Job Selection Model Based on Svm and Bpneural Network

Sitong Jiang[1], Jingwen Ding[2], Yuxi Liu[3], Qingyi Hu[4]

[1]Beijing National Day School, Beijing, 100000, China

[2]The Affiliated High School of Capital Normal University, Beijing, 100000, China

[3]Beijing No. 80 Middle School, Beijing, 100000, China

[4]The High School Alliliated to Renmin University of China, Beijing, 100000, China

Keywords: Random forest, Questionnaire, Svm, Bp, Summer job selection model

Abstract: There are various kinds of summer jobs for high school students, which makes them get into confused because they don't know how to choose the best job for themselves. Different students have a different preference about the job. For example, some students prefer a high salary because in this case, they can save a huge number of money and then purchase some goods they want; while others prefer doing a meaningful summer job like the volunteer because they can get a unforgettable experience and help the people needed. So, there is an urgent demand to scientifically and comprehensively establish a summer job selection model to help these high school students find the best summer job. In this paper, our group designed a series of mathematical model to help high school students to find the best summer job.

1. Introduction

There are more and more high school students who hope to find a good job during the summer vacation. Because in school, they have been busy with their courses and lack many life experiences. However, different high school students have different expectations for the summer job. So, there is an urgent demand for high students to find the best summer job. In order to help out these high school students, our group designed a series of models to help them find the best summer job[1].

As society is becoming well-developed, a lot of new firms appeared. In this case, the competition between people also becomes extremely drastic. Summer jobs are a good time for high school students to improve their abilities and to gain some working experience that can be helpful in the future. Therefore, in order to find the most suitable summer job for high school students, many different factors should be considered well before we make decisions.

This topic is about determining the "best" choice of a summer job, which means students need to converge all the potential factors to consider which one is their fit. By analyzing the summer job options and evaluating the choices for students, they can have a better understanding of how to choose the jobs and know more extensive about those job fields, which may even contribute to their future life.

In order to solve the problem, we will use some methods by making models give a general recommendation and discuss the results. Then we will design and develop the concept of a user-friendly application in order to meet the personal preference of each individual. The application will be based on the results we make.

Here is how we do. In the beginning, we will do some online research about how to find a job and make interviews with our friends. After comprehensively considering the literature research results and actual interview results, our group selected the significant factors. And then we designed the questionnaire to collect research data and tested the validity and reliability of the questionnaire. In order to predict the choice of summer job of high school students more accurately, two methods (SVM and BP) are adopted to establish the summer job, selection model. According to the results, we finally made a simple ideal diagram of the job selection model for students to easily understand.

  DOI: 10.25236/mmmce.2021.029

2. Assumption and Definition

2.1 Assumptions

Assumption 1: The questionnaire has good validity and reliability.

Justification: Because the credible questionnaire survey plays an important role in the evaluation of the model proposed in this paper.

Assumption 2: The data we collected can depict the main characteristics of high school students.

Justification: Since the main characteristics of high school students are the basic factors to consider in choosing jobs, the main characteristics are of great significance for the model of job selection.

Assumption3: Our consideration doesn't include the situation that happened on way to work.

Justification: Students are in various countries and cities around the world, so the situation like climate may very drastically, which might also affect student's choices for the job. However, the situations on the way to work are too incidentally and uncertain for us to collect so such data in such a short period of time. Hence, we decided not to consider these situations.

2.2 Definition of Variables

Symbol	Explanation
S	Salary
H	Hours
I	Intensity
Se	Security
W	Workplace
F	Facilities
η	Job Selection

3. Factor Selection and Elimination

3.1 Factor Selection and Quantitative Analysis

Facing various summer jobs, high school students will consider numerous factors when they are choosing these jobs. After comprehensively considering the literature research results and actual interview results, our group selected the following 7 factors and divided these factors into 3 categories.

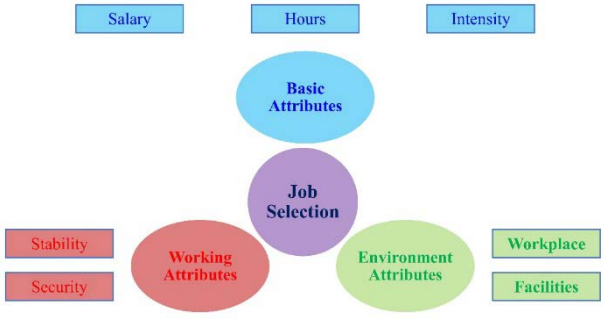

Fig.1 Selected Factors

As shown in Figure 1, the three categories are basic attributes, work attributes and environment attributes, respectively, and the seven factors are salary, hours, intensity, workplace, facilities, stability, and security, respectively.

The basic attributes contain 3 general factors that no matter students or adults will consider when they are choosing a job. It is common sense that salary is the most important factor influencing one to choose his or her job. For better quantitative analysis of salary, this paper divided the salary into 3 levels. The relationship can be expressed as follows:

$$S = \begin{cases} 1, & salary \leq \$500 \\ 2, & \$500 < salary \leq \$1000 \\ 3, & salary > \$1000 \end{cases} \quad (1)$$

S is the quantified salary, and its values are 1, 2, and 3. “1” represents the low level, “2” represents the medium level, and “3” represents the high level.

Working hours are also an important indicator for different people looking for summer jobs. Some people like to work for fewer hours, because they can have more free time to do other things, such as playing with friends and so on. However, the other parts of people prefer to work with more hours, because they can live a more fulfilling life at work. Similar to the quantitative analysis of salary, this paper divided working hours into 3 levels. The results are as follows:

$$H = \begin{cases} 1, & hours \leq 6 \\ 2, & 6 < hours \leq 8 \\ 3, & hours > 8 \end{cases} \quad (2)$$

H is the quantified working hour, and its values are 1, 2, and 3. “1” represents the low level, “2” represents the medium level, and “3” represents the high level.

Work intensity includes workload during the day, overtime, and night shift work. We can also quantify the work intensity. In this case, I is the quantified work intensity, and its values are 1, 2, and 3. “1” represents busy, “2” represents moderate, and “3” represents easy.

The environmental attributes also have a certain impact on the job selection. Some people prefer to work online at home, while others prefer to work outside. In addition, the impact of work facilities should not be ignored.

In this paper, two conditions of the workplace were considered when high school students chose a job. We defined “P” as the quantified factor of the workplace and its values are 1 and 2. “1” represents “working at home” and “2” represents “working outside”. Similarly, we defined “F” as the quantified facilities and its values are 1, 2, and 3. In this case, “1” represents very good, “2” represents moderate, and “3” represents very bad.

If one wants to work safely and stably, the working attributes undoubtfully must be considered. Some high students are not willing to get a dangerous job, so it is necessary to take security into account. To avoid frequently changing jobs, the stability of a job also should be considered when one is choosing a job. This paper quantifies the two factors. “St” was defined as the quantified stability and its values are 1, 2, and 3. “1” represents very stable, “2” represents moderate, and “3” represents very precariously. “Se” was defined as the quantified security and its values are 1, 2, and 3. “1” represents very safe, “2” represents moderate, and “3” represents very dangerously.

3.2 Random Forest Model to Eliminate Factors

3.2.1 Random Forest Model

Random forest is an algorithm based on a decision tree[2-6]. Its main idea is to form some decision trees by using the self-help method, and the final result can be obtained by using the voting method. A decision tree is a prediction model suitable for regression and classification problems. It can learn and train complex data for prediction analysis. First, the self-help method is used to conduct sampling. N training sets are extracted from the original data set, and the size of each training set is about 2/3 of the original data set. Secondly, a classification regression tree was established for each training set to generate a forest composed of M decision trees. When each decision tree was growing, $n(n \leq N)$ attributes were randomly selected from all characteristic variables, and the optimal attribute was selected from the n attributes to branch in the internal node. Finally, the predicted results of the set M decision tree determine the category of the new sample by voting. About two-thirds of the data is extracted from each sample. See the flowchart below for details.

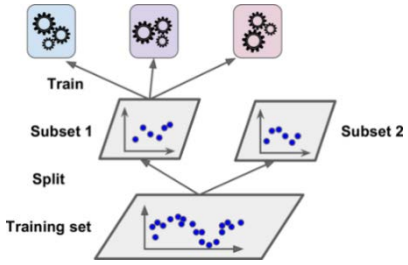

Fig.2 Rf Flowchart

The random forest model has the advantages of improving prediction accuracy, reducing overfitting, and simply processing a large amount of quantitative or qualitative data.

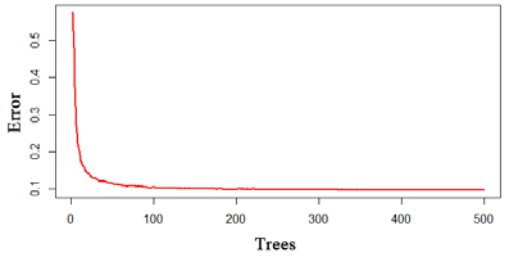

Fig.3 Relationship between Rf Decision Tree and Error

In this paper, the random forest algorithm is implemented by the R language. In order to test the number of decision trees corresponding to the optimal model, we first used 200 trees for testing, and then drew the relationship curve between the error rate and the number of decision trees (As shown in Figure 3).

When the number of trees is more than 300, the error rate tends to be stable. Therefore, we take $ntree = 300$ in this paper, and there will be no underfitting or overfitting phenomenon in the random forest.

3.2.2 Factor Screening Results

The random forest algorithm judges the contribution of each feature to each tree in the random forest and takes the average value, and then compares the contribution between the features. The classification trees constructed by the random forest may be different each time. In general, hundreds or even thousands of classification trees are randomly generated by the random forest, and then the trees with the highest degree of repetition are selected as the final result, so that the importance ordering of variables can be obtained quantitatively.

First, the out-of-pocket error e_i of each decision tree in the random forest is calculated according to the out-of-pocket data. Then, the j^{th} characteristic variable X_j of out-of-pocket data was randomly changed, and the new out-of-pocket error e_i^j was calculated. Finally, the importance $(V(X_j))$ of variable X_j is expressed as follows:

$$V(X^j) = \frac{1}{N}\sum_{i=1}^{N}(e_i^j - e_i)(3)$$

The greater the increase of out-of-pocket error caused by the change of variable X_j, the more the accuracy decreases, indicating that the variable is more important.

In this paper, 7 feature variables are selected in Section 3.1, but not every feature variable has a significant effect on job selection. Based on random forest algorithm, the importance of feature variables is extracted. According to its importance and the actual result (the response variable), the characteristic variable can be selected. In the process, a random forest filter is used for variable selection. The filter adopts the 10-fold cross validation method, which is repeated 10 times. See the appendix for the source code. The cumulative score rate of the final variables is shown in Figure 4:

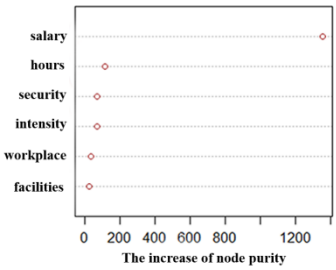

Fig.4 The Cumulative Score Rate of Each Factor in the Random Forest Model

After screening, six factors including salary, hours, security, intensity, workplace, and facilities are retained, in which the importance of the attribute of salary dominates. Job stability was removed, which is consistent with the reality because summer jobs are a shorter process compared to a normal job and students care much less about job stability in the short term, especially for a summer job.

4. Job Selection Model Based on Svm and Bp Neural Network

4.1 Principle of Svm

SVM was proposed in 1963 and developed rapidly after the 1990s, and a series of improved and extended algorithms were derived, including multi-classification SVM, least-square SVM (LS-SVM), Support Vector Regression (SVR) and so on[7-9]. Its outstanding advantages are as follows:

Based on the principle of structural risk minimization and the THEORY of the VC dimension, it has a good generalization ability, that is, the small error obtained from the limited training sample can ensure that the independent test set still has a small error.

The solution problem of support vector machines is a convex optimization problem, so the locally optimal solution is the global optimal solution.

The nonlinear problem can be transformed into linear problem-solving owing to the successful application of kernel function.

The maximization of classification interval makes the SVM algorithm more robust. Because of its outstanding advantages, SVM has been used by numerous researchers as a strong learning tool to solve many difficult problems in various fields, such as pattern recognition and regression estimation.

The key idea of SVM is a nonlinear transform defined by inner product function will low-dimensional input space transformation to a higher dimensional space. In the high dimensional space, it is able to search for a linear relationship between input and output variables, then regard the optimal linear regression hyperplane algorithm as solving a convex programming problem under constraint conditions, and finally SVM can get the global optimal solution. In this case, the method of linear learning machine can be applied in the feature space to solve the highly nonlinear regression problem in sample space.

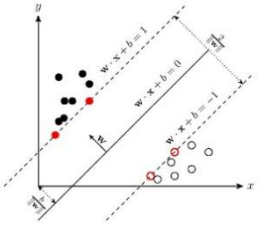

Fig.5 Principle of Svm Algorithm

In the SVM regression algorithm, the aim is to train hyperplane $y = w^T x + b$, and use $y = w^T x_n + b$ as the predictive value. In order to obtain a sparse solution, that is, calculating the hyperplane parameter w, b does not rely on all sample data but some data (such as in the SVM classification algorithm, support vector is defined).

According to practical problems, the job selection process of input variables of the support vector machine in this paper is as follows:

First, the 400 sets of questionnaire data collected were divided into 300 groups as training sets and 100 groups as test sets. Six factors (Salary, Hour, Security, Intensity, Workplace, Facilities) are selected as the input variable, and the actual job number is selected as the response variable. The specific structure diagram of the support vector machine is shown in Figure 8:

4.2 Principle of Bp Neural Network

The key idea of BP (Back Propagation) Neural Network Model to process information: Input signal X_i through the intermediate node (hidden) can be applied to the output nodes[10-12]. After a nonlinear transformation, it can produce output signals Y_k, and each sample in the training process contains the input vector X, the desired output T, and the deviation between network output value Y and desired output T. By adjusting connection intensity $W_{i,j}$ between the input nodes and hidden layer nodes, connection intensity $T_{j,k}$ between hidden nodes and output nodes and its threshold value, the error can be reduced along the gradient direction. After repeated training, the minimum error of the corresponding network parameters (weights and threshold) could be determined, and then the training stops. At this time, the trained neural network can process similar input information, and then output the information that has been nonlinearly transformed with the minimum error.

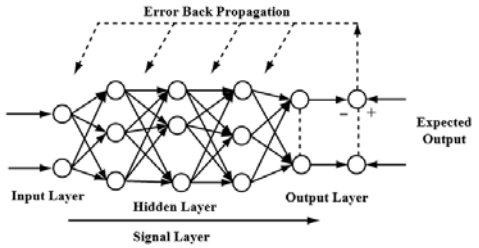

Fig.6 Principle of Bp Neural Network

4.3 Results

In order to more accurately grasp the choice of a summer job for high school students, this paper adopts two methods (SVM and BP) to establish the summer job selection model. The research data come from the 400 questionnaires. Six questions were designed in the questionnaire, mainly including salary, hour, security, intensity, workplace and facilities. In addition, there are 8 types of jobs (Private tutor, Waiter, Salesman, Courier, Fitness Coach, Tour-Guide, Research Assistant, Volunteer) among these high school students. In this paper, 6 factors and 8 types of the job were substituted into the SVM and BP models respectively as input variables and response variables. 300 sets of data were selected for training and 100 sets of data for testing. The final results are shown in the figure below:

As shown in Figure 7, in the SVM model, the true value is in good agreement with the predicted value, while under BP neural network, there will be a large deviation compared to SVM. Among them, the accuracy of SVM training set is 94%, test set 90%, BP training set 92%, and test set 88%. In conclusion, the prediction results of SVM are better than BP neural network. Finally, we select the SVM with higher accuracy and use all valid data as test sets to build a complete mathematical model.

Ten-fold cross validation was used to predict the 10 test sets. Figure 8 is a line diagram of the prediction results of the 10 training sets and the predicted results. Wherein, the training sets accuracy rate was greater than or equal to 94.0% but less than or equal to 96%; the results of the testing sets were greater than or equal to 91.0% but less than 95%, and the average accuracy was 93.5%. The accuracies of the test set were all lower than the accuracies of the training set except the eighth, and the accuracy showed a higher level. These results show that the SVM model has good stability and can be used to predict job selection accurately.

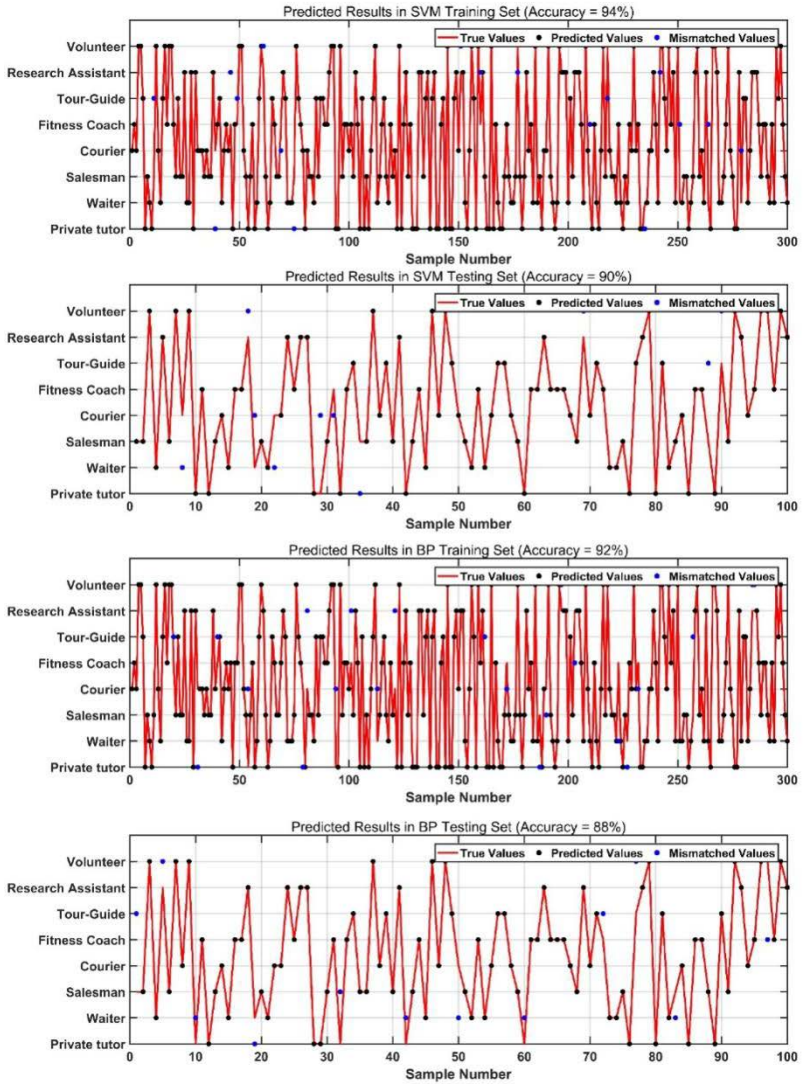

Fig.7 Job Selection Results in Bp & Svm

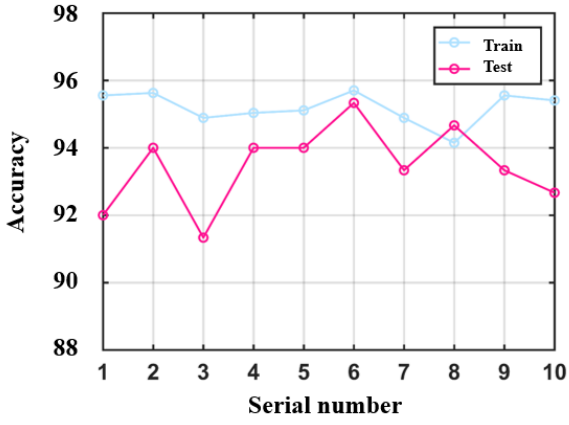

Fig.8 Accuracy of 10 Svm Tests

Take the Volunteer for example to illustrate the point. There are about 50 students choosing the volunteer as their summer job. These students have a low expectation on the salary and the facilities, and most of them are strongly desired to help others so that they do not care much about the benefits of the job itself but cherish the experience in the job. According to these characteristics, our model can accurately output the job selection of these students after the training. As for other job types, surely, there are special characteristics that our model can grasp and then obtain accurate

results. In conclusion, the job selection model we established in this paper can simulate the actual result, which means that the model can be widely promoted and applied.

5. Sensitivity Analysis

In the above analysis, we have attained the job selection model based on SVM trained with 400 questionnaire data. However, if the sample number is not enough, the model will be unstable. Here, we adjust the sample number and trial number to test the sensitivity of this model. The results are shown in Figure 9:

As shown in the figure above, the accuracy of the model changes with the change of experiment number and training sample number. When the number of samples is small and the number of experiments is small, the accuracy of the model is extremely low. With the increase of the number of samples and the number of experiments, the accuracy of the model increases gradually. When the number of samples is greater than 200, the accuracy reaches a stable value, which is about 90%. Therefore, the SVM-based work selection model needs to collect a large amount of sample data for training, otherwise, the results obtained will have a large deviation from the actual situation. In addition, if more impact factors are considered in the model, more samples will be needed.

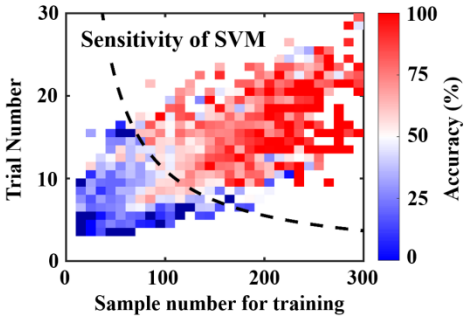

Fig.9 Sensitivity Test

6. Advantages and Disadvantages of Our Model

6.1 Advantages

1). The Random Forest Filter is Used to Screen Variables, Which Can Avoid the Objective Factors That Affect the Artificial Screening Index.

2). The Data Collected through Questionnaires Are More Scientific Than the Fabricated Data.

3). Compared with the Traditional Linear Model, Svm and Bp Neural Network Model Have Higher Accuracy and Can Solve High Dimensional Problems.

6.2 Disadvantages

1). Many Practical Factors Are Not Taken into Account and the Model is Relatively Simple.

2). The Data Used in This Paper May Be Not Enough and May Not Be Representative, and the General Conclusions Obtained by Using These Data May Have Deviated from Reality.

Acknowledgment

Sitong Jiang, Jingwen Ding, Yuxi Liu, Qingyi Hu, The 4 authors are ranked in no particular order

References

[1] Flieber, Ron. “Why a Teenage Bank Teller May Have the Best Summer Job. “ New York Times (2015).

[2] Jiansheng Wu, Hongde Liu, Xueye Duan, Yan Ding, Hongtao Wu, Yunfei Bai and Xiao Sun*. “Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature.” Bioinformatics 25.1(2009):30-35.

[3] Ma, Xin , et al. “Prediction of RNA‐binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature.” Proteins-structure Function & Bioinformatics 79.4(2015):1230-1239.

[4] Chen, Lu. “Prediction of hot spots in protein interfaces using a random forest model with hybrid features.” Protn Engineering, Design and Selection 25.3(2012):119-126.

[5] Chen, Lu. “Prediction of hot spots in protein interfaces using a random forest model with hybrid features.” Protn Engineering, Design and Selection 25.3(2012):119-126.

[6] Brokamp, Cole , et al. “Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model.” Environmental ence & Technology 52.7(2018):4173-4179.

[7] Cherkassky, Vladimir , and Y. Ma . “Practical selection of SVM parameters and noise estimation for SVM regression. “ Neural Networks 17.1(2004):113-126.

[8] Duan, Kai Bo , and S. S. Keerthi . “Which is the best multiclass SVM method? An empirical study. “ Proc.inte.works.mcs05 Seaside Ca Usa 3541(2005):278-285.

[9] Pal, M. , and G. M. Foody . “Feature Selection for Classification of Hyperspectral Data by SVM.” IEEE Transactions on Geoence & Remote Sensing 48.5(2010):2297-2307.

[10] Sadeghi, B. H. M. . “A BP-neural network predictor model for plastic injection molding process.” Journal of Materials Processing Technology 103.3(2000):411-416.

[11] A, Zhengxue Li , W. W. A , and Y. T. B . “Convergence of an online gradient method for feedforward neural networks with stochastic inputs.” Journal of Computational and Applied Mathematics 163. 1(2004):165-176.

[12] Zhi-Jie, Zhu . “Prediction of Coal and Gas Outburst Based on PCA-BP Neural Network.” China Safety ence Journal 23.4(2013):45-50.