Research on Automatic Estimation Method of College Students' Employment Rate Based on Internet Big Data Analysis

Xiao-hui Zhang(✉), Li-wei Jia, and Fa-wei Zhou

Henan Medical College, Zhengzhou 451191, Henan, China
lrl23201712@163.com

Abstract. In order to solve the problem of large error and inaccuracy in employment rate estimation, an automatic employment rate estimation method based on Internet big data analysis is proposed. This method can be divided into four steps: Firstly, the data integration model based on XML middleware is used to select the sample data of employment rate estimation. Secondly, the decision tree C4.5 algorithm is used to classify the attributes of the sample data. Thirdly, the improved KPCA algorithm is used to extract the feature vectors of employment information and calculate the distance between the forecasted samples and all samples. Fourthly, non-linear mapping method is used to transform employment structure data into corner data, and grey theory is used to establish employment rate estimation model. The results show that the average employment rate estimation error of this method is 4.81% lower than that of the statistical method based on support vector machine.

Keywords: Internet big data analysis · Employment rate · Estimation method

1 Introduction

Employment has always been one of the key issues of national concern. With the increasing number of college graduates in China, the employment situation has become more and more serious. Many graduates are facing the problem of unemployment. In this case, how to effectively estimate the employment rate of students has become a major problem to be solved in the field of education. The method of estimating the employment rate of students can estimate the future employment rate of students according to the employment data of historical graduates. At present, the employment rate of colleges and universities is mainly counted by manual method, but the process of this method is complex, which requires a lot of manpower and material resources, and the cost of statistics is relatively high. With the continuous development of intelligent technology, the use of intelligent statistical methods for college employment statistics [1]. Mainstream intelligent statistical methods of employment rate include the statistical methods of employment rate in Colleges and Universities Based on particle swarm optimization algorithm; Statistical method of university employment rate based on genetic algorithm; Statistical methods of college employment rate based on support vector machine algorithm, etc. Among them, the most commonly used method is based

G. Gui and L. Yun (Eds.): ADHIP 2019, LNICST 301, pp. 326–335, 2019.
https://doi.org/10.1007/978-3-030-36402-1_35

on support vector machine (SVM). However, due to the statistical error of this method, this study proposes a new method of automatic employment rate estimation based on large data analysis of the Internet. Firstly, it integrates the employment rate data information of colleges and universities, determines the sample data needed for employment rate estimation, classifies them, extracts the characteristic vector of the employment information of students to be predicted, calculates the distance between students to be predicted and all samples, and constructs the employment rate estimation model. Finally, the experiment proves that this method has less error and higher accuracy in estimating employment rate, and achieves satisfactory results [2].

2 Automatic Estimation Method of Employment Rate Based on Big Data Analysis

"Statistical data are the reflection of objective things' quantity. Quantitative understanding of statistics must be based on qualitative understanding of objective things. Statistical research is closely related to the quality of phenomena to study its quantity and reflect the quality of phenomena through quantity". Therefore, the concept of "employment" needs to be defined scientifically before the employment rate statistics. However, at present, the concept of "employment" has not been recognized and unified in China. According to the provisions of China's employment policy, labourers with labor capacity can combine with means of production through certain ways, and obtain certain remuneration or income from engaging in a legal social work. This is employment. Although the above definitions are different, they have common elements, namely, labourers with labor capacity, workers with the same means of production, and workers with remuneration or labor. Income. Accordingly, it can be strictly said that the employment rate refers to "the total number of employed graduates * 100% of the total number of employed graduates". Essentially, the employment rate is a statistical data and an important indicator to measure the comprehensive teaching ability of colleges and universities. The employment situation of college graduates is becoming more and more serious, and many professionals are in a saturated state, which makes it difficult for college graduates to find jobs [3]. Therefore, it is necessary to make statistics on the employment rate of different majors in universities and distribute educational resources rationally so as to guide students'majors effectively and ultimately promote the improvement of employment rate. The statistical method of employment rate in Colleges and universities has become a hot issue in the field of education, which has attracted the attention of many scholars.

2.1 Principle of Estimating Employment Rate

In the process of establishing the model of estimating the employment rate of students, the employment data of historical students are obtained, the employment samples of students are given and classified, and the data characteristics of students to be predicted are extracted, which are transformed into feature vectors, the distance between students to be predicted and all samples is calculated, and the model of estimating the employment rate of students is established. Detailed steps are shown in Fig. 1 below.

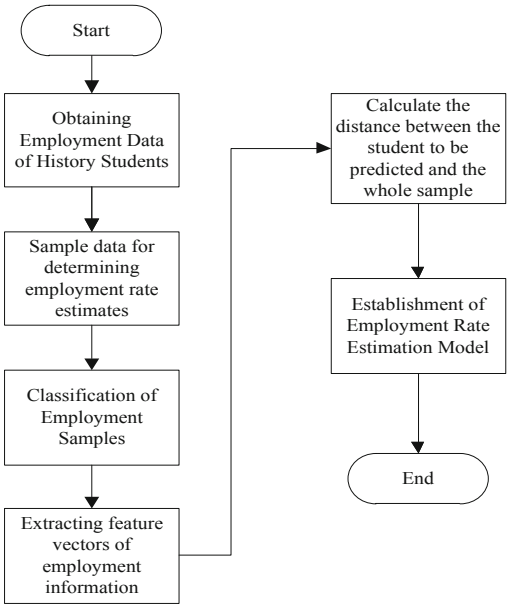

Fig. 1. Design flow of employment rate estimation method

2.2 Sample Data for Estimating Employment Rate

Due to the influence of phases, technology and other economic and human factors in the construction and implementation of data management system in Colleges and universities, a large number of employment-related data with different storage modes have been accumulated in the development process of colleges and universities. From simple file databases to complex network databases, they constitute heterogeneous data sources in Colleges and universities. Therefore, data integration is the only way to establish employment samples, and data integration needs a data integration model based on XML middleware to complete. The model is constructed through the following four steps:

(1) Establishing Data Integration Middleware

Data integration middleware is used to transfer data from data source to any target data source that needs data. External applications that want to access data can also access data from different heterogeneous data sources in a unified form through data integration middleware [4].

(2) Establishing relevant XML data model

In order to deal with all kinds of data sources in a unified way, the system must describe data from different data sources in a common mode. Generally speaking, the global data model of heterogeneous data source integration system must satisfy the following two points: First, it can describe various data formats, whether structured or semi-structured, whether it supports all query languages or simple text queries. Second,

it is easy to publish and exchange data. The integrated data can be easily published and exchanged in various formats.

(3) Establishing the mapping from specific data model to common data model

Mapping between specific storage mechanisms and public data models is required, and each data source must establish a mapping from itself to the XML public model. On the basis of completing the mapping process, it is also necessary to establish a data conversion program from the specific storage format of each data source to the XML format.

(4) Solving Semantic Heterogeneity

After resolving the problem of schema heterogeneity, the problem of semantic heterogeneity is very prominent. How to clean up the data, according to certain standards, error data, duplicate data and contradictory data from different systems. It is very important to unify appropriate transformation rules, which directly affects the integrity, consistency and sharing of data after integration.

This paper presents a heterogeneous data integration model based on XML middleware, as shown in Fig. 2. Using the method of heterogeneous data integration based on middleware, the middleware system is responsible for data access, query and coordination among heterogeneous data sources, and centralizes to provide high-level retrieval services for heterogeneous data sources. According to the method of middleware heterogeneous data integration, XML Schema is used to describe the schema information and global schema information of each local data source, and XPath query is used to query data based on global schema in a unified way [5].

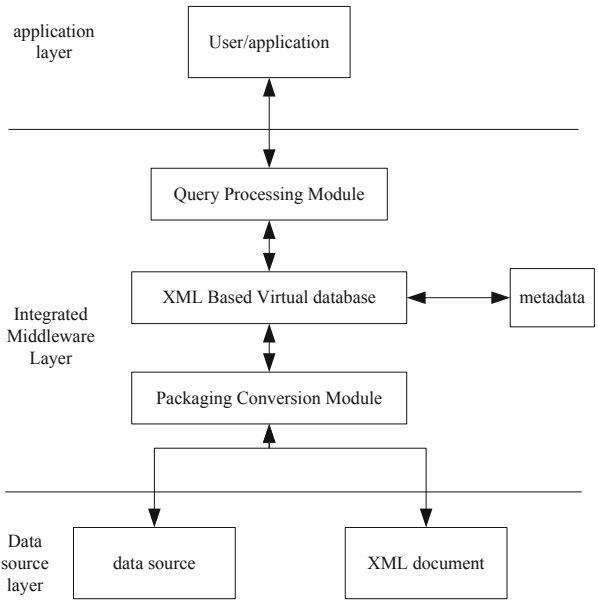

Fig. 2. Heterogeneous data integration model based on XML middleware

The whole model architecture is divided into three layers: data source layer, integration middleware layer and application layer. The integration middleware layer is the most important layer of the whole model and the key to realize heterogeneous data integration.

Data Source Layer: At the lowest level, it is a data provider, which should include various types of databases (relational database and object-oriented database), files and other information. Integration Middleware Layer: Coordinating database systems downwards, providing unified data schema and common interface for data access for applications accessing integrated data upwards, providing necessary data conversion functions or tools, transforming data into XML format, storing data into XML data space, and maintaining mapping relationship between XML data space and heterogeneous data sources [6].

Application layer: User interface layer, according to the specific application and user computing environment, adopt appropriate information access technology or application software. The application layer can access data in the application server layer of integrated data by Web browser or special client. Whether the application is in C/S mode or B/S mode, as long as the interface specification of the interface layer is followed, the underlying data sources can be operated effectively and transparently.

2.3 Data Attribute Classification

After the sample data is determined by the data integration model based on XML middleware mentioned above, it is necessary to classify different student characteristics information.

Decision Tree is a tree graph composed of a series of nodes and branches, in which branches are composed of nodes and sub-nodes. Nodes represent attributes that need to be considered in learning or decision-making, and different branches are composed of different attributes. By using the attribute value of an instance, from the root node of the decision tree to the leaf node, the case can be learned and the decision can be made. The final result of learning or decision-making is represented by leaf nodes [7].

Decision tree has the ability of multi-concept learning through analysis and induction of case sets. It is easy to use and has a wide range of applications. Decision tree algorithm is the best choice for classifying large-scale case data represented by unstructured attribute values. At present, ID3, C4.5, SLIQ and SPRINT are the most widely used decision tree learning algorithms. In this chapter, C4.5 algorithm is selected to complete the classification of sample data.

C4.5 algorithm constructs decision tree by top-down greedy algorithm for learning. The construction process starts with selecting the attributes to be tested at the root node, and then chooses the attributes to be tested at each node in the decision tree according to the criteria of maximum information gain and minimum entropy. The object set is divided by the test results. This process is repeated until the leaf nodes of a cotyledon are in the same class as the classification criteria.

2.4 Extraction of Employment Information Feature Vector

KPCA feature extraction method is a principal component analysis method based on the kernel transformation method. It maps the original data space to the high-dimensional space through the non-linear function (kernel function), and then processes the data of principal component analysis in the high-dimensional space to extract the linear and non-linear features of samples. The most important part of the method is to introduce the kernel function, which can simplify the calculation greatly by calculating the inner product in the high-bit space through the kernel function in the original space [8].

The purpose of KPCA algorithm is to preserve the feature information of samples as much as possible, and to simplify the representation of data as much as possible. The cumulative contribution rate is used to select feature vectors to reduce the dimension of feature space, which has a good effect on data dimension reduction. However, the whole process only considers the total feature information of samples without considering the category information of samples. Relevant literature also shows that the feature vectors extracted by the traditional KPCA method according to the cumulative contribution rate have no good effect on the classification of abnormal data, that is, only considering the main components of the matrix from a mathematical point of view without considering the classification of actual samples. So in this section, we improve the traditional KPCA algorithm, that is, to measure the class information by the degree of clustering within the class and the degree of dispersion between the classes of the feature vectors, which not only retains a good dimension reduction effect, but also is more conducive to the subsequent pattern classification. The improved KPCA algorithm runs as follows (Fig. 3):

According to the above process, assuming that P and O represent different types of data eigenvectors, formula (1) is used to calculate the distance d between the students to be predicted and the whole sample:

$$d = \frac{\sqrt{P+O}}{C_i \otimes X} \tag{1}$$

In the formula, C_i represents any category of student employment samples, and X represents unknown student employment samples [9].

2.5 Employment Rate Estimation Model Formation

In the process of establishing the employment rate estimation model, based on the distance between the students to be predicted and the whole sample obtained in Sect. 1.4, the employment structure data are transformed into corner data by using the non-linear mapping method, and the employment rate prediction model is established by integrating grey theory. Detailed steps are as follows:

Assuming that the original data of employment structure contains three dimensions and b represents two-dimensional corner data, it is based on the distance between the

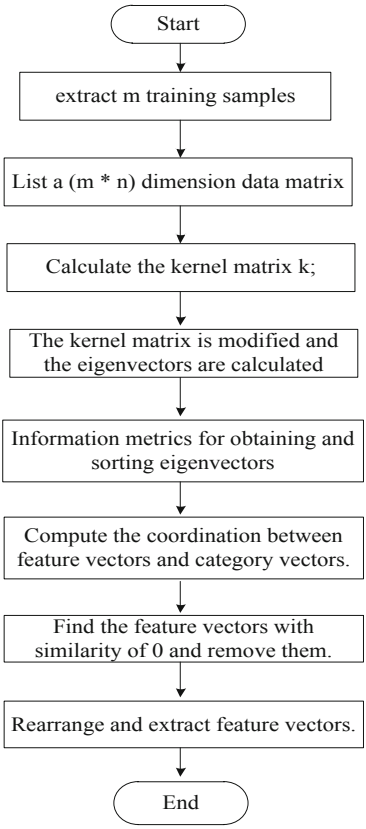

Fig. 3. Flow chart of improved KPCA algorithm

students to be predicted and the whole sample obtained in Sect. 1.4. Using the nonlinear mapping method, the employment structure data are transformed into corner data by formula (2):

$$\begin{cases} \alpha_3 = b \cdot d \\ \alpha_2 = \frac{d \cdot b}{\alpha_3} \end{cases} \tag{2}$$

In formula, α_2, α_3 represents the employment structure at different stages.

Assuming that e(0) represents the original employment forecasting time series, the grey theory is used to generate the series accumulatively. Formula (3) is used to obtain the accumulative time series of one-time employment forecasting:

$$e^{(1)} = \frac{e^{(0)} \otimes \left\{ e^{(1)}(1), e^{(1)}(2), \ldots, e^{(1)}(n) \right\}}{\alpha_3 \cdot \alpha_2} \tag{3}$$

The whitening equation of GM (1,1) model is obtained by approximating the change trend of a cumulative sequence with differential equation. The whitening equation of GM (1,1) model is expressed by formula (4):

$$f_{\mathrm{GM}(1,1)} = \frac{\{g,h\}}{W} + ge^{(1)} = h \tag{4}$$

In the formula, g represents the development coefficient of different employment stages, h represents the grey function of employment prediction. W represents a given sample of observed employment time series data.

Using Formula (5) to obtain the corresponding sequence of employment forecasting in different time periods:

$$\hat{e}^{(1)}(t+1) = \left[e^{(0)}(1) - \frac{g}{h}\right] \cdot q \tag{5}$$

In the formula, t stands for time; q stands for the certainty that the labor force has been employed.

Formula (6) is used to obtain the time response function of employment estimation in different stages:

$$\hat{e}^{(0)}(t+1) = [1-r]\left[e^{(0)}(1) - \frac{g}{h}\right] \cdot q + (\alpha_2, \alpha_3)\hat{e}^{(1)}(t+1) \tag{6}$$

In the formula, r represents the weighted coefficient of the overall employment forecast.

Assuming that j represents the mean square deviation of employment estimation and u represents the correction coefficient of prediction, the employment estimation model is established by formula (7).

$$y(t) = u(\lambda_1) \otimes \lambda_2 + \lambda_3 \ldots \frac{\delta_1(g)}{\delta_2(h)} \otimes (j,u) \tag{7}$$

In the formula, $y(t)$ represents the estimated value of t-time, δ_1, δ_2 represents the adjustment factor of constant, and λ_1, λ_2, λ_3 represents the transition process of each stage of employment estimation [10].

3 Experimental Analysis

The experimental data were collected from the statistical records of the employment situation of graduates in Liaoning Province from 2010 to 2018. In this study, the estimation method and the statistical method based on support vector machine are used to estimate the employment rate of graduates from 2010 to 2018, and the estimation errors of the two methods are detected. The results are shown in Fig. 4 below.

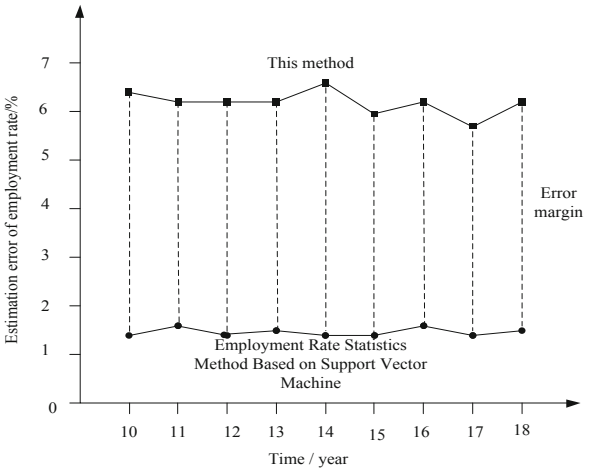

Fig. 4. Estimation error of two methods

Figure 4 shows that the average employment rate estimation error of this method is 1.34%, while that of the statistical method based on support vector machine is 6.15%. By comparing the two methods, we can see that the estimation of this method is more accurate, which makes up for the shortcomings of the statistical method of employment rate based on support vector machine.

4 Conclusion

To sum up, this study proposes an automatic employment rate estimation method based on large data analysis of the Internet, aiming at the problem of large error in employment rate estimation of statistical methods based on support vector machine. The simulation results show that the method solves the problems of the employment rate statistics method based on support vector machine, and lays a foundation for the development of employment guidance in Colleges and universities.

References

1. Wang, S.: Based on large data analysis of disadvantaged groups in colleges and universities graduates employment situation in recent years. China Univ. Students Career Guide **2**(17), 55–58 (2016)
2. Yang, H., Li, L.: Viewpoint on the employment rate of graduate students in agriculture and forestry. Heilongjiang Agric. Sci. **58**(7), 78–82 (2017)
3. Yan, Y., Deng, F.: Employment situation and characteristics for college graduates—discovery from data of 42 colleges and universities in Chengdu. J. Southwest JIAOTONG Univ. (Soc. Sci.) **17**(1), 1–8 (2016)
4. Yao, J., Chen, Y.: Employment status of graduates of food quality and safety specialty in West Anhui University. Anhui Agric. Sci. Bull. **22**(6), 177–179 (2016)

5. Liu, H.: Verification method for the maximal employment rate of university. J. Guangdong AIB Polytech. Coll. **33**(3), 48–53 (2017)
6. Zhao, X., Chen, X.: Clustering analysis based on graduates' employment quality pluralistic evaluation. J. Xinyang Agric. Coll. **26**(4), 149–151 (2016)
7. Zhang, R., Zhang, W., Wu, C.: Survey and analysis of employment status of graduates in mechanics—based on the statistic data of graduates in the Department of Engineering Mechanics at Dalian University of Technology during the past five years. China Univ. Students Career Guide **9**(11), 59–64 (2016)
8. Kong, P., Huang, C., Lu, Y.: Employment statistics and Countermeasures for aquaculture graduates—taking Guangdong Ocean University as an example. Agric. Dev. Equipments **36**(11), 38 (2016)
9. Lei, Y.: Further improve professional art school graduates' employment quality evaluation system –thinking based on MyCOS company on employment-based data analysis of music schools. Guide Sci. Educ. **12**(1), 167–169 (2016)
10. Huang, N.: On the employment situation and countermeasures of international Chinese education in application-oriented universities—a case study of Huangshan University. J. Huangshan Univ. **18**(1), 118–122 (2016)