is encountered, the optimal classification surface needs to be improved accordingly. What the researcher uses is to add a relaxation term $\xi_i\geq 0$ to the constraints, then the optimization problem of solving the optimal plane becomes solving for the minimum value of Eq. (15):

$$\begin{cases}\phi(\omega,\xi)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i\\ s.t.y_i\left(\omega x_i+b\right)\geq 1-\xi_i,\xi_i>0\end{cases} \tag{15}$$

where C is a constant used to penalize the misclassified samples to a greater degree. As the penalty for misclassification increases, the value of C increases. This is done to enhance the learning model to make it more generalizable.

For the very small value problem of Eq. (15), the approach taken is also to utilize the Lagrangian dyadic form, with the difference being that the constraints are a little bit different from linear differentiability. Namely:

$$\max Q(\alpha)=\sum_{i=1}^{n}\alpha_i-1/2\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i * x_j\right) \tag{16}$$

The constraints are:

$$\begin{gathered}0\leq\alpha_i\leq C,i=1,2,\ldots,n\\ \sum_{i=1}^{n}\alpha_i y_i=0\end{gathered} \tag{17}$$

The formula, α_i has three possible values (1) $\alpha_i=0$. (2) $0<\alpha_i<C$. (3) $\alpha_i=C$.

x_i satisfying (2) is the standard support vector and x_i satisfying (3) is called the boundary support vector. From the above analysis, it is clear that only the support vectors play a role in determining the optimal hyperplane and the decision function. Meanwhile, we can see from equation (14) optimal

classification function that there is a dot product between the classification samples and the support vectors. In the case of binary classification, it can be well understood; however, if it is on the high-dimensional space, an inner product should be sought to replace the dot product, so that the original feature space can be transformed to a new feature space, and here $K(x,x')$ is used to replace the dot product in the optimal classification surface. The optimization function at this point then becomes:

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-1/2\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(x_i * x_j\right) \tag{18}$$

and the corresponding discriminant function Eq. (14) should also become:

$$f(x)=\operatorname{sgn}\left(\omega^* x+b^*\right)=\operatorname{sgn}\left\{\sum_{i=1}^{n}\alpha_i^* y_i K\left(x_i * x\right)+b^*\right\} \tag{19}$$

Therefore, the idea of SVM can be summarized as the use of a suitable inner product function to deal with nonlinear problems, which aims to map the input space to high dimensions to find the optimal classification surface.

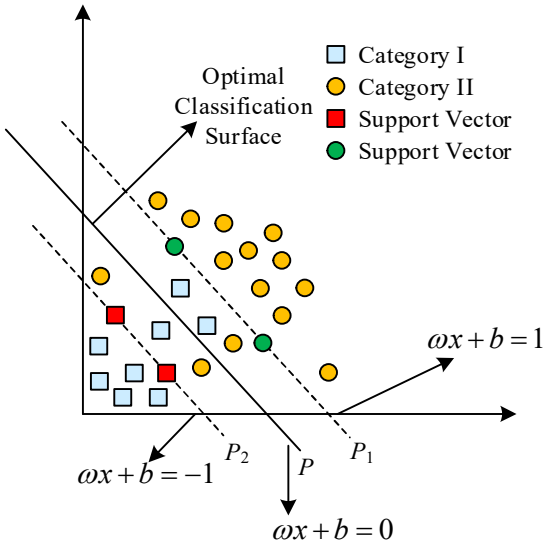


Fig. 3. The optimal hyperplane of linear inseparably

2.2.3. Kernel function. The reason why SVM can improve the handling of nonlinear problems and enhance the generalization ability of the learning model, these are because of the introduction of the kernel function, the kernel function allows the mapping of low-dimensional indivisible problems to high-dimensional linearly divisible, and the whole mapping process is shown in Figure 4.

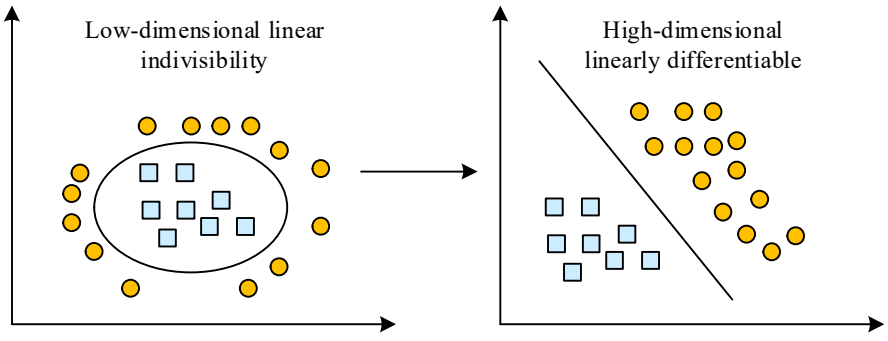


Fig. 4. Low dimensional linear inseparably mapping to high dimensional separable

In the nonlinear mapping process, a suitable $\varphi(\cdot)$-transformation function is introduced, and the set of training samples is set to be:

$$\Omega=(x_i,y_i),i=1,\ldots,l,x\in R^d,y\in\{-1,1\} \tag{20}$$

Transformation from R^d-space to Hilbert space H $x=\varphi(x)$:

$$R^d\to H,x\to x=\varphi(x) \tag{21}$$

by Eq. (20) makes the set of training samples Ω to be:

$$\Omega_\varphi=(\varphi(x_i),y_i),i=1,\ldots,l,x\in H,y\in\{-1,1\} \tag{22}$$

In terms of Eq. (22), it can be seen that the transformation $\varphi(\cdot)$ is expressed in the form of an inner product of functions:

$$K(x,x')=(\varphi(x)\cdot\varphi(x')) \tag{23}$$

In the actual classification process, we are concerned with the realization of the classification results rather than the data representation from the low-dimensional space mapped to the high-dimensional space, the kernel function of the advantage is that you can even do not need to know the specific functional form of $\varphi(\cdot)$, as long as there is a kernel function $K(\cdot,\cdot)$ to satisfy Mercer's conditions, which corresponds to a certain mapping space of the dot product.

3. Data collection and processing

3.1. Data sources

1) Student academic data. This data contains students' course grades, course elective information, test score distribution, and homework completion. These data reflect students' performance and preference in specialized knowledge learning. For example, by analyzing a student's performance advantage in science and engineering courses, it is possible to initially determine his potential in technology development and other related innovative and entrepreneurial fields.
2) Student Behavior Data. This data covers students' records of various activities on campus, such as the type and frequency of participation in club activities, experience of participating in research projects or competitions, and library borrowing records. Students who participate in innovation and practice club activities frequently may be more interested and motivated in innovation and entrepreneurship practices.
3) Student psychological and career assessment data. In the process of collecting this data, students were tested and guided to vote on the types of support by using the common means of seven major types of psychological tests, including intellectual inclination, personality test, vocational interest, personality, ability, values and motivation, and developmental assessment test, respectively, and the Hollander's vocational interest test (SDS) among the vocational interest tests was finally selected by combining the scenarios in which the seven types of tests were applicable. This test produces results through its partial test scale, constructing a hexagon consisting of conventional, corporate, social, artistic, research, and realistic types, etc., which makes a direct correlation between occupations and interests, and thus speculates on the type of work for which the test taker is suited. Collecting students' psychological and vocational assessment data helps to gain a deeper understanding of students' intrinsic traits and vocational tendencies, and determining the types of students' vocational interests based on the SDS test results provides a reference for career planning and entrepreneurial method selection.

3.2. Data cleansing and standardization

3.2.1. Data cleansing. The purpose of data cleansing is to remove noisy data from the data, such as incorrect records, duplicate data, etc. There may be incorrectly entered grades in the student achievement data, which need to be cleaned by data verification and correction. For missing data, this paper uses interpolation to fill in. In the process of data cleaning, the validity of the data is checked, such as student grades must be greater than 0, etc. At the same time, the collected student information may contain height, weight and other related information, which are of less value when applied to the subsequent analysis, and will lead to a waste of data analysis resources, so this paper excludes this type of data from the process.

3.2.2. Data standardization. Since data from different sources may have different scales and value ranges, data standardization is required to facilitate subsequent data analysis and model training. The data standardization method used in this paper is the $z\text{-}score$ -normalization method, and its calculation formula is shown below:

$$x_{new} = \frac{x - \mu}{\sigma} \tag{24}$$

where x is the raw data, μ is the mean of the data, and σ is the standard deviation of the data.

3.3. Feature Selection and Extraction

3.3.1. Feature data preprocessing. Information gain can consistently and effectively describe the meaning of information in an information system, and the more balanced the data dispersion in an information system, the greater the corresponding information gain [38]. Therefore, information gain can be used to describe balanced information data. Let the discrete random variable be $S(A)$:

$$S(A) = -\sum_{i=1}^{A} p_i \tag{25}$$

Where, p_i is the probability of discrete random variable A in the supply chain terminal high dimensional data.

Let the outcome variable B be a binary variable, i.e., $B = 0,1$; the subset of primary dependent variables in the sample data is C and the subset of secondary ones is D, which can be expressed in the following equation:

$$C = \{c : A_c \text{ is related to } EB \mid b\} \tag{26}$$

$$D = \{1, 2, \cdots, i\} \tag{27}$$

Examining the relationship between dependent variable A_c and outcome variable B in an information system, the information gain of dependent variable A_c can be expressed in the following equation:

$$S(A_c) = -\sum_{i=1}^{A_c} p_i C \tag{28}$$

where $p_i = 1$, $c = 1, 2, \cdots, p$.

Meanwhile, the information gain obtained for the dependent variable with greater correlation with the outcome variable B can be expressed in equation (29):

$$S(A_c) \mid B = -\sum_{i=1}^{A_c} p_i B \tag{29}$$

Using the sample data estimator to calculate the correlation between dependent variable A_c and outcome variable B, we can analyze the proportionate size of the importance of discrete random variables in the sample data to do characteristic screening and obtain the final simulation model.

3.3.2. Feature selection for high-dimensional data:

1) Feature Sorting. After completing the preprocessing of high-dimensional data of career planning and innovation industry, the information gain is sorted based on the improved genetic algorithm [39], the feature data beyond the edge range is removed, and the information density of each group information feature is set in order to effectively regulate the distribution of the subsequent data groups in the feature screening.

Let the set of discrete random variables be $X=\{x_1,x_2,\cdots,x_m\}$ and the set of unclassified characteristic variables of the sample be $Y=\{y_1,y_2,\cdots,y_{n-1}\}$, where the Mth data in the set of discrete random variables X is characterized by $Y_M(1\le l\le n-1)$ and the N th data is characterized by $Y_N(1\le l\le n-1)$, where $N<M$. The information density of Y_M can be expressed by the following equation:

$$\rho_M=\frac{H(Y_M)}{\sum_{M=1}^{n-1}H(Y_M)},M=1,2,\cdots,n-1 \tag{30}$$

Where $H(Y_M)$ is the information entropy of the Mnd data feature, according to equation (30) can be obtained between the two data information feature data density difference value, then set the information feature data density difference value between Y_M and Y_N is J_{M-N}, can be expressed in the following equation:

$$J_{M-N}=\frac{H(Y_M)-H(Y_N)}{\sum_{M=1}^{n-1}H(Y_M)},N,M=1,2,\cdots,n-1 \tag{31}$$

As a result, the larger $H(Y_M)$ is, the larger the corresponding information density ρ_M is, the smaller the value J_{M-N} of the difference in data density of information features between Y_M and Y_N is, the more compact the high-dimensional data is in some regions, and furthermore, the more likely it is that Y_M and Y_N will be computed in the same data distribution group.

2) Feature coding. In the initial stage of feature screening of high-dimensional data of career planning and innovation industry, the improved genetic algorithm is used to encode the data distribution group that is sorted according to the grouping of data features. The size of the space occupied by the data distribution group is calculated based on the data coding obtained by the algorithm, and the initial race is selected based on the race size. At the same time, discrete random variables are used to generate initial races for data groups with the same lower limit as the population size, focusing on encoding high-dimensional data.

3) Algorithm Evaluation. Let the set of data within the race be $H_h=[h_1,h_2,\cdots,h_m]$, $h=1,2,\cdots,m$, then the flexibility function obtained by the improved genetic algorithm for the data within the race that has completed feature coding can be expressed by equation (32):

$$H_h=\frac{1}{N}\sum_{i=1}^{N}\alpha_i \tag{32}$$

where H_h is the hnd data individual of the data set within the race, and α_i is the classification criterion corresponding to a particular data in the data classification group.

Let the data set outside the race be $H_j=[h_1,h_2,\cdots,h_n]$, $j=1,2,\cdots,n$. The flexibility function obtained by the improved genetic algorithm in feature screening of high dimensional data can be expressed in equation (33):

$$F = \frac{1}{N}\sum_{i=1}^{N} H_i \beta \tag{33}$$

where H_i is the set of data for which feature coding has been completed outside of races, and β is a condition that restricts the distribution of data features between races.

4. Career planning and innovation and entrepreneurship model construction and optimization

4.1. Modeling

1) Determine input and output variables. The processed students' academic performance feature vector, behavioral feature vector and psychological assessment feature vector are composed into a comprehensive feature vector, which is used as the input variable of the support vector machine X, and the type of students' career planning (employment, further education, entrepreneurship, etc.) or the results of the assessment of innovation and entrepreneurship ability (high, medium, low) are used as the output variable Y.
2) Selection of kernel function and parameter setting. Selecting the appropriate kernel function according to the characteristics of the data is especially important for the career planning and innovation and entrepreneurship model based on the support vector machine algorithm. Since the student characteristic data presents a complex nonlinear structure, selecting the traditional linear kernel function will lead to a decline in the model performance, which will ultimately reduce the accuracy of the model output results, so this paper chooses to use the Gaussian kernel function, and the parameters of the Gaussian kernel function as well as the regularization parameter C, etc., for tuning.

4.2. Model optimization

1) Cross-validation. In this paper, a k- fold cross-validation method is used to evaluate the performance of the model and perform parameter tuning. The dataset is divided into k subsets of similar size, k - 1 subsets are used as the training set each time, and the remaining one subset is used as the test set, which is repeated k times to calculate the performance metrics such as average accuracy, recall, and F1 value. The parameter settings with the best performance are selected by cross-validation under different parameter combinations.
2) Model Updating and Improvement. With the continuous generation of new data, the support vector machine model is regularly updated. In this paper, an incremental learning approach is used to gradually incorporate new data into the model training, while the model is adjusted and optimized to adapt to changes in the characteristics of the student population and the educational environment. For example, when a new innovation and entrepreneurship education program is implemented, data from students' learning and practical activities in related courses can be used to update the model so that the model can better reflect the new educational effects and student development.

5. Educational paths based on support vector machine algorithms

5.1. Personalized career planning guidance

1) Student classification and labeling. Using the trained support vector machine model to continue the classification of students, according to the model output of the type of career planning labels, such as the students are divided into different categories such as suitable for employment type, further study and entrepreneurship type. For students who are suitable for employment, further analyze their suitable career fields, such as technology, management, service, etc.

2) Personalized Programming. Develop personalized career planning programs for different types of students. For entrepreneurial students, we provide entrepreneurial project recommendation, entrepreneurial resource matching, and entrepreneurial training course arrangement. For example, according to students' interests and ability characteristics, recommend matching entrepreneurial projects, such as science and technology entrepreneurial projects recommended for students with science and technology background and innovation ability. Provide students with resources of entrepreneurship mentors, whose area of specialization matches students' entrepreneurial direction. Arranging entrepreneurship training courses, including courses on business plan writing, marketing, financial management, etc. The depth and breadth of the course content is adjusted according to the students' foundation and learning ability.

5.2. Design of practical activities for innovation and entrepreneurship education

1) Model-based student team formation. In innovation and entrepreneurship practice activities, such as entrepreneurship competitions, scientific research project team formation, etc., reasonable team formation is carried out based on the results of the analysis of students' innovation and entrepreneurship abilities and specialties by the support vector machine model. Students with innovative thinking ability, students with technical expertise, students with organizational and management ability, etc. are combined together to form a team with complementary advantages and improve team competitiveness.
2) Activity content customization. Customize the content of practical activities of innovation and entrepreneurship education according to the characteristics and needs of the student groups, for example, for the student groups that show strong innovative ability in the model but insufficient practical experience, design more activities with practical operation links, such as enterprise field research, entrepreneurial simulation of the actual battle, and so on. For students with certain entrepreneurial foundation and experience, organize activities such as high-end entrepreneurship forums and international entrepreneurship exchange programs to broaden their horizons. Enhance their entrepreneurial level.

5.3. Evaluation of Educational Effectiveness and Feedback

1) Construction of assessment index system. Establish a multi-dimensional evaluation index system of educational effect, including the indicators of students' knowledge and skill enhancement, such as the degree of mastery of entrepreneurial knowledge and the ability to utilize innovative methods. Indicators of students' attitude and value change, such as the degree of entrepreneurial willingness and the effect of innovation spirit cultivation. Indicators of students' actual achievements, such as the number of entrepreneurial projects launched, the number of innovative achievements transformed, etc.
2) Model-based assessment and feedback. The support vector machine model is used to analyze the educational effect assessment data and predict the impact of educational interventions on the development of students' career planning and innovation and entrepreneurship abilities, even if problems and deficiencies in the educational process are found. For example, if the model predicts that a certain type of educational activity does not have a significant effect on the enhancement of innovation and entrepreneurship ability of a certain group of students, even if it adjusts the content and mode of the activity or increases the investment of educational resources, etc., and uses the feedback information to optimize the design of subsequent educational paths.

6. Experimental results and analysis

6.1. Feature Selection and Extraction

In this section, in order to validate the effectiveness of the proposed algorithms, comparison experiments of the algorithms will be carried out on 12 sub-datasets from the collected student characteristics data. Nine other algorithms are selected for comparison, which are SGAI algorithm, MIMIC-FS algorithm, SRS algorithm, RS algorithm, TIE algorithm, IAMB algorithm, PCMB algorithm, HITON-MB algorithm and MMMB algorithm. During the experiments, these algorithms are done with the assistance of 3 different classifiers respectively, the 3 classifiers are KNN, NB and SVM.The hardware equipment used in the experiments are Core i7 10510U 3.5 GHz processor, 16 GB DDR 4 2666 RAM, and Windows 10 operating system are performed on HP workstation.

6.1.1. Experimental data. A total of 12 benchmark sub-datasets were selected for the experiment, and the details of each dataset are shown in Table 1. In the table, four sub-datasets, including G1, G2, G3, and G4, were derived from students' academic performance characteristics, including final grades, usual grades, practical grades, and course choices, while five sub-datasets, including B1-B5, were the types of clubs participated, the frequency of club participation, the number of competitions, the number of scientific research projects, and the number of library borrowings, and the three sub-datasets of H1-H3 were personality test scores, occupational interest test scores, and mental health test scores from students' psychological assessment characteristics, respectively. These 12 sub-datasets cover all the data related to students' career planning and innovation and entrepreneurship, and the dimensions of the datasets range from 20 to 10394, the sample sizes range from 52 to 4572, and the categories of the datasets range from 1 category to 9 categories. For each dataset in the table, the limitations and peculiarities of fixed division of datasets can be avoided during model evaluation by conducting 50 experiments on each training dataset individually and finally taking the average classification accuracy. The optimal feature subset is selected by different feature selection methods, and then the selected optimal feature subset is applied to the test data and the classification accuracy of the selected feature subset is evaluated using a classifier.

Table 1. Experimental data set

Datasets	Instances	Features	Class	Datasets	Instances	Features	Class
G1	262	20	3	B3	547	27	1
G2	374	33	7	B4	238	45	4
G3	108	52	6	B5	4572	58	8
G4	52	201	2	H1	395	439	3
B1	2076	498	4	H2	163	647	9
B2	619	6397	5	H3	109	10394	7

6.1.2. Experimental parameterization. In the course of the experiments, three classifiers are used to evaluate the feature subset, namely, the NB classifier and KNN classifier provided by Matlab Statistical Toolbox, and the SVM classifier provided by LibSVM library. Among the nine comparative algorithms of the experiment, MIMIC-FS algorithm, SRS algorithm, SGAI algorithm, RS algorithm, and TIE algorithm are feature selection algorithms based on common correlation coefficients; IAMB algorithm, PCMB algorithm, HITON-MB algorithm, and MMMB algorithm are feature selection algorithms based on the state-of-the-art Markov Blanket Filter. During the experiments, the parameters of the algorithms are designed as follows:(1) In the SRS algorithm, parameters λ_1 and λ_2 in the algorithm are varied over the interval [0,0.1] in steps of 0.05. (2) The significance level of IAMB algorithm, PCMB algorithm, HITON-MB algorithm and MMMB algorithm are all set to 0.05. (3) Parameter selection for the improved genetic algorithm: maximum number of iterations $T_{max}=500$, initial population size

$N=15$. For the inertia weight ω, the linear decreasing method is usually used to adjust ω. However, the process of population genetics is not a linear process. In this paper, a nonlinear method is used to adaptively adjust the inertia weights ω to increase the diversity, i.e:

$$\omega=\omega_{min}+(\omega_{max}-\omega_{min})\cdot(-\frac{1}{1+(1+\frac{t}{T_{max}})}) \tag{34}$$

6.1.3. Analysis of experimental results. In order to validate the feature selection and extraction performance of the Improved Genetic Algorithm (IGA) in this paper, experiments are conducted on 12 different datasets and compared with 9 algorithms. The results of the experiments are analyzed in terms of both the classification accuracy of the experiments and the size of the selected feature subset, where the classification accuracy of the experiments and the number of selected features are averaged after 50 independent experiments, respectively.

1) Comparison of classification accuracy. Tables 2 - 4 show the results of comparison of IGA algorithm with MIMIC-FS algorithm, SRS algorithm, SGAI algorithm, RS algorithm, TIE algorithm, IAMB algorithm, PCMB algorithm, HITON-MB algorithm and MMMB algorithm on KNN, NB and SVM classifiers respectively. As can be seen from the table, the experimental dataset contains both binary and multiclassification, for the KNN classifier, the experimental results on 12 datasets show that the IGA algorithm consistently outperforms the other 9 compared feature selection algorithms on 8 datasets (G1, G3, G4, B1, B3, B5, H1, H3), and the average of the experimental results of this algorithm on 12 datasets is 0.863 , which is 0.050, 0.078, 0.070, 0.087, 0.073, 0.050, 0.062, 0.042, 0.077 better than the MIMIC-FS algorithm, the SRS algorithm, the SGAI algorithm, the RS algorithm, the TIE algorithm, the IAMB algorithm, the PCMB algorithm, the Hiton-MB algorithm and the MMMB algorithm, respectively. On the NB classifier, the IGA algorithm outperforms the other nine compared algorithms on eight datasets (G1, G3, G4, B2, B3, B4, H2, and H3), and on the remaining four datasets (G1, B2, B5, and H1). The IGA algorithm is ranked first with an average classification accuracy of 0.872 on the 12 datasets, i.e., the classification performance is optimal compared to the other nine algorithms. On the SVM classifier, each feature selection achieves better performance than on the other two classifiers, and the classification accuracy of the IGA algorithm in this paper is significantly better than that of the other nine comparative algorithms on 10 datasets (G1, G2, G3, G4, B1, B3, B4, H1, H2, and H3), with an average classification accuracy of 0.943 on 12 datasets, which is comparable with the other two classifiers the mean values of 0.080 and 0.071 respectively. It shows that the improved genetic algorithm and SVM model selected in this paper are more suitable, and applying it to the feature selection and extraction of the model constructed based on the support vector machine algorithm in this paper can make the model achieve better results.

Table 2. The classification accuracy of the feature selection algorithm(KNN)

Datasets	IGA		MIMIC-FS		SRS		SGAI		RS	
	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.
G1	0.810	0.017	0.749	0.022	0.724	0.085	0.766	0.027	0.724	0.046
G2	0.961	0.011	0.939	0.019	0.924	0.027	0.945	0.025	0.934	0.022
G3	0.819	0.072	0.771	0.034	0.709	0.031	0.715	0.034	0.693	0.011
G4	0.832	0.014	0.762	0.034	0.799	0.087	0.829	0.105	0.777	0.060
B1	0.938	0.024	0.858	0.018	0.878	0.028	0.915	0.027	0.841	0.051
B2	0.862	0.018	0.823	0.103	0.833	0.046	0.825	0.026	0.825	0.030
B3	0.867	0.022	0.758	0.019	0.702	0.064	0.689	0.022	0.730	0.042
B4	0.611	0.018	0.617	0.088	0.568	0.041	0.541	0.088	0.556	0.097
B5	0.944	0.025	0.861	0.088	0.818	0.062	0.756	0.022	0.781	0.064
H1	0.885	0.067	0.865	0.061	0.852	0.091	0.831	0.020	0.865	0.079
H2	0.965	0.019	0.929	0.021	0.941	0.069	0.955	0.023	0.944	0.038
H3	0.860	0.032	0.818	0.017	0.675	0.024	0.753	0.064	0.641	0.073
Mean	0.863	/	0.813	/	0.785	/	0.793	/	0.776	/
Rank	1		3		9		6		10	
Datasets	TIE		IAMB		PCMB		HITON-MB		MMMB	
	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.
G1	0.766	0.089	0.697	0.015	0.697	0.081	0.708	0.091	0.697	0.059
G2	0.938	0.062	0.962	0.031	0.940	0.030	0.969	0.034	0.934	0.050
G3	0.769	0.029	0.812	0.054	0.742	0.092	0.802	0.076	0.791	0.019
G4	0.814	0.047	0.817	0.014	0.792	0.040	0.769	0.076	0.765	0.038
B1	0.786	0.026	0.808	0.016	0.863	0.026	0.911	0.042	0.911	0.023
B2	0.847	0.033	0.857	0.017	0.872	0.031	0.872	0.059	0.833	0.079
B3	0.702	0.029	0.786	0.013	0.730	0.064	0.758	0.015	0.675	0.036
B4	0.585	0.084	0.568	0.021	0.545	0.051	0.556	0.013	0.556	0.056
B5	0.728	0.032	0.850	0.021	0.879	0.040	0.848	0.056	0.838	0.023
H1	0.839	0.043	0.865	0.031	0.800	0.083	0.852	0.078	0.865	0.062
H2	0.955	0.025	0.926	0.016	0.976	0.061	0.976	0.092	0.931	0.032
H3	0.750	0.083	0.808	0.031	0.780	0.058	0.830	0.020	0.637	0.067
Mean	0.790	/	0.813	/	0.801	/	0.821	/	0.786	/
Rank	7		3		5		2		8	

Table 3. The classification accuracy of the feature selection algorithm(NB)

Datasets	IGA		MIMIC-FS		SRS		SGAI		RS	
	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.
G1	0.831	0.023	0.777	0.147	0.697	0.105	0.716	0.042	0.697	0.077
G2	0.966	0.018	0.905	0.057	0.932	0.054	0.952	0.098	0.941	0.095
G3	0.929	0.090	0.753	0.101	0.791	0.100	0.910	0.089	0.780	0.090
G4	0.941	0.022	0.754	0.093	0.798	0.164	0.847	0.099	0.757	0.090
B1	0.933	0.028	0.725	0.104	0.974	0.055	0.974	0.021	0.911	0.036
B2	0.861	0.018	0.773	0.080	0.787	0.037	0.777	0.099	0.780	0.040
B3	0.911	0.012	0.758	0.039	0.814	0.018	0.680	0.105	0.831	0.096
B4	0.751	0.024	0.653	0.050	0.599	0.078	0.580	0.029	0.660	0.041
B5	0.648	0.017	0.558	0.104	0.638	0.016	0.649	0.024	0.638	0.105
H1	0.892	0.012	0.919	0.093	0.839	0.054	0.805	0.105	0.883	0.064
H2	0.955	0.018	0.938	0.103	0.899	0.016	0.919	0.093	0.919	0.029
H3	0.849	0.012	0.694	0.079	0.678	0.049	0.698	0.044	0.648	0.054
Mean	0.872	/	0.767	/	0.787	/	0.792	/	0.787	/
Rank	1		9		5		4		5	
Datasets	TIE		IAMB		PCMB		HITON-MB		MMMB	
	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.
G1	0.697	0.064	0.697	0.014	0.697	0.040	0.708	0.052	0.697	0.057
G2	0.971	0.054	0.962	0.015	0.940	0.093	0.969	0.047	0.934	0.035
G3	0.802	0.126	0.788	0.039	0.850	0.087	0.798	0.048	0.841	0.076
G4	0.810	0.059	0.817	0.016	0.792	0.058	0.769	0.036	0.765	0.026
B1	0.959	0.209	0.808	0.029	0.593	0.014	0.911	0.054	0.911	0.070
B2	0.787	0.058	0.812	0.016	0.852	0.061	0.842	0.055	0.834	0.025
B3	0.901	0.017	0.730	0.022	0.823	0.027	0.860	0.015	0.808	0.041
B4	0.704	0.029	0.530	0.015	0.554	0.035	0.599	0.070	0.559	0.064
B5	0.629	0.066	0.634	0.057	0.577	0.050	0.638	0.066	0.638	0.078
H1	0.826	0.054	0.805	0.017	0.839	0.029	0.865	0.042	0.826	0.015
H2	0.941	0.035	0.944	0.015	0.908	0.016	0.911	0.015	0.944	0.016
H3	0.838	0.105	0.698	0.038	0.728	0.026	0.838	0.013	0.668	0.015
Mean	0.822	/	0.769	/	0.763	/	0.809	/	0.785	/
Rank	2		8		10		3		7	

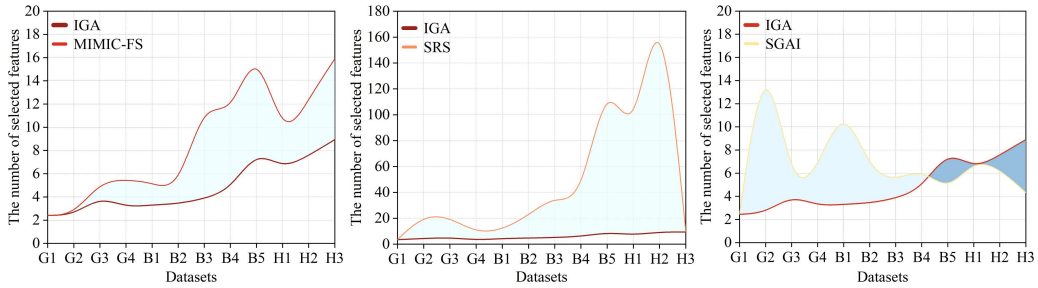
Table 4. The classification accuracy of the feature selection algorithm(SVM)

Datasets	IGA		MIMIC-FS		SRS		SGAI		RS	
	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.
G1	0.916	0.111	0.856	0.102	0.795	0.159	0.795	0.106	0.795	0.110
G2	0.993	0.128	0.916	0.113	0.976	0.104	0.958	0.365	0.929	0.109
G3	0.998	0.105	0.855	0.165	0.807	0.318	0.937	0.106	0.928	0.108
G4	0.999	0.120	0.886	0.111	0.915	0.152	0.971	0.128	0.893	0.104
B1	0.947	0.105	0.938	0.141	0.755	0.301	0.848	0.129	0.839	0.148
B2	0.865	0.101	0.926	0.141	0.865	0.101	0.943	0.112	0.856	0.148
B3	0.999	0.105	0.836	0.108	0.845	0.106	0.704	0.167	0.762	0.140
B4	0.791	0.105	0.751	0.146	0.777	0.104	0.725	0.116	0.747	0.133
B5	0.888	0.119	0.705	0.105	0.919	0.129	0.907	0.124	0.932	0.138
H1	0.973	0.134	0.907	0.115	0.848	0.119	0.809	0.135	0.809	0.109
H2	0.999	0.107	0.923	0.189	0.934	0.174	0.973	0.178	0.954	0.139
H3	0.950	0.101	0.903	0.130	0.905	0.128	0.905	0.182	0.895	0.155
Mean	0.943	/	0.869	/	0.862	/	0.873	/	0.862	/
Rank	1		3		5		2		5	
Datasets	TIE		IAMB		PCMB		HITON-MB		MMMB	
	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.	ACC	Std.
G1	0.795	0.105	0.795	0.115	0.795	0.118	0.788	0.103	0.795	0.142
G2	0.876	0.123	0.901	0.11	0.882	0.132	0.872	0.119	0.849	0.116
G3	0.937	0.121	0.91	0.106	0.878	0.149	0.873	0.167	0.818	0.139
G4	0.889	0.152	0.885	0.104	0.889	0.119	0.889	0.098	0.886	0.151
B1	0.847	0.114	0.877	0.116	0.701	0.149	0.839	0.179	0.839	0.156
B2	0.943	0.112	0.874	0.103	0.865	0.159	0.874	0.161	0.858	0.131
B3	0.991	0.157	0.817	0.114	0.845	0.116	0.845	0.139	0.762	0.128
B4	0.781	0.185	0.704	0.112	0.731	0.106	0.704	0.147	0.747	0.184
B5	0.716	0.136	0.719	0.103	0.751	0.122	0.933	0.116	0.933	0.098
H1	0.809	0.164	0.952	0.105	0.835	0.124	0.822	0.136	0.848	0.175
H2	0.893	0.12	0.857	0.104	0.837	0.133	0.894	0.182	0.863	0.177
H3	0.905	0.153	0.635	0.105	0.905	0.141	0.895	0.144	0.885	0.167
Mean	0.865	/	0.827	/	0.826	/	0.852	/	0.840	/
Rank	4		9		10		7		8	

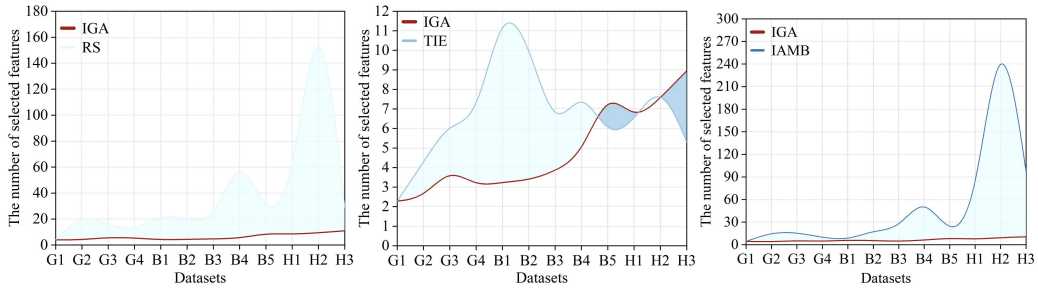
2) Comparison of the size of selected feature subsets. The average number of features selected by the IGA algorithm is shown in Table 5 and Figure 5 shows the comparison of the number of features selected by the IGA algorithm with the MIMIC-FS algorithm, SRS algorithm, SGAI algorithm, Rs algorithm, TIE* algorithm, IAMB algorithm, PCMB algorithm, HITON-MB algorithm and MMMB algorithm, respectively. On most of the datasets, the number of features selected by the IGA algorithm is less than the other nine algorithms. Although on some datasets, the number of features selected by IGA algorithm is greater than the other algorithms.IGA maximizes the retention of the more salient features associated with the class. In terms of redundancy and separation performance of features, the IGA algorithm can remove redundant features more effectively, select features with stronger separation performance, further reduce the shrinkage space of features, and achieve the highest classification performance with fewer features.

Table 5. The number of features selected by the IGA algorithm

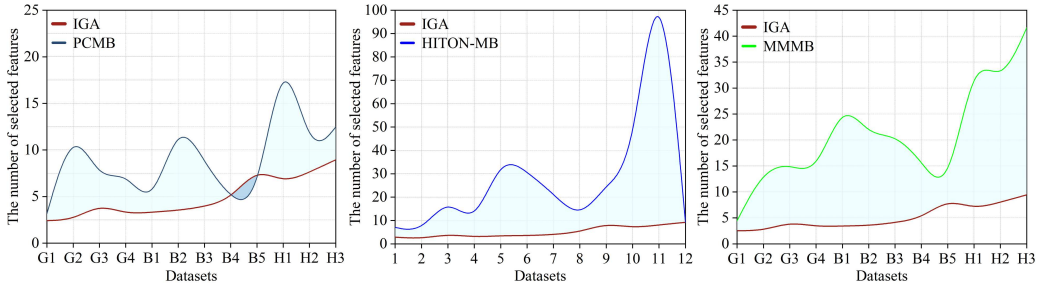
Datasets	# of selected features	Datasets	# of selected features
G1	2.72	B3	2.18
G2	3.87	B4	2.36
G3	2.93	B5	3.11
G4	3.54	H1	4.28
B1	7.17	H2	6.19
B2	7.29	H3	8.47



(a) IGA&MIMIC-FS (b) IGA&SRS (c) IGA&SGAI



(d) IGA&RS (e) IGA&TIE (f) IGA&IAMB



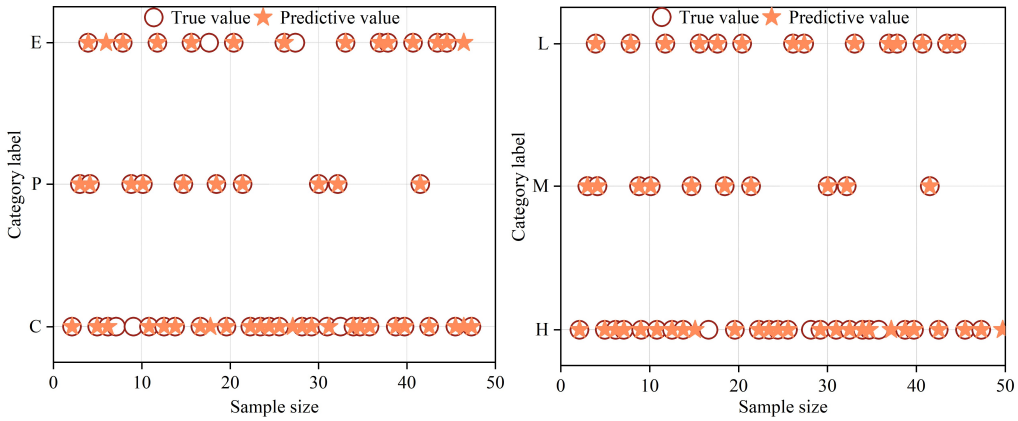
(g) IGA&PCMB (h) IGA&HITON-MB (i) IGA&MMMB

Fig. 5. IGA algorithm compares number of features with nine algorithms

6.2. Performance Analysis of Career Planning and Innovation and Entrepreneurship Models

6.2.1. Model classification forecast analysis. The division ratio of training set and test set in the student career planning and innovation and entrepreneurship dataset is 8:2, and the prediction test analysis of the student career planning and innovation and entrepreneurship education model based on SVM algorithm is carried out in the same hardware environment as that of the feature selection experiments, and the classification results of the test are shown in Fig. 6, and the category labels C, P, and E denote the employment, further education, and entrepreneurship. Figure 6(b) shows the results

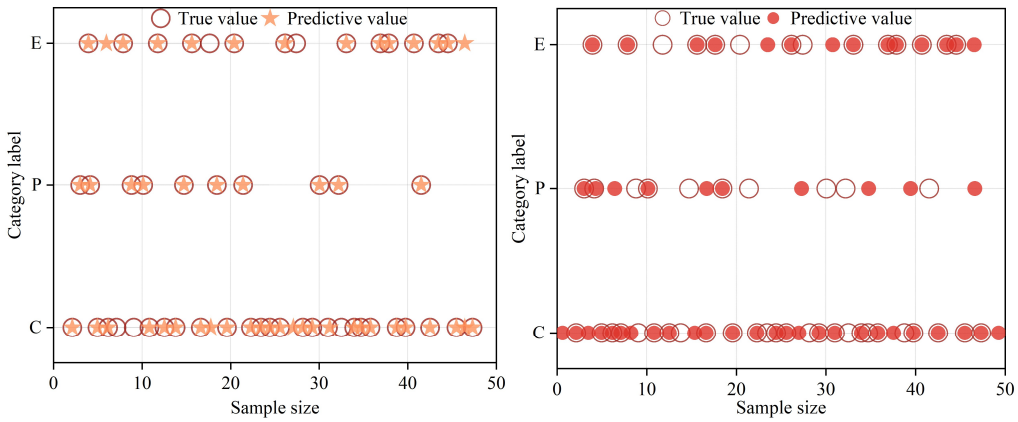
of the assessment of students' innovation and entrepreneurship ability, and the category labels H, M, and L represent students' high, medium, and low innovation and entrepreneurship ability, respectively. From the figure, it can be seen that the model achieves 100% categorization prediction accuracy for the category of promotion (P) in career planning as well as the categories of medium (M) and low (L) in the assessment of innovation and entrepreneurship ability. Out of the total 50 sets of test classification samples, 45 sets of samples were correctly classified for the career planning test, with a model classification prediction accuracy of 90%. There are 47 groups of correctly categorized samples for innovation and entrepreneurship ability assessment, and the classification prediction accuracy is 94%. Overall, the classification prediction accuracy of students' career planning and innovation and entrepreneurship education model constructed based on SVM algorithm in this paper can reach 90% and above in both career planning and innovation and entrepreneurship ability assessment using students' data, and it shows good performance in the experiment.



(a) Career planning (b) Innovative entrepreneurial ability

Fig. 6. Model classification prediction results

In order to verify the superiority of this paper's model, this paper's model is experimentally compared with three classification prediction models, namely, DBO-SVM, SSA-SVM, and Random Forest (RF), and the results are shown in Figs. 7 and 8, where (a)-(d) represent the prediction results of this paper's model, DBO-SVM, SSA-SVM, and Random Forest model, respectively. It is obvious from the figure that the classification prediction effect of this paper's model is better than the other three classification prediction models, both in terms of students' career planning and in terms of students' innovation and entrepreneurship ability assessment. It shows that the model constructed based on SVM algorithm in this paper is more suitable for students' career planning and innovation and entrepreneurship education scenarios.



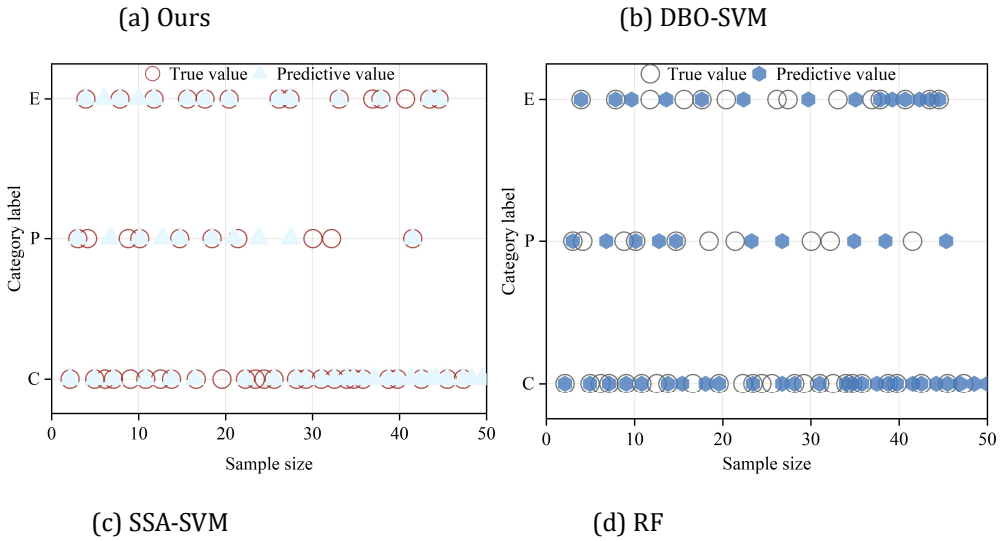


Fig. 7. The classification prediction effect of different models in career planning

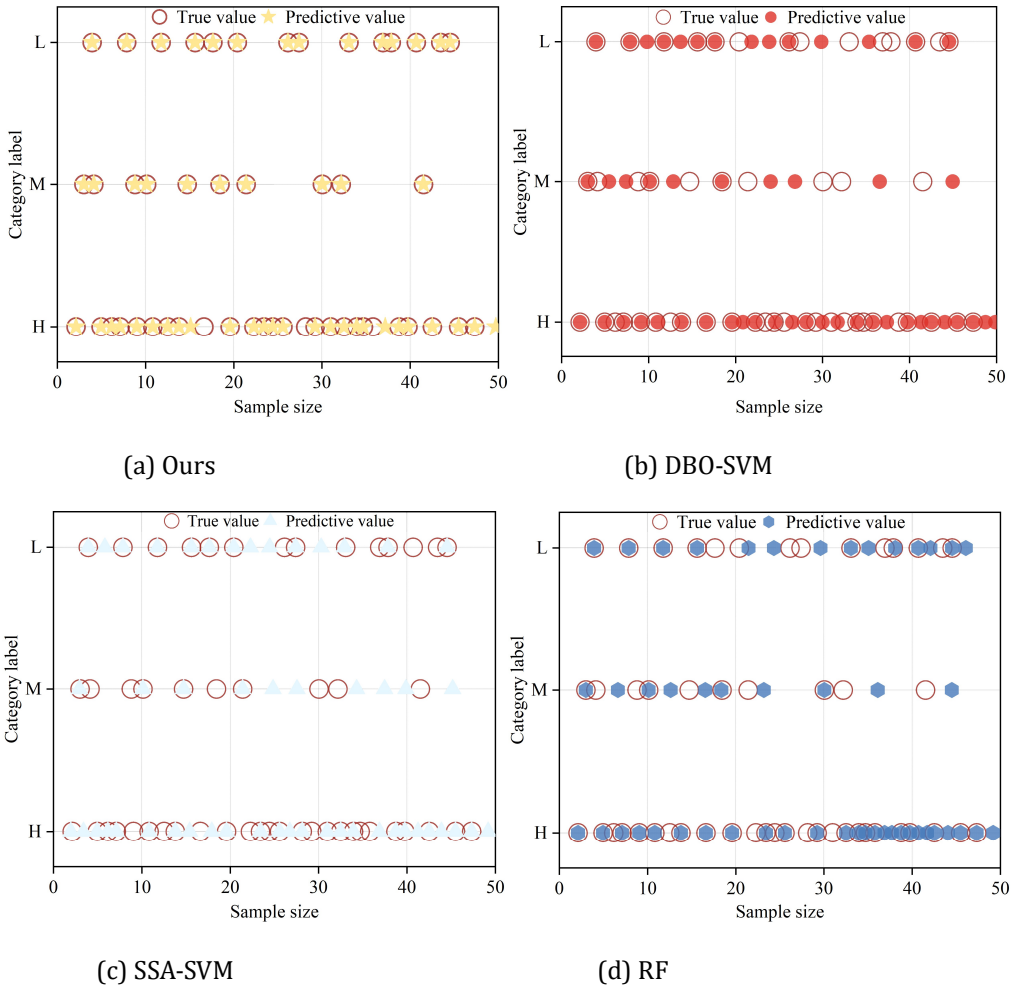


Fig. 8. Classification prediction effect in innovative entrepreneurial ability

6.2.2. Model performance analysis. Using accuracy (A), precision (P), recall (R) and F1 score as the evaluation indexes of model performance, the model in this paper is compared with DBO-SVM, SSA-SVM and random forest model in the comparison experiments. The results are shown in Table 6 and Table 7. By analyzing the results in the tables, it can be seen that this paper's model has the highest

accuracy rate in career planning and innovation and entrepreneurship ability assessment, which are 90% and 94% respectively, and it improves 16% and 32% compared to the SSA-SVM model with the next best performance. Meanwhile, the values of precision rate, recall rate and F1 score of this paper's model on the two types of tasks are all 1, indicating that this paper's model is able to maintain optimal model performance stability while ensuring high precision rate. Compared with the other three predictive classification models, the performance of this paper's model is the most advanced in career planning and honest entrepreneurial ability assessment.

Table 6. Model performance contrast(Career planning)

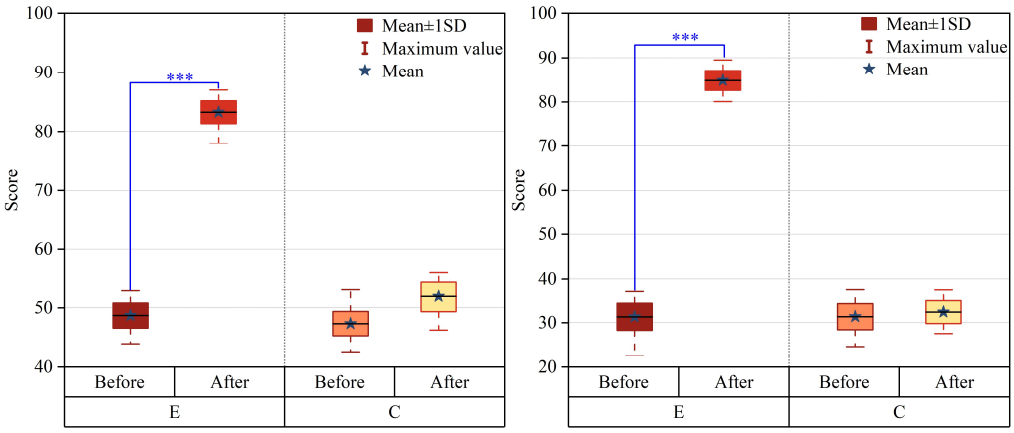
Model	Categories	P	R	F1	A/%
Ours	C	1.00	1.00	1.00	90.00
	P	0.98	0.92	0.96	
	E	0.93	0.97	0.96	
	Mean	0.97	0.96	0.97	
DBO-SVM	C	0.75	0.77	0.73	58.00
	P	0.62	0.58	0.61	
	E	0.67	0.63	0.62	
	Mean	0.68	0.66	0.65	
SSA-SVM	C	0.78	0.74	0.76	74.00
	P	0.67	0.69	0.66	
	E	0.79	0.76	0.77	
	Mean	0.75	0.73	0.73	
RF	C	0.57	0.52	0.53	54.00
	P	0.55	0.51	0.54	
	E	0.56	0.52	0.55	
	Mean	0.56	0.52	0.54	

Table 7. Model performance contrast(Innovation and entrepreneurship assessment)

Model	Categories	P	R	F1	A/%
Ours	H	1.00	1.00	1.00	94.00
	M	0.99	0.98	0.98	
	L	0.96	0.94	0.95	
	Mean	0.98	0.97	0.98	
DBO-SVM	H	0.71	0.64	0.67	56.00
	M	0.54	0.58	0.56	
	L	0.63	0.61	0.61	
	Mean	0.63	0.61	0.61	
SSA-SVM	H	0.73	0.76	0.74	62.00
	M	0.53	0.57	0.52	
	L	0.66	0.64	0.62	
	Mean	0.64	0.66	0.63	
RF	H	0.73	0.70	0.71	60.00
	M	0.51	0.53	0.52	
	L	0.61	0.63	0.61	
	Mean	0.62	0.62	0.61	

6.3. Analysis of the effectiveness of educational pathways

Based on the support vector machine algorithm, this paper constructs an educational model of students' career planning and innovation and entrepreneurship ability, and proposes an educational path of career planning and innovation and entrepreneurship ability combined with the support vector machine algorithm. In order to fully verify the effectiveness of the educational path of this paper, 100 students were randomly selected in a university for a one-year educational experiment. According to the ratio of 1:1, the students were divided into experimental group and control group, and the experimental group used the educational path proposed in this paper to cultivate the students, while the control group used the traditional way of cultivating career planning and innovative entrepreneurial ability. Students' satisfaction scores on career planning and assessment scores on innovation and entrepreneurship abilities were measured before and after the educational experiment. Among them, the satisfaction score was scored by students independently through the form of questionnaire, and the innovation and entrepreneurship ability was obtained from the model of this paper. According to the assessment category of innovation and entrepreneurship ability, the low innovation and entrepreneurship ability is classified to the score interval of [0,33], the medium ability to the interval of (33,67], and the high ability to the interval of (67,100], so that both the satisfaction score and the assessment score of innovation and entrepreneurship ability can be expressed in percentage. The results obtained from the experiment are shown in Fig. 9, with (a) and (b) denoting career planning and innovation and entrepreneurship ability assessment, respectively, and E and C denoting the experimental and control groups, respectively. It is obvious from the results that adopting the educational path proposed in this paper can significantly improve students' satisfaction with career planning and innovative entrepreneurial ability, in which the mean values of career planning satisfaction and innovative entrepreneurial ability scores of the experimental group students before the experiment are 48.71 and 31.33, respectively, and after the experiment are 83.24 and 84.82, respectively, which is an improvement of 70.89% and 170.73%. After the experiment, students' innovation and entrepreneurship ability increased from low grade to high grade, which fully proved the effectiveness of the educational path proposed in this paper.



(a) Career planning (b) Innovation and entrepreneurship

Fig. 9. Education path experiment results

7. Conclusion

This paper constructs the scientific education path of students' career planning and innovation and entrepreneurship based on the branch vector machine algorithm, and analyzes the effectiveness of this paper's education path with the evaluation indexes of career planning satisfaction and innovation and entrepreneurship ability score. The results show that the average values of the experimental group's satisfaction with career planning and students' innovation and entrepreneurship ability before using this paper's educational path are 48.71 and 31.33, respectively, and after using this paper's educational path, the average values of the scores are increased by 70.89% and 170.73%, respectively, compared with those before the experiment. This result fully demonstrates that the educational path proposed in this paper has a significant positive effect on the rational planning of students' career and the enhancement of innovation and entrepreneurship ability. On the contrary, students in the control group, who used the traditional career planning and innovation and entrepreneurship training methods, did not have a significant increase in career planning satisfaction and innovation and entrepreneurship scores before and after the educational experiment, which further highlights that the educational path constructed in this paper is effective.

References

[1] Jackson, D., & Tomlinson, M. (2020). Investigating the relationship between career planning, proactivity and employability perceptions among higher education students in uncertain labour market conditions. Higher education, 80(3), 435-455.

[2] Shabur, M. A. (2024). The potential and implications of artificial intelligence in Bangladesh's early career planning education. Discover Global Society, 2(1), 50.

[3] Rogers, M. E., & Creed, P. A. (2011). A longitudinal examination of adolescent career planning and exploration using a social cognitive career theory framework. Journal of adolescence, 34(1), 163-172.

[4] Stebleton, M. J., Kaler, L. S., Diamond, K. K., & Lee, C. (2020). Examining career readiness in a liberal arts undergraduate career planning course. Journal of employment counseling, 57(1), 14-26.

[5] Wei, L. Z., Zhou, S. S., Hu, S., Zhou, Z., & Chen, J. (2021). Influences of nursing students' career planning, internship experience, and other factors on professional identity. Nurse education today, 99, 104781.

[6] Shury, J., Vivian, D., Turner, C., & Downing, C. (2017). Planning for success: Graduates' career planning and its effect on graduate outcomes. London: Department for Education.

[7] Litynska, V., Kondratska, L., Romanovska, L., Kravchyna, T., Chagovets, A., & Kalaur, S. (2023). Career planning of scientific-pedagogical personnel in higher education institutions. European Journal of Sustainable Development, 12(4), 437-437.

[8] Yang, L., & Wong, L. P. (2020). Career and life planning education: Extending the self-concept theory and its multidimensional model to assess career-related self-concept of students with diverse abilities. ECNU Review of Education, 3(4), 659-677.

[9] Wei, X., Liu, X., & Sha, J. (2019). How does the entrepreneurship education influence the students' innovation? Testing on the multiple mediation model. Frontiers in psychology, 10, 1557.

[10] Papadopoulos, P. M., Burger, R., & Faria, A. (Eds.). (2016). Innovation and entrepreneurship in education. Emerald Group Publishing.

[11] Johansen, V. (2018). Innovation cluster for entrepreneurship education. Lillehammer, Norway, Østlandsforskning/Eastern Norway Research Institute. Available online: http://icee-eu.eu/component/attachments.

[12] Feng, X., Wang, X., & Qi, M. (2024). A comparison study on innovation and entrepreneurship education in and out of China from a bibliometric perspective. Library Hi Tech, 42(6), 1810-1838.

[13] Čapienė, A., & Ragauskaitė, A. (2017). Entrepreneurship education at university: Innovative models and current trends. Research for rural development, 2(23), 284-291.

[14] Mavlutova, I., Lesinskis, K., Liogys, M., & Hermanis, J. (2020). Innovative teaching techniques for entrepreneurship education in the era of digitalisation. WSEAS Transactions on Environment and Development, 16(1), 725-733.

[15] Gao, H., Qiu, Z., Liu, Z., Huang, L., & San, Y. (2016). Research and practice on college students' innovation and entrepreneurship education. In Social Computing: Second International Conference of Young Computer Scientists, Engineers and Educators, ICYCSEE 2016, Harbin, China, August 20-22, 2016, Proceedings, Part II 2 (pp. 36-44). Springer Singapore.

[16] Xiaoxing, Q. (2020). Research on innovation and entrepreneurship education. Higher Education Research, 10(4), 209-213.

[17] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile networks and applications, 19, 171-209.

[18] Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), 2053951716631130.

[19] Wu, W. (2022). Probabilistic linguistic promethee i and ii methods for evaluation of the reform scheme of postgraduate innovation and entrepreneurship education talent training mode under the big data environment. Mathematical Problems in Engineering, 2022(1), 8341052.

[20] Zhao, X., & Cheng, L. (2022, December). Construction of Big Data Analysis Model for High School Students' Career Planning Interest Events Based on Weight Algorithm. In 2022 International Conference on Computer Science, Information Engineering and Digital Economy (CSIEDE 2022) (pp. 526-532). Atlantis Press.

[21] Zhang, W. (2022). Quality improvement of college students' innovation and entrepreneurship education based on big data analysis under the background of cloud computing. Scientific Programming, 2022(1), 8734474.

[22] Quanli, W. (2020, December). Survey of College Students' Career Planning Based on Big Data Statistical Analysis. In Proceedings of the 2020 3rd International Conference on E-Business, Information Management and Computer Science (pp. 363-369).

[23] Nie, M., Xiong, Z., Zhong, R., Deng, W., & Yang, G. (2020). Career choice prediction based on campus big

data—mining the potential behavior of college students. Applied Sciences, 10(8), 2841.

[24] Wang, Y., Yang, L., Wu, J., Song, Z., & Shi, L. (2022). Mining campus big data: Prediction of career choice using interpretable machine learning method. Mathematics, 10(8), 1289.

[25] Zhang, N., & Wu, C. (2024). Application of deep learning in career planning and entrepreneurship of college students. Journal of Computational Methods in Science and Engineering, 24(4-5), 2927-2942.

[26] Sun, C. (2024, February). Design and implementation of career planning system based on machine learning. In 2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA) (pp. 68-73). IEEE.

[27] Ye, D., Li, M., Han, M., & Wang, X. (2022, January). SWOT Analysis and Strategies of Postgraduate Precise Career Education in the Context of Big Data. In 2022 3rd International Conference on Education, Knowledge and Information Management (ICEKIM) (pp. 876-879). IEEE.

[28] Guo, J., & Qi, L. (2022). Visualization research of college students' career planning paths integrating deep learning and big data. Mathematical Problems in Engineering, 2022(1), 6006968.

[29] Lu, L. (2023, December). Research on the Path of Building the Innovation and Entrepreneurship Education Ecosystem of Local Universities Based on the Big Data Platform. In International Conference on Computer Science and Education (pp. 167-177). Singapore: Springer Nature Singapore.

[30] Chen, L., & He, J. (2023). Application of big data in entrepreneurship and innovation education for higher vocational teaching. International Journal of Information Technology and Web Engineering (IJITWE), 18(1), 1-16.

[31] Yuan, F., & Hu, D. (2021, April). Research on the application of big data technology in college students' innovation and entrepreneurship guidance service. In Journal of Physics: Conference Series (Vol. 1883, No. 1, p. 012171). IOP Publishing.

[32] Liu, T. (2021, May). Thinking of College Students' innovation and Entrepreneurship Education under the background of big data. In 2021 6th International Conference on Smart Grid and Electrical Automation (ICSGEA) (pp. 546-550). IEEE.

[33] Aijun, D. (2020, February). Analysis on the evaluation model of the teaching effect of innovation and entrepreneurship course based on big data. In 2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) (pp. 848-851). IEEE.

[34] Deng, B., & Wu, J. (2021). The cultivation of innovation and entrepreneurship skills and teaching strategies for college students from the perspective of big data. Arabian Journal for Science and Engineering, 1-14.

[35] He, L. (2020). Innovation and Entrepreneurship Education in Higher Vocational Colleges Under Big Data. In Data Processing Techniques and Applications for Cyber-Physical Systems (DPTA 2019) (pp. 979-984). Springer Singapore.

[36] Housen Li & Frank Werner. (2020). Empirical risk minimization as parameter choice rule for general linear regularization methods. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques(1),405-427.

[37] Bo Yu,Jian Liang & Jiann Wen Woody Ju. (2025). Classification method for crack modes in concrete by acoustic emission signals with semi-parametric clustering and support vector machine. Measurement116474-116474.

[38] Baohang Zhang,Ziqian Wang,Haotian Li,Zhenyu Lei,Jiujun Cheng & Shangce Gao. (2024). Information gain-based multi-objective evolutionary algorithm for feature selection. Information Sciences120901-.

[39] Abdullah Saeed Ghareb,Azuraliza Abu Bakar & Abdul Razak Hamdan. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems With Applications31-47.