Advances in Computer Science Research, volume 91

International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019)

Study of Employment Salary Forecast using KNN Algorithm

Junyu Zhang and Jinyong Cheng*

School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences) Jinan, China

*Corresponding author

Keywords: KNN, Prediction, Salary.

Abstract. In the market analysis of the 2018 national recruitment data and the employment situation in a city, it was found that the current college students have serious imbalances in their actual ability and employment salary expectations. In order to predict the salary of employment, doing this research. For the Java back-end engineer position, there are the following seven influencing factors: computer network scores, Java programming basic results, database principle results, java web scores, framework programming scores, Linux scores, education. The above seven factors affect the salary of the java back-end engineers, using different levels of salary as a marker to build a sample set, using the KNN algorithm to build a salary level prediction classifier. The conclusion is as follows: When K=7, the classifier has better prediction effect with an accuracy of 88.10%. This model has better predictive effect when the degree is undergraduate.

Introduction

With the development of social productivity, the living conditions and learning conditions of college students are constantly improving. The salary and social status are the college students' first choice of careers, leading to a mismatch in their ability and salary in selection. More and more people are blindly pursuing high academic qualifications to raise their salary levels. Under such social reality, the employment of college students has become more and more unbalanced. There is a common phenomenon in employers: there are many candidates with severely unsatisfactory positions in high-paying positions, and no one with low salary is concerned. This phenomenon has also caused many college students to lose their jobs: the general expectation salary of college students is higher than the salary that employers are willing to pay, which leads to employers preferring to hire some lower-educated personnel and not willing to hire college students. Under the education of four years or more, college students are reluctant to accept grassroots jobs with lower salaries and do not have a correct estimate of the salary they should receive for their actual level. These problems have directly led many college students to complain about employment difficulties and even unemployment. Based on the above-mentioned social status quo, it is decided to conduct actual research and analysis in related fields.

In order to ensure the authenticity of the results, 200,000 Java back-end development recruitment information of five recruitment websites national and 10,000 student employment salary and achievement information of Java back-end development in the city were obtained, after that how to handle big data sets should be considered[1,2].

Brief Description of KNN Algorithm

The full name of KNN is the K-Nearest Neighbor classification algorithm[3,4]which is one of the top ten algorithms for data mining[4] and be researched from 1967[5]. The core idea is to represent the sample according to the most one of its closest k points[6]. It belongs to a classification algorithm and is a non-parametric supervised learning algorithm[7,8]. The training set is represented by 8000 samples, which is expressed as S. As follows:

$$S=\begin{Bmatrix} X11\cdots X1j,\ Y1j+1 \\ X21\cdots X2j, Y2j+1 \\ \cdots \\ Xi1\cdots Xij, Yij+1 \end{Bmatrix}. \quad (1)$$

Among them, Xij represents seven attributes of computer network scores, basic results of Java programming, database principle scores, java web scores, framework programming scores, Linux scores, and academic qualifications. Yij indicates salary levels. According to the actual situation, 2000 yuan is defined as E level, 2000-4000 yuan is defined as D level, 4000-6000 yuan is defined as C level, 6000-8000 yuan is defined as B level, and 8000 yuan or more is defined as A level. Algorithm flow: input the prediction sample, calculate the distance between the test data and each training data, sort according to the increasing relationship of distance, select the K points with the smallest distance, determine the frequency of occurrence of the category of the first K points, and return the first K of the highest frequency category in the point is used as the predictive classification of the test data, that is, the level of the salary level that is the most common in the sum of the flag amounts, which is regarded as the category of K[9].

Flow Chart

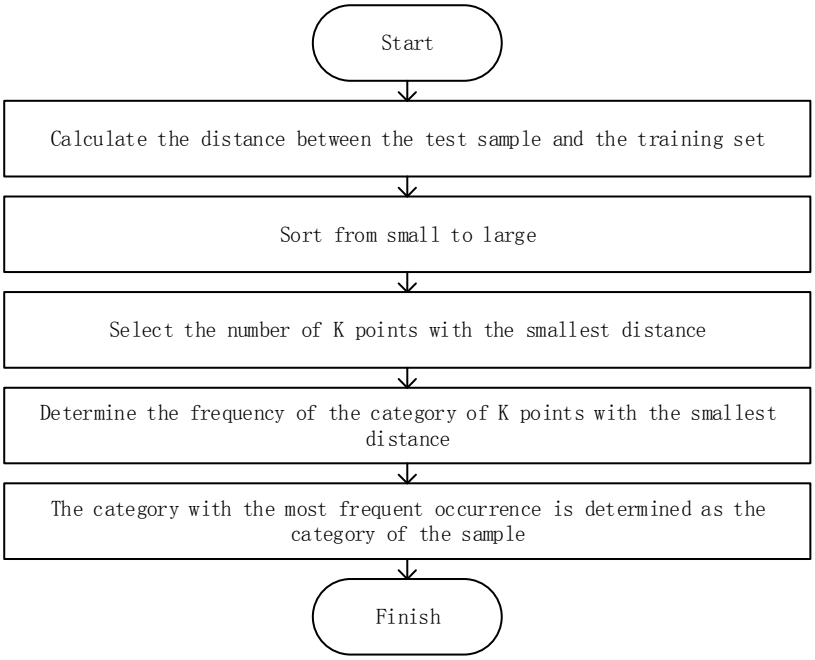

Figure 1. The Flow Charts of KNN to get Category

Selection of K Value in KNN Algorithm

The K-near number is the prediction of the sample when k adjacent points are selected for prediction. The selection of the k-value is directly related to the accuracy of the algorithm. If the value of K is too small, once the presence of noise components will have a greater impact on the prediction, for example, when the K value is 1, once the nearest point is noise, then there will be a deviation, K value Decreasing means that the overall model becomes complicated and over-fitting is easy to occur; if the value of K is too large, it is equivalent to prediction with a training instance in a larger neighborhood, and the approximate error of learning increases. At this point, the instance that is farther away from the input target point will also work on the prediction, causing the prediction to have an error. An increase in the K value means that the overall model becomes simple. If K=N, then all the instances are taken, that is, to take the most points in a certain category in the instance, there is no practical significance for the prediction; the value of K should be taken as odd as possible to ensure that the calculation is performed. The result will eventually produce a larger category, and if you take even numbers, it may produce an equal situation, which is not conducive to prediction. K's method: The commonly used method is to estimate the error rate of the classifier using the test set

starting from k=1. This process is repeated, each time K is incremented by 1, allowing the addition of a neighbor. Select the K that produces the smallest error rate. Generally, the value of k does not exceed 20, and the upper limit is the square root of n. As the data set increases, the value of K also increases.

Selection of Distance in KNN Algorithm

For the measurement of distance, commonly used are: Euclidean distance, cosine (cos), correlation (correlation), Manhattan distance (Manhattan distance) or other.

Experiment Analysis

Selection of K Value

This article uses cross-validation to determine the K-worth size. Cross Validation, also known as Rotation Estimation. The basic idea is to group the raw datasets in a sense, one part as a train set and the other part as a validation set or test set. First, use the training set to classify the classifier. Training, and then use the verification set to test the model obtained by training, as a performance indicator for evaluating the classifier.

After cross-validation, K is 5, 7, and 9 respectively, and the classification is accurate. As shown in the following table, the exact class is above 85%. When K=7, the accuracy is up to 88.10%. Therefore, the selection area K=7 was carried out in the experiment.

Table 1. Accuracy when K takes different values (%)

Education and K value	K=5	K=7	K=9
Bachelor degree	87.8	89.9	88.6
Master degree	84.4	86.3	85.7
Average accuracy rate	86.10	88.10	87.15

Analysis of Experimental Results

Under the undergraduate degree, the exact categories of predictive analysis are shown in the following table: When the salary level is A, the prediction accuracy rate is 93.3%, the false positive B-level probability is 4.5%, the false positive C-level probability is 1.3%, and the remaining false positive probability is less than 1.0%; When the salary level is B, the prediction accuracy rate is 85.2%, the false positive A-level probability is 5.2%, the false positive C-level probability is 6.1%, and the remaining false positive probability is less than 2.5%; When the salary level is C, the prediction accuracy is 81.1%, the false positive B-level probability is 8.2%, the false-reported D-level probability is 6.3%, the false-reported A-level probability is 3.1%, and the false-reported E-level probability is 1.3%; When the salary level is D, the prediction accuracy is 76.3%, the false positive C-level probability is 11.3%, the false-reported E-level probability is 6.3%, the false-reported B-level probability is 4.3%, and the false-reported A-level probability is 1.8%; When the salary level is E, the prediction accuracy is 73.1%, the false positive D-level probability is 13.8%, the false positive C-level probability is 8.7%, the false-reported B-level probability is 2.6%, and the false-reported A-level probability is 1.8%.

Table 2. Undergraduate degree forecast analysis accuracy rate (%)

	Class A	Class B	Class C	Class D	Class E
Class A	93.3	4.5	1.3	0.6	0.3
Class B	5.2	85.2	6.1	2.2	1.3
Class C	3.1	8.2	81.1	6.3	1.3
Class D	1.8	4.3	11.3	76.3	6.3
Class E	1.8	2.6	8.7	13.8	73.1

It can be seen from the above results that the accuracy rate is higher when the predicted salary is higher, and the probability of false positives is the lowest. There is still a certain degree of error and impact in the division of salary levels, but the level of salary can be predicted more accurately.

KNN-based Employment Analysis and Forecasting System

The Qt and KNN are written in Python language to realize the employment analysis and forecasting system. The recruitment information of computer related posts on several major recruitment websites of the Internet can be collected in real time, and relevant data analysis can be made. The employment salary forecast is made for the student's school grade and academic qualifications. Let students in school understand the demand for talents of enterprises and provide assistance for future employment. On the other hand, according to the trend of social needs, give future reference for students, and can provide a reference for the number of colleges and universities in terms of professional setting and enrollment. data.

Summary

This paper studies the method of employment salary forecast based on KNN algorithm, and draws the following conclusions: Seven variables such as computer network scores, Java programming basic scores, database principle scores, java web scores, framework programming scores, Linux scores, and academic qualifications constitute the characteristic attributes of the KNN classifier, and the salary level is selected as the marker amount. Cross-validation shows that when K=7, the accuracy of the KNN classifier is up to 88.10%, when K=5, the accuracy is 86.1%, and when K=9, it is 87.15%. In the analysis and prediction of the undergraduate degree, it can be concluded that the higher the salary level, the stronger the ability, the higher the prediction accuracy. Because the data used in this paper are all 2018 data, there is a certain error in the accuracy of the data, but it also reflects the current social status to some extent. In order to achieve higher accuracy, you can consider enriching the data set or improving the KNN algorithm. The application of data mining in colleges and universities in China is in a stage of rapid development. To a certain extent, the research on the factors affecting the employment of college graduates will gradually deepen and increase, and such education will be gradually improved.

Acknowledgement

This research was supported by the Project of Natural Science Foundation of Shandong Province (ZR2017MF056, ZR2018LF004, ZR201702220380), China.

References

[1] Deng Zhenyun, Zhu Xiaoshu, Cheng Debo. Efficient kNN classification algorithm for big data, Neurocomputing, 2016, pp.143-148.

[2] Shan S. Big data classification: problems and challenges in network intrusion prediction with machine learning, ACM Sigmetrics Performance Evaluation Review, 2014, pp.70-73.

[3] Ma, Bin, A New Kind of Parallel K_NN Network Public Opinion Classification Algorithm Based on Hadoop Platform, Applied Mechanics and Materials, 2014, pp. 644-650

[4] La Lei, Guo Qiao, Yang Dequan, Multiclass Boosting with Adaptive Group-based kNN and Its Application in Text Categorization, Mathematical Problems in Engineering, 2012, pp.473-486.

[5] Wu Xindong, V. Kumar, J.R. Quinlan, Top 10 algorithms in data mining, Knowledge and Information Systems, 2008.

[6] T. Cover, P. Art, Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 1967.

[7] V.N. Vapnik, The nature of statistical learning theory, New York: Springer, 2000.

[8] Tan Pangning, M. Steinbach, V. Kumar, Introduction to data mining, New Jersey: Addison Wesley, 2005.

[9] P. Harrington, Machine learning in action, Shelter Island, NY: Manning Publications Co, 2012.