applied sciences

MDPI

Article

Career Choice Prediction Based on Campus Big Data—Mining the Potential Behavior of College Students

Min Nie [1], Zhaohui Xiong [1], Ruiyang Zhong [2], Wei Deng [3,4,*] and Guowu Yang [1,*]

1 Big Data Research Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; niemin@ebigdata.org (M.N.); mehuizaia@163.com (Z.X.)
2 Faculty of Business and Economics, University of Hong Kong, Hong Kong 999077, China; cicely_zzz@yahoo.com
3 Center for Statistical Research, Southwestern University of Finance and Economics, Chengdu 611130, China
4 Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
* Correspondence: dengwei@swufe.edu.cn (W.D.); guowu@uestc.edu.cn (G.Y.)

Received: 22 February 2020; Accepted: 14 April 2020; Published: 20 April 2020

Abstract: Career choice has a pivotal role in college students' life planning. In the past, professional career appraisers used questionnaires or diagnoses to quantify the factors potentially influencing career choices. However, due to the complexity of each person's goals and ideas, it is difficult to properly forecast their career choices. Recent evidence suggests that we could use students' behavioral data to predict their career choices. Based on the simple premise that the most remarkable characteristics of classes are reflected by the main samples of a category, we propose a model called the Approach Cluster Centers Based On XGBOOST (ACCBOX) model to predict students' career choices. The experimental results of predicting students' career choices clearly demonstrate the superiority of our method compared to the existing state-of-the-art techniques by evaluating on 13 M behavioral data of over four thousand students.

Keywords: campus behavior; career choice prediction; cluster centers; XGBOOST

1. Introduction

According to Erikson's theory [1], identity development primarily relates to career identity, which is mainly developed during adolescence. A student's career identity is probably shaped by adequate career exploration and consecutive commitment at school [2]. Therefore, career counseling services at universities are significant in helping students find their career goals, which is the reason for many special job counseling centers having been established. The major challenge is to reveal important factors that affect students' career planning. From the psychological point of view, collecting, screening, and evaluating relevant personal information is a cognition-based approach to providing career counseling service [3]. Specifically, students are supposed to develop abilities and skills in understanding themselves to be able to participate in occupational decision-making. However, due to the complexity of each person's goals and ideas, it is difficult for students to clearly determine their postgraduation destinations. In contrast, from an empirical point of view, the students' inner interests and future postgraduation destinations can be effectively ascertained by exploring behavioral data of students at school, which makes students' behavioral data essential for their career planning.

The self-perception theory presumes that human behavior can be used to infer a person's goals and intrinsic motivation [4]. Due to the development of information technology, in modern universities there is a growing trend to augment physical facilities with sensing, computing, and communication

capabilities. This means that all behavioral data of students on campus can be recorded in real time through the campus information system. Such behavioral data can reflect the students' unique habits, abilities, preferences, and state of mind [5]. Furthermore, accumulating such data continually provides a way for students to better understand themselves by using data-mining techniques [6]. In contemporary research, differences and regularities of behaviors of various types of graduates of a school have been analyzed by using a data-mining classification algorithm [7]. Additionally, theories can be applied in practice. For example, we can not only establish a set of teaching approaches according to the actual circumstances of students at school but also ensure that students can be better educated according to their own personal conditions. As a result, students can plan their own careers based on their actual personal circumstances to effectively alleviate the problem of difficulty finding employment [8].

Using behavior data to predict students' career choices is a challenging task. Although existing studies use various machine learning algorithms, problems of low precision and models' poor performance exist. Hence, motivated by the social influence theory [9], we further analyze the correlation of each student's career choice with choices of students behaviorally similar to him/her. There are three challenges to this process. First, career choices can be divided into four major categories—employment, postgraduate studies, further studies abroad, and others. Thus, this process is a multiclass learning task. To enhance the performance of multiclass learning, prototypical cluster centers are calculated as priori information for each college. In this paper, the promotion of prototypical cluster centers to multiclass learning is testified by experiments.

Second, as there is aggregation in student groups [10], cluster centers is used to help the model capture information in behavioral data. Prototype is widely used in machine learning, and it aims to let us make use of priori knowledge, thereby achieving better results. In this paper, we propose a new regularization method to compensate for the gap between the examples of students and prototypical cluster centers. More specifically, the output of each instance of our model and its corresponding cluster center should be similar. Then, such normalization will naturally encourage the local smoothness of the learning function and will hence achieve the purpose of improving the accuracy based on the original model.

Lastly, behavior data is massive and mixed, including completely different types of data, such as library records, dormitory entries and exits, consumption at campus locations, book borrowing, and academic achievements [11]. In order to predict students' career choices based on behavior data, data mining approaches such as feature engineering [12] is introduced. Before training our model, we cluster students according to their college information. Afterwards, we establish many new behavior-based representative factors that affect a student's career selection. Inspired by Reference [13], a behavioral entropy index is established to measure the regularity of student behavior. Later, we will discuss in detail how we establish new factors.

Specifically, in this paper we collect (mainly through campus smart cards) 4634 students' longitudinal behavioral data spanning almost three years. Based on the statement that clustering the data by feature "college" can capture the connections between students, cluster centers are calculated as prototypes for each college [14,15]. For all instances in a cluster, the cluster center, which is a feature vector, represents their average band. In other words, the cluster center can be a new instance with the average label. Such prototype approach brings multiple important advantages for multiclass learning.

The framework of our model is shown in Figure 1. First, four types of behavioral features are generated based on campus behavioral data: mastery of professional skills, behavior regularity, reading interest, and family economic status. Next, the Approach Cluster Centers Based on XGBOOST (ACCBOX) model can be obtained by our prototypical cluster center generation method and a novel regularization item. Finally, the resulting predicted career choice is presented. We use actual career choice data to evaluate our model.

In summary, we make the following contributions:

(1) We collect behavioral data of students at school. Using the location and timestamp of the student's card at the school, the designed algorithm constructs a behavioral entropy index that

measures the regularity of student behavior. The behavioral entropy index are used to describe the regularity of students' behavior at school, and the differences in the behavioral patterns of students in different graduating classes are analyzed. Finally, a feature that considers data with and without labels measures the data shift with respect to different years.

(2) Based on student behavior, we propose four factors that reflect student traits. Our study shows that these factors are significantly correlated with students' achievements and career choices.

(3) We perform an extensive evaluation on a real-world dataset covering over 4634 students. To make our results credible, we perform numerous experiments. A methodology called the ACCBOX model is proposed to model behavioral information of students belonging to different clusters. We verify the effectiveness of our method of career choice prediction through experiments on students' behavior dataset.

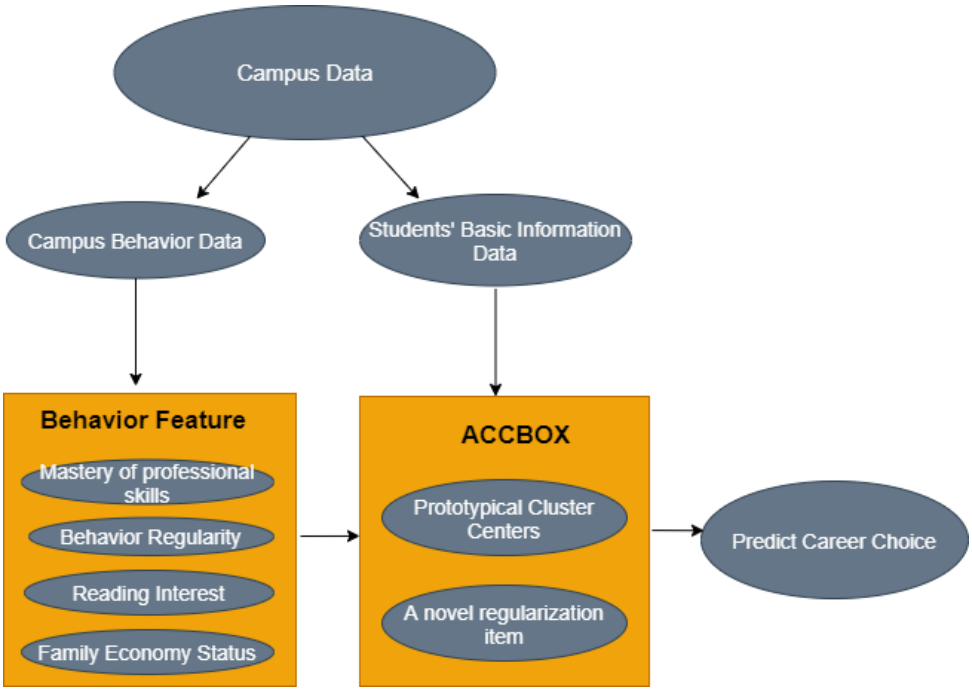

Figure 1. Framework of the proposed Approach Cluster Centers Based On XGBOOST (ACCBOX) model.

2. Related Studies

2.1. Analysis of Factors Influencing Career Choices

Many existing studies focus on factors that influence college students' career choices. In Reference [16], the influence of parents' careers on students' choices was studied. In Reference [17], it was discovered that salary and quality of career advice were the most common factors influencing career choice. In addition, the influence of peers, gender, print media, and interests on career choices has been investigated [18]. One drawback, however, is that all of these studies used questionnaires to study the influencing factors, and ignored students' daily behaviors. In this study, students' daily behavioral data are used to analyze factors that influence career choice.

2.2. Career Choice Prediction

With the rapid development of science and technology, many studies use data generated by online platforms to predict the performance of students based on behavior. For example, in Reference [19], students' academic performance was analyzed by studying students' behavioral data generated on an online platform, namely, a course learning management system (LMS). Additionally, transfer learning

was used to predict individuals' professional expertise through online behavioral data [20]. However, in this study, offline behavioral data is our main focus in predicting students' career choices.

Several studies used offline behavioral features to predict students' future. However, all of them focused on predicting academic performance. The authors of Reference [21] compared the effectiveness of multiple linear regression, a multilayer perceptron network, a radial basis function network, and a support vector machine in predicting academic performance, and the support vector machine was observed to attain the best predictive performance. A questionnaire was utilized to collect data about students' social media use for collaboration and communication, and that data was subsequently used to analyze the influence of social media use on students' academic performance [22]. In addition, multilevel regression based on LMS data was used in predicting academic performance [23]. Though the existing methods use mainstream machine learning models to predict students' future, they all ignore the distribution differences between different student clusters of, for example, students in different colleges, which has important implications for predicting students' future. In our study, generating prototypical cluster centers is proposed to capture this information, thus improving the performance of our learner.

3. Behavioral Factors

To mine information correlated with career choices from students' behavioral data, four types of behavioral features are generated as follows.

3.1. Analysis of Learning the Level of Mastery of Professional Skills

In this paper, students' professional skills can be extracted from course scores they have earned during school days. However, we still face numerous questions in doing so. First, there are thousands of courses in a university. If the score of each course is taken as a feature, that feature representation will face the challenge of sparsity. In addition, if a student performs well in several courses, we believe that he or she has effectively attained some professional skills. Extracting professional skills is, however, a difficult problem indeed. Therefore, to extract such features, we use a matrix factorization based on a dimensionality reduction algorithm. As to students, we denote the matrix of the grades of students by $R \in \mathbb{R}^{M \times E}$, where each element $r_{i,j}$ in R represents the grade of student u_i in class c_j. We divide sparse matrix R into two matrices, denoted by $P_{M \times K}$ and $Q_{K \times E}$:

$$R_{M \times E} \approx P_{M \times K} \times Q_{K \times E} = \hat{R}_{M \times E}. \tag{1}$$

In these two matrices, M represents the number of students, E represents the number of classes, and K refers to K kinds of potential professional skills. In this case, matrix $P_{M \times K}$ represents the professional skills of each student, and $Q_{K \times E}$ embodies the relationships between these K kinds of skills and each class in this college.

To ensure that the product of the two matrices $P_{M \times K}$ and $Q_{K \times E}$ approximate $R_{M \times E}$, we transform this question into a regression problem in data mining. We define the squared error between the original matrix $R_{M \times E}$ and the new matrix $\hat{R}_{M \times E}$ as the loss function. Additionally, to increase our model's generalizability, we add the L_2 norm of the two matrices to the function as a regularization item. Hence, the objective function is defined as follows:

$$\min_{P,Q} \sum_{i,j} I_{i,j} \left(r_{i,j} - p_i' q_j \right)^2 + \lambda \left(\sum_i \| p_i \|_2 + \sum_j \| q_j \|_2 \right). \tag{2}$$

In the objective function, $I_{i,j}$ indicates whether student u_i has taken class c_j. Vector $p_i \in \mathbb{R}^K$ is the ith row vector in matrix $P_{M \times K}$, while $q_j \in \mathbb{R}^K$ is the jth column vector in matrix $Q_{K \times E}$.

Next, we try to find the solution of $P_{M \times K}$ and $Q_{K \times E}$ in the objective function by stochastic gradient descent, using the following gradient update functions:

$$\begin{aligned} p_i &= \left(\lambda I_K + \textstyle\sum_j I_{i,j} q_j q_j'\right)^{-1} \left(\textstyle\sum_j I_{i,j} r_{i,j} q_j\right) \\ q_j &= \left(\lambda I_K + \textstyle\sum_i I_{i,j} p_i p_i'\right)^{-1} \left(\textstyle\sum_i I_{i,j} r_{i,j} p_i\right) \end{aligned} . \tag{3}$$

We optimize the loss function by using the above gradient update function and calculate the professional skill vector $p_{u,k}$; afterwards, we add this vector into the forecasting model as a feature factor. As a given course may be taught by a different teacher every time, the final grades may not be comparable, and the standard of evaluation may vary. Hence, we need to normalize different grades to make them comparable, using the following processes:

Suppose that both teachers A and B teach course C; teacher A gave student u_i a grade of g_{u_i}, and student u_j while being taught by teacher B earned grade $\hat{g}_{u_j}$; the grade after the normalization is

$$\begin{aligned} g_{ui}' &= g_{ui} - \frac{1}{H} \sum_{i=1}^{H} g_{ui} \\ \hat{g}_{uj}' &= \hat{g}_{uj} - \frac{1}{L} \sum_{j=1}^{L} \hat{g}_{uj} \end{aligned} , \tag{4}$$

where H and L represent the numbers of students in classes of teacher A and teacher B, respectively.

3.2. Analysis of Behavior Regularity

According to the Big Five personality traits (namely, openness, conscientiousness, extraversion, agreeableness, and neuroticism), conscientiousness plays an important role in job and academic performance [24]. The factors we consider are eating breakfast, going to the library, and bathing for the first time every day. Specifically, an entropy of probability that a behavior occurs within specific time intervals can measure the regularity of daily behaviors. Assume that period T is divided into n time intervals: $T = \{t_1, \ldots, t_n\}$.

For each student, the probability of behavior $v \in V = \{\text{“}eating\ breakfast\text{”}, \text{“}going\ to\ the\ library\text{”}, \text{“}bathing\text{”}\}$ occurring within a given time interval t_i can be calculated as

$$P_v\left(T = t_i\right) = \frac{n_v\left(t_i\right)}{\sum_i n_v\left(t_i\right)}, \tag{5}$$

where $n_v\left(t_i\right)$ refers to the frequency of occurrence of behavior v within a given time interval t_i. The entropy of behavior v is then calculated as

$$E_v = -\sum_{i=1}^{n} P_v\left(T = t_i\right) \log P_v\left(T = t_i\right). \tag{6}$$

We know that the regularity of a behavior is low due to its probability's uniform distribution over time intervals, while its entropy is high. In the computation of entropy, time periods and time interval spans can be different according to different behavior characteristics. For example, the time period of breakfast behavior is set from 6 a.m. to 10 a.m., and its time interval is half an hour.

3.3. Analysis of Book Reading Interests

We can learn what a student is interested in from borrowing archives' records of libraries that partly correlate with future vocational choices. There are millions of books in the library, and each student may borrow only one of them. Directly counting the quantity of each student's book borrowing for each book will cause the problem of vector sparsity. However, each book has several very rich attributes, such as book classification. Therefore, we can use the Chinese library classification to define the respective categories for each book. Considering the accuracy of the final partition and the sparsity of the vector, the second-level partition of the Chinese library classification is used as the criterion,

and there are approximately 250 dimensions in total. We compute the frequency of borrowing books of each student within more than 200 book categories, and define the frequency as a feature vector that characterizes a student's personal interests; that is,

$$\mathbf{S} = \left(G_1, G_2, G_3, \cdots, G_\zeta\right). \tag{7}$$

In this equation, S denotes the feature vector of a student's interests, G_j represents the number of books borrowed by students in category j, and there are ζ categories of books in total. Considering that the frequency of borrowing varies across students, to make each element in the feature vector comparable, it is necessary to normalize the frequency with the following equation:

$$G'_j = \frac{G_j}{\sum_{j=1}^{\zeta} G_j}. \tag{8}$$

The feature vector after normalization is $S' = \left(G'_1, G_{2'}, G'_{3'}, \ldots, G'_\zeta\right)$; we then add the new feature vector to the prediction model to improve the accuracy and interpretability of the model.

3.4. Analysis of Family Economic Status

We can assess students' economic conditions by using a questionnaire, but because students may not be able to estimate their family economic conditions well, and because of geographical differences, it is difficult for us to unify the criteria. Therefore, we calculate daily expenditures through the consumption of students in the cafeteria and supermarkets. Afterwards, we use first- and second-order descriptive statistics, including torsion, discrete, median, mean, quartile range, standard deviation, and kurtosis values, to assess each family's economic conditions.

Second, we calculate the ratio of transitions on weekends and weekdays, and subsequently perform a fast Fourier transform (FFT); we obtain the consumption cycle by calculating the total energy as the sum of squares of components of each FFT, which provides more information about families' economic status. To eliminate the influence of consumption level, each value of sequence $[x_1, x_2, \cdots, x_n]$ should be reduced by the mean value of the sequence. We define the energy based on the converted sequence $[\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n]$ as follows:

$$\begin{array}{c} \text{Energy} = \frac{\sum_{i=0}^{n-1} |F_i|^2}{n} \\ F_i = e^{-j2\pi \frac{ki}{n} \tilde{x}_i} \end{array}, \tag{9}$$

where $\tilde{x}_i = x_i - \sum x_i / n$, and symbol j denotes the imaginary unit.

4. Model Introduction

At present, the traditional machine learning method is limited in solving the problems of students' postgraduation plans. For a group of students, there must be a certain connection between them, and this problem could be understood better by mining the relationship between students. First, we find the connections among students through the clustering method; however, large differences between students will affect the performance of the clustering algorithm. Fortunately, we can obtain this a priori information from each college. By building information bridges, we naturally connect related students. In this section, we present our approach called the "Approach Cluster Centers Based On XGBOOST" (ACCBOX). ACCBOX proceeds by taking three elementary steps: prototypical cluster center generation, model training, and optimization.

4.1. Problem Statement

In this subsection, we will introduce some notation and then formally define our main contribution of this paper. In a university, let $U = \{U_1, U_2, U_3 \ldots, U_C\}$ denote the set of colleges. For every student

i, we denote the feature vector and the student's career choice by $x_i \in \mathbb{R}^{1\times p}$ and y_i, respectively. Parameter p represents the total number of dimensions of students' features after using the methods of feature generation mentioned in Section 3.

Let $x = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{N\times p}$ and $y = [y_1, y_2, \ldots, y_N] \in \mathbb{R}^N$ denote the feature matrix and career choices of all students, where N is the total number of all students. $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ denotes students' behavioral data and their career choices, where n is the number of students. The detail of features will be covered in the next section. We then formally define our method of career choice prediction as follows.

Career Choice Prediction: Given the feature vector x_i of every student, we are supposed to predict the corresponding career choice y_i. To build our ACCBOX model, we first introduce the prototypical cluster center. Then the optimization method of training model parameters is been expounded.

4.2. Prototypical Cluster Center

As shown in Figure 2, based on the statement that the most remarkable characteristics of the class are supposed to be reflected by the main samples of a category, we determine the cluster center for college U. The size of college U is C. Next, we calculate a main sample for each cluster. For clusters $U = \{U_1, U_2, \ldots, U_C\}$, cluster center z_j is defined as

$$z_j = \frac{1}{|U_j|} \sum_{i=1}^{n} x_i \cdot \mathbb{I}\left(x_i \in U_j\right), \tag{10}$$

where $\mathbb{I}(\cdot)$ is an indicator function, that is, $\mathbb{I}\left(x_i \in U_j\right)$ equals 1 if $x_i \in U_j$ is true and equals 0 otherwise. Similarly, let t_j denote the labeling information of U_j; then, we treat the the average vector t_j of students' career choices from the college U_j as the career choice prototype of each college.

$$t_j = \frac{1}{|U_j|} \sum_{i=1}^{n} y_i \cdot \mathbb{I}\left(x_i \in U_j\right). \tag{11}$$

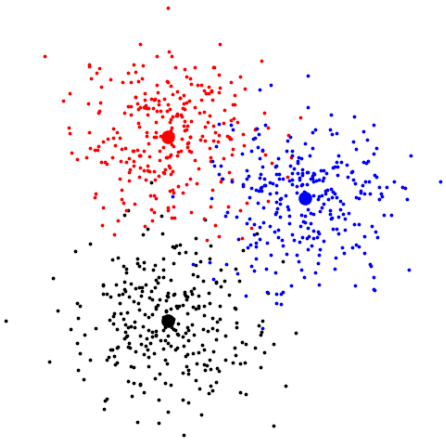

Figure 2. The most remarkable characteristics of the class are supposed to be reflected by the main samples of a category.

Accordingly, the prototypes of colleges are defined as $\mathcal{D}' = \{\mathbf{z}_j, \mathbf{t}_j\}_{j=1}^C$. The students from the same college usually have similar career choices. The prototype of each college can reveal the priori information of the career choice of students from the corresponding college. Hence, we also require the model to approach the prediction of students' career choice to that of the prototype. Experiments in Section 5.2 shows that this effectively improves the learning ability of our model.

4.3. Model Training

According to [25], XGBOOST model $\phi(x_i)$ can be described as:

$$\hat{\mathbf{y}}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i),\ f_k \in \mathcal{F}, \tag{12}$$

where $\hat{y}_i$ is the prediction of students' career choice, K is the number of decision tree in XGBOOST, and f_k is the k-th decision tree in CART tree set $\mathcal{F}$. Followed by XGBOOST [25], the objective function could be designed as follows:

$$\min_{\phi} \frac{1}{2} \sum_{i=1}^{n} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2 + \frac{\alpha}{2} \Omega(\phi), \tag{13}$$

where the first item is the conventional loss function and the second item is the regularization item to avoid overfitting. The trade-off hyperparameter α reduce the model capacity. $\Omega(f)$ of each tree can be described as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{14}$$

where T is the number of leaves in the tree, w is the scores of all leaves, both γ and λ are the trade-off parameters.

We use the soft labels of each college that are treated as career choice prototypes to regularize the model's predictions for the students of each college. Thus, the novel regularization item is specified as:

$$\min_{\phi} \sum_{i=1}^{n} \|\phi(\mathbf{x}_i) - \sum_{j=1}^{c} \mathbf{t}_j \cdot \mathbb{I}(\mathbf{x}_i \in \mathbf{U}_j)\|_2^2. \tag{15}$$

Combining Equations (13) and (14) leads to the following final objective function:

$$\min_{\phi} \frac{1}{2} \sum_{i=1}^{n} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2 + \frac{\alpha}{2} \Omega(\phi) + \frac{\beta}{2} \sum_{i=1}^{n} \|\phi(\mathbf{x}_i) - \sum_{j=1}^{c} \mathbf{t}_j \cdot \mathbb{I}(\mathbf{x}_i \in \mathbf{U}_j)\|_2^2, \tag{16}$$

where β is a trade-off parameter that controls the importance of the regularization item.

4.4. Optimization

XGBOOST is a commonly used algorithm in machine learning that performs very well on most classification tasks. Unfortunately, it is not very good at the task of predicting students' postgraduation plans. However, applying the method we propose to XGBOOST can improve its effectiveness.

Similarly to other boosting algorithms, XGBOOST is an iterative decision tree algorithm; its base learner is a classification and regression tree (CART) and constructs an integration model in a phase-iterative manner. The ACCBOX algorithm's iterative update process is shown in Algorithm 1.

Algorithm 1: Iterative update of ACCBOX

Input:
the students' behavior dataset D, the colleges' prototype dataset D′, the college clusters U.
Loss function:

$$\min_{\phi} \frac{1}{2}\sum_{i=1}^{n} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2 + \frac{\alpha}{2}\Omega(\phi) + \frac{\beta}{2}\sum_{i=1}^{n} \|\phi(\mathbf{x}_i) - \sum_{j=1}^{c} \mathbf{t}_j \cdot \mathbb{I}(\mathbf{x}_i \in \mathbf{U}_j)\|_2^2. \tag{17}$$

Output: CART tree ensemble $\phi(x, z)$

Initialize the model:

$$F_0(\mathrm{x}, \mathrm{z}) = wf_0(X, Z) \tag{18}$$

k=1

while(k<=K)

Calculate the residuals:

$$g_i = \phi(\mathbf{x}_i) - y_i \tag{19}$$

$$\hat{g}_i = \phi(\mathbf{x}_i) - \sum_{j=1}^{c} \mathbf{t}_j \cdot \mathbb{I}(x_i \in U_j). \tag{20}$$

Fit a CART tree f_k to the above three residuals. Find the optimal weight for f_k by minimizing the following loss function:

$$\begin{aligned} w_k = \arg\min_{w} & \frac{1}{2}\sum_{i=1}^{n} \|F_{k-1}(x_i) + wf_k(x_i) - \mathbf{y}_i\|_2^2 \\ & + \frac{\alpha}{2}\Omega(\phi) \\ & + \frac{\beta}{2}\sum_{i=1}^{n} \|F_{k-1}(x_i) + wf_k(x_i) - \sum_{j=1}^{c} \mathbf{t}_j \cdot \mathbb{I}(\mathbf{x}_i \in \mathbf{U}_j)\|_2^2. \end{aligned} \tag{21}$$

Update the model:

$$F_k(x, z) = F_{k-1}(x, z) + \alpha w_k f_k(x, z), \tag{22}$$

where α refers to the learning rate of our model k=k+1;

end while;

output

$$\phi(x, z) = F_K(x, z). \tag{23}$$

At each iteration, ACCBOX uses the integrated model obtained at that stage to calculate the residual of the model's predicted and true values. There are two parts in the model: the residual of the student's career choice and the predicted value, the residual between the prototypical cluster centers' data and the predicted value, as shown in Equations (19) and (20).

To update the existing integration model, it is necessary to train a CART tree in each iteration to fit the above three residuals and add that tree to our integration model. To ensure that the addition of that tree can benefit our model, we need to continuously optimize the parameters of that tree. The optimal parameters of that tree can be obtained by minimizing the loss function as shown in Equation (21). As a result, we can obtain the optimal tree ensemble by K-round iteration, as shown in Equation (23).

In summary, we have fully introduced our ACCBOX model through three steps: prototypical cluster center generation, model training, and optimization. The pseudocode of the model algorithm is shown in Algorithm 2.

Algorithm 2: The Approach Cluster Centers Based On XGBOOST (ACCBOX) algorithm.

Input:
$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$ the students' behavior dataset
α, β : the regularization hyperparameters
U : the college clusters
$\hat{x}$: the student test set
Output:
$\hat{y}$: the prediction of student's career choice
1 divide x into different colleges' clusters;
2 calculate college cluster centers' prototypes $\{z_j\} \in R^{c \times d}$ according to Equation (10);
3 calculate prototypes' career choice labels $\{t_j\} \in [0,3]^{c \times 1}$ according to Equation (11);
4 train the best tree ensemble model structure according to Algorithm 1;
5 using the tree ensemble model to predict student's career choice $\hat{y}$;

5. Experiments

In this section, we first introduce the dataset and the settings used for evaluation. Afterwards, we report experimental results and discuss them.

5.1. Dataset and Settings

The evaluation uses a dataset of smart card data of 4634 students of the same grade from 16 colleges. The total number of consumption records is 13,122,696. We collect data from one university during 2010/09/01 to 2014/06/30. This dataset consists of four types of data: academic performance data, basic information data of students, behavior data, and career choice data. The students have borrowed 95,493 books, resulting in 391,637 book loan records. They have taken 1358 courses and generated 336,353 course grade records. The numbers of library records and dormitory entries and exits are 1,048,576 and 727,260, respectively. Students' behaviors include library and dormitory entries and exits, consumption at campus locations (e.g., a canteen and a supermarket), book borrowing, and academic achievements.

The experimental training group and the test group are divided according to a 70%–30% ratio. All hyperparameters are tuned based on accuracy in 5-fold cross validation.

In the experiment, we divide the training group into 16 colleges, and the central label of each college is the mean feature of students in that college. The final objective function of our proposed approach that incorporates prototypical cluster center generation and a novel regularization item for career choice prediction is shown in Equation (16). We search for regularization hyperparameters α, and β in the interval of $\left[10^{-4}, 1\right]$ with a step size of 10.

In our experiments, we use the accuracy, recall, precision, and micro-F1 as our evaluation metrics.

5.2. Result and Discussion

5.2.1. Feature Importance

We construct the following four types of features for career choice prediction.

- R: Students' interest in reading.
- M: Mastery of professional skills.
- B: Behavior regularity.
- F: Family economic status.

As shown in Table 1, the model performance under different feature combinations by utilizing one study center's data is compared. We generate a total of 16 feature combinations from these four features (M: Mastery of Professional Skills in Section 3.1, B: Behavior Regularity in Section 3.2, R: Reading

Interests in Section 3.3, F: Family Economic Status in Section 3.4) in Section 3 in a full combination way (C(4,4) = 16). To compare the model performance under these 16 feature combinations, first we input the four types of features into the classification algorithm so that we can obtain the performance of each type of features. It is clear that each type of feature helps refine students' post graduation predictions, with mastery of professional skills having the greatest impact. This should be quite intuitive, as common sense tells us that the main goal of college students is to learn professional skills that will help them in their future careers. Additionally, a person's interest in reading reflects that person's preference for learning, and life, behavioral regularity, and family economic conditions affect lifestyle, which will affect career choice prediction.

Table 1. Performance evaluation under different feature conditions (showing performance on the XGBOOST model).

	Feature				**Metrics**			
	R	**M**	**B**	**F**	**Acc**	**Micro-F1**	**Precision**	**Recall**
1					0.462	0.484	0.481	0.487
2	✓				0.493	0.511	0.494	0.530
3		✓			0.555	0.567	0.557	0.547
4			✓		0.486	0.506	0.488	0.525
5				✓	0.487	0.512	0.494	0.531
6	✓	✓			0.570	0.590	0.573	0.608
7	✓		✓		0.506	0.521	0.501	0.543
8	✓			✓	0.507	0.524	0.495	0.557
9		✓	✓		0.568	0.589	0.572	0.607
10		✓		✓	0.575	0.584	0.570	0.599
11			✓	✓	0.505	0.514	0.498	0.531
12		✓	✓	✓	0.581	0.600	0.575	0.627
13	✓		✓	✓	0.526	0.547	0.510	0.590
14	✓	✓		✓	0.586	0.597	0.576	0.620
15	✓	✓	✓		0.572	0.586	0.575	0.597
16	✓	✓	✓	✓	0.604	0.622	0.617	0.627

5.2.2. Prediction Performance

We compare the proposed algorithm with the following competing baselines. All methods use the four types of features listed in Table 1.

- Decision tree used for career choice prediction.
- SVM: A kind of generalized linear classifier that classifies data based on supervised learning.
- Random forest used for career choice prediction.
- Logistic regression used for career choice prediction.
- XGBOOST [25] is a boosting-tree-based method.
- ACCBOLR is a new method we propose based on logistic regression, incorporating prototypical cluster center generation, and a novel regularization item.
- ACCBOX represents the application of our proposed method based on XGBOOST, incorporating prototypical cluster center generation and a novel regularization item.

As shown in Table 2, adding the regularization item and prototype examples improves the predictive accuracy of logistic regression and XGBOOST models. By using our method, the accuracy of ACCBOX increases from 0.604 to 0.638, the Micro-F1 score rises from 0.622 to 0.647, Micro-Precision increases from 0.617 to 0.629, and Micro-Recall rises from 0.627 to 0.666. The accuracy of ACCBOLR increases from 0.605 to 0.623, the Micro-F1 score rises from 0.607 to 0.624, Micro-Precision increases from 0.601 to 0.626, and Micro-Recall rises from 0.615 to 0.623. In contrast, the performance of other models is much worse. The accuracy of SVM is 0.606, the value of Micro-F1 is 0.621, Micro-Precision is 0.610, and Micro-Recall is 0.632. The accuracy of the random forest method is 0.623, the value of Micro-F1 is 0.633, Micro-Precision is 0.620, and Micro-Recall is 0.647. The performance of the

decision tree is the worst: its accuracy is 0.532, the Micro-F1 score is 0.562, Micro-Precision is 0.541, and Micro-Recall is 0.584.

As displayed in Table 3, we perform a further ablation study to demonstrate the contribution of each design in ACCBOX. The second row represents the XGBOOST method trained with reducing the model capacity. In the third row, only the novel regularization approach is added in XGBOOST. It is clear that the novel regularization term makes a good contribution to regularizing the learner.

The experimental results prove that our approach outperforms the state-of-the-art counterparts. It is clear that by using the prototypical cluster center generation approach and the novel regularization item contribute to the performance of the model.

Table 2. Comparison of ACCBOLR and ACCBOX with baselines.

Model	Acc	Micro-F1	Precision	Recall
Decision Tree	0.532	0.562	0.541	0.584
SVM	0.606	0.621	0.610	0.632
Random Forest	0.623	0.633	0.620	0.647
Logistic Regression	0.605	0.607	0.601	0.615
XGBOOST	0.604	0.622	0.617	0.627
ACCBOLR	0.623	0.624	0.626	0.623
ACCBOX	0.638	0.647	0.629	0.666

Table 3. Ablation study of contributions of different modules.

Model	Acc	Micro-F1	Precision	Recall
XGBOOST	0.604	0.622	0.617	0.627
XGBOOST+RC	0.613	0.633	0.621	0.646
XGBOOST+NR	0.622	0.634	0.619	0.650
ACCBOX	0.638	0.647	0.629	0.666

5.3. Effect of α and β

As Figure 3 shows, as α increases in the ACCBOX model, the model attains the highest accuracy if α is 10^{-2}; subsequently, the accuracy decreases gradually. As β increases, the accuracy of the model increases gradually. If it is 10^{-2}, the accuracy reaches the highest level, and subsequently decreases gradually.

In conclusion, α should be 10^{-2}, and β should also be 10^{-2}.

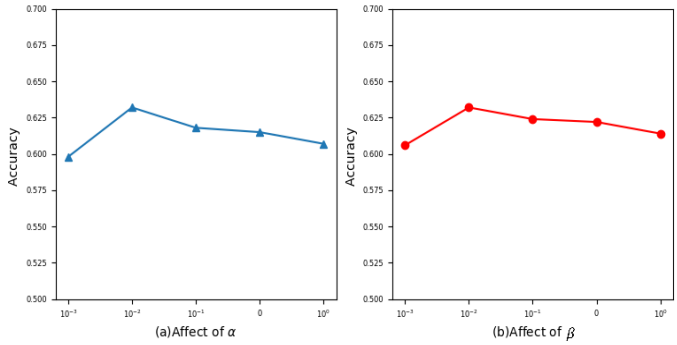

Figure 3. (**a**) Effect of reducing the model capacity. (**b**) Effect of a novel regularization term.

6. Conclusions

In this paper, we have studied college students' career choices based on their professional skills, behavior regularity, and other related behaviors. Additionally, the study has offered several important insights into improving the model.

We have proposed a prototypical cluster center generation approach to use the priori information from each college. Motivated by the cluster assumption that examples in the same cluster should have the same label, we have introduced a novel regularization item to bridge the gap between the real-world examples and prototypical cluster centers. The results of multiple experiments demonstrate that our approach is superior to other approaches to career choice prediction.

In future studies, three directions can be followed with interest. First Cluster Centers can be discovered in a more precise method. In addition, our model can be extended from using only behavioral data to using multimodal data, such as adding school achievement and questionnaire data. Furthermore, it is meaningful to improve our model to not only predict career choices but also advise on career planning, such as advising on the courses required.

Author Contributions: All authors have contributed equally to the work. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China under grant 2016YFB1000905, and the National Natural Science Foundation of China under Grant No. 61572109.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Erikson, E.H. *Identity: Youth and Crisis*; WW Norton & Company: Manhattan, NY, USA, 1994; pp. 176–200.
2. Marcia, J.E.; Waterman, A.S.; Matteson, D.R.; Archer, S.L. *Ego Identity: A Handbook for Psychosocial Research*; Springer Science & Business Media: New York, NY, USA, 2012.
3. Gati, I.; Krausz, M.; Osipow, S.H. A taxonomy of difficulties in career decision making. *J. Couns. Psychol.* **1996**, *43*, 510–526. [CrossRef]
4. Bem, D.J. Self-perception theory. *Adv. Exp. Soc. Psychol.* **1972**, *5*, 1–62.
5. Albion, M.J.; Fogarty, G.J. Factors influencing career decision making in adolescents and adults. *J. Career Assess.* **2002**, *10*, 91–126. [CrossRef]
6. Festinger, L. A theory of social comparison processes. *Hum. Relations* **1954**, *7*, 117–140. [CrossRef]
7. Nie, M.; Lian, D. Advanced Forecasting of Career Choices for College Students Based on Campus Big Data. *Front. Comput. Sci.* **2018**, *12*, 494–503. [CrossRef]
8. Hunt, D. *Matching Models in Education: The Coordination of Teaching Methods with Student Characteristics*; Ontario Institute for Studies in Education: Toronto, ON, Canada, 1971; Monograph 87.
9. Yao, H.; Nie, M.; Su, H.; Xia, H.; Lian, D. Predicting Academic Performance via Semi-supervised Learning with Constructed Campus Social Network. In Proceedings of the Database Systems for Advanced Application(DASFAA), Suzhou, China, 27–30 March 2017; pp. 597–609.
10. Keesling, J.W. A problem in the aggregation of student data to the level of school. In Proceedings of the American Educational Research Association Meeting, New Orleans, LA, USA, 25 February–1 March 1973; Volume 1.
11. Thompson, M.N.; Subich, L.M. The relation of social status to the career decision-making process. *J. Vocat. Behav.* **2006**, *69*, 289–301. [CrossRef]
12. Kuhn, M.; Johnson, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*; CRC Press: Boca Raton, FL, USA, 2019; p. 9.
13. Boer, E.R. Behavioral Entropy as an Index of Workload. In Poceedings of the Human Factors and Ergonomics Society Annual Meeting, San Diego, CA, USA, 29 July–4 August 2000; pp. 125–128.
14. Jain, A.K.; Dubes, R.C. Algorithms for clustering data. *Technometrics* **1988**, *32*, 227–229.
15. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. *ACM Comput. Surv.* **1999**, *31*, 264–323. [CrossRef]
16. Suryadi, B.; Sawitri, D.R.; Hayat, B.; Putra, M.D.K. The Influence of Adolescent-Parent Career Congruence and Counselor Roles in Vocational Guidance on the Career Orientation of Students. *Int. J. Instr.* **2020**, *13*, 45–60. [CrossRef]
17. Oshodi, O.S.; Aigbavboa, C.; Babatunde, O.K.; Arijeloye, B.T. Apprenticeship: A Narrative Review of Factors Influencing Career Choice of Young People. In Proceedings of the Construction Industry Development Board Postgraduate Research Conference, Johannesburg, South Africa, 28–30 July 2019; pp. 90–99.

18. Kazi, A.S.; Akhlaq, A. Factors Affecting Students' Career Choice. *J. Res. Reflections Educ.* **2017**, *11*, 187–196.
19. Alalwan, N.; Al-Rahmi, W.M.; Alfarraj, O.; Alzahrani, A.; Yahaya, N.; Al-Rahmi, A.M. Integrated Three Theories to Develop a Model of Factors Affecting Students' Academic Performance in Higher Education. *IEEE Access* **2013**, *7*, 98725–98742. [CrossRef]
20. Li, J.; Liu, C.; Liu, B.; Mao, R.; Wang, Y.; Chen, S.; Yang, J.-J.; Pan, H.; Wang, Q. Diversity-aware retrieval of medical records. *Comput. Ind.* **2015**, *69*, 81–91. [CrossRef]
21. Huang, S.; Fang, N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Comput. Educ.* **2013**, *61*, 133–145. [CrossRef]
22. Brozina, C.; Knight, D.B.; Kinoshita, T.; Johri, A. Engaged to Succeed: Understanding First-Year Engineering Students' Course Engagement and Performance Through Analytics. *IEEE Access* **2019**, *7*, 163686–163699. [CrossRef]
23. Conijn, R.; Snijders, C.; Kleingeld, A.; Matzat, U. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Trans. Learn. Technol.* **2017**, *10*, 17–29. [CrossRef]
24. Dudley, N.M.; Orvis, K.A.; Lebiecki, J.E.; Cortina, J.M. A meta analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *J. Appl. Psychol.* **2006**, *91*, 40–57. [CrossRef]
25. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.