Prediction of the Data Demand Model of University Teachers and Students

Xinping Lv

{15854585772@163.com}

No. 1 Daxue Road, Shandong Normal University, Changqing District, Jinan City, Shandong Province, China

Abstract. This survey mainly takes teachers and students in colleges and universities as the survey objects, investigates the data needs of teachers and students in colleges and universities, and draws conclusions and puts forward reasonable suggestions by combining the gradual regression model, Bernoulli distribution, short-term memory model, K-means clustering model and negative binomial model. On the basis of the research conclusions, this paper considers how to build a data service platform suitable for teachers and students in universities, so as to provide reference for the subsequent construction.

Keywords: Sampling stepwise regression,LSTM neural network,K-means clustering, Data requirement, Service platform, College teachers and students

1 Introduction

At present, many college teachers and students still face the problems of time-consuming and labor-intensive data processing, lack of relevant technology, and difficulty in seeking professional guidance. At the same time, the number of professionals in the big data industry is gradually increasing, and they are eager to improve their professional skills and obtain additional income through practice, but there is no suitable way. In order to solve this market difficulty and pain point, fill the market gap, promote the academic research level of college teachers and students, improve the efficiency of personnel management, and promote the scientific research progress of our country, we specially carried out this survey college teachers and students as the survey object, investigate the data needs of college students and students, put forward practical suggestions, and actively put into practice.

2 Main body

2.1 The user's processing demand prediction model is constructed based on sampling stepby-step regression

a.Model construction

First, This text study the relationship between different users' data processing needs and their individual attributes. The data processing requirements of different users are affected by the individual attributes of users. [1] In the construction of the model, the team took the major

PMBDA 2023, December 15-17, Nanjing, People's Republic of China
Copyright © 2024 EAI
DOI 10.4108/eai.15-12-2023.2345369

universities as a first-level indicator to calculate the individual attributes of teachers and students in the major universities.

The individual attributes of teachers and students in universities are denoated as the data processing demand rate of students in matrix X=(xij), denoated as yi, where i represents the identity of users; j indicates the user's major. The functional relationship between the data processing needs of different users (dependent variable) and the individual attributes of users (independent variable) is constructed as shown in the equation.

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{122} x_{i,122} + \varepsilon_i \quad \text{.....(1)}$$

Where α represents the intercept term, ε represents the residual term, and βj represents the regression coefficient of the JTH variable in the regression equation.

b.Model analysis

Next, building a variable selection model based on spik-slab sparse function and stepwise regression to solve equation (1). Since there are many independent variables in the equation and multicollinearity exists between variables, ordinary OLS cannot accurately estimate the influence coefficient of each variable on the dependent variable. [2] At the same time, we want to keep all the influencing factors in the model as much as possible, unless the variable has no impact on the user's data processing needs. Specifically, the solution steps are as follows:

Random sampling of independent variable X is performed based on Spike-slab sparse function to obtain a sample of X, denoting Xr;

Build a stepwise regression model and use Xr to regression the dependent variable y. The regression results were obtained at the specified significance level (significance level was 0.05).

Repeat steps 1 and 2 enough times to average all regression results that pass the test and obtain estimates of all parameters in the equation.

Here's how to do it. First, remember the coefficients of the argument X as the column vector $\beta=(\beta_j)$. Construct $\gamma=(\gamma_j)$ according to β, where, when $\beta_j=0$, make $\gamma_j=0$; Conversely, when $\beta_j\neq 0$, let $\gamma_j=1$. It is usually possible to construct 8 based on the Bernoulli distribution, as shown in equation (2) :

$$\gamma \sim p_j^{\gamma_i}(1-p_j)^{1-\gamma_i} \quad \text{.....(2)}$$

Where, for each stepwise regression model, when we want the expected number of independent variables to be m,Let $p_j=\frac{m}{n}$ and n represent the total number of independent variables. Then, according to the formula, a prior γ is obtained by sampling, and the corresponding variable x_j of $\beta_j\neq 0$ is selected according to $\gamma_j=1$ and recorded as set X_γ, which is the independent variable sample of the current stepwise regression

Secondly, a stepwise regression model is constructed to estimate the values of each parameter in model $y=f(X_\gamma)$. The fitting values of regression coefficients β and α corresponding to current X_γ are obtained. The stepwise regression method can ensure that the variables retained

in the model are independent variables that have significant influence on the dependent variables, and the multicollinearity is eliminated.

Finally, repeat sampling enough times to ensure model convergence. The regression parameter of the i step regression is $\varphi^{(i)} = (\alpha, \beta)^{(i)}$, so a series of fitting results $(\varphi^{(i)})$ can be obtained. Take the mean of all regression coefficients and use it as the final estimated coefficient of the independent variable, let $\overline{\varphi} = \sum_{i=1}^{N} \varphi^{(i)} / N$. Therefore, the relationship model between the inflow and outflow of public bicycles at dot i and the land use attribute of the dot is shown in equation (3) :

$$\hat{y}_i = \overline{\alpha} + \overline{\beta}_1 x_{i,1} + \overline{\beta}_2 x_{i,2} + \cdots + \overline{\beta}_{122} x_{i,122} \quad \text{.....(3)}$$

c.Result analysis

Finally, the regression coefficients of the respective variables (individual attributes) to the dependent variables (data processing requirements) were calculated, as shown in the following table1:

Table 1. Regression coefficient

	Gather Ingredients	preconditioning	analyze	visualization
sex	0.0046	0.0052	0.0048	0.0051
standing	0.2026	0.2068	0.2113	0.2097
profession	0.1926	0.1879	0.1903	0.1968
degree	0.2268	0.2158	0.2214	0.2256

The regression equation of dependent variable and independent variable can be obtained from the above table:

$$\begin{aligned} y1 &= 0.0046x1 + 0.2026x2 + 0.1926x3 + 0.2268x4 \\ y2 &= 0.0052x1 + 0.2068x2 + 0.1879x3 + 0.2158x4 \\ y3 &= 0.0048x1 + 0.2113x2 + 0.1903x3 + 0.2214x4 \\ y4 &= 0.0051x1 + 0.2097x2 + 0.1968x3 + 0.2256x4 \end{aligned} \quad \text{.....(4)}$$

According to the regression equation, the user's identity, the user's major and the user's highest degree are positively correlated with the data processing demand rate. Among them, the user's highest degree has the most significant impact on the data processing demand rate, followed by the user's identity, and finally the user's major. Higher degrees have a higher demand for data processing than lower degrees, and teachers often have a higher demand rate for data processing than students.

2.2 Based on LSTM neural network, the user data processing demand dilemma prediction model is constructed

a.Long-term Memory Model (LSTM)

User's data processing demand dilemma prediction model is composed of LSTM model to predict user's data processing demand dilemma under different individual attributes and data processing needs. LSTM model is a derivative of RNN model, which is a "processor" that judges whether the information is useful on the basis of RNN algorithm. [3] xt is the input user's personal attributes and data processing demand data, ht is the hidden layer, yt is the

user's data processing demand dilemma to predict the output result, which is jointly determined by the current user's input xt and the previous user's hidden layer ht-1. The hierarchy of hidden layer ht is expanded as shown in the figure. St represents the memory of the user's data processing needs dilemma at the user t, St = f(WSt-1 +UXt), W represents the weight of the input, U represents the weight of the input data at the moment, and V represents the weight of the output data.

At the first user, initialize input S0=0, randomly initialize W, U, V, and perform the following formula calculation: f(.) And g(.) Both are activation functions:

$$\begin{aligned} h_1 &= Ux_1 + WS_0 \\ S_1 &= f(h_1), f(.)\ sigmoid \\ o_1 &= g(VS_1), g(.) soft\max \end{aligned} \quad \text{.....(5)}$$

Model training is advanced, at this time S1, as the memory state of the user, participates in the prediction activity of the next user, and we can get:

$$\begin{aligned} h_2 &= Ux_2 + WS_1 \\ S_2 &= f(h_2) \\ o_2 &= g(VS_2) \end{aligned} \quad \text{.....(6)}$$

After continuous training, you can finally get:

$$\begin{aligned} h_t &= Ux_t + WS_{t-1} \\ S_t &= f(h_t) \\ o_t &= g(VS_t) \end{aligned} \quad \text{.....(7)}$$

The more specific structure of the LSTM is the cell.Each cell has three doors: input door, output door, and forget door. Oblivion gate decides to discard the user's data processing needs dilemma data information, it reads the ht-1 and xt information, and then enters the sigmoid function, output a value between 0 and 1:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \text{.....(8)}$$

ht−1 indicates the output of the previous cell, xt indicates the input of the current cell, and σ indicates the sigmod function.

The input gate controls the data processing requirements of the user t transmitted to the cell. The calculation formula is as follows:

$$\begin{aligned} i_i &= \sigma(W_i \cdot [h_{i-1}, x_i] + b_i) \\ VC_i &\tanh(W_c \cdot [h_{i-1}, x_i] + b_c) \end{aligned} \text{.....(9)}$$

The output gate determines the state vector of the user's data processing requirements that user tcell actually outputs.

$$\begin{aligned} Ct &= f_i \times C_{i-1} + i_i \times VC_i \\ h_i &= o_i \times \tanh(C_i) \\ o_i &= \sigma(W_o \cdot [h_{i-1}, x_i] + b_o) \end{aligned} \text{.....(10)}$$

Where, Wf, Wi, Wc, Wo represent the weight matrix, bf, bi, bc, bo are bias vectors of each network layer.

b.Model parameter setting

For the user's data processing demand dilemma prediction model for the user's prediction data set, the specific Settings are as follows: 1. The input layer of the neural network is the user's individual attributes and data processing needs. The output layer is the user's data processing needs dilemma. 2. The number of hidden layers of the model is set to 3 layers, and the activation function is set to RELU according to the data output characteristics. [4] To avoid overfitting problems, we define the rejection rate of nodes in each layer as 0.2. 3. The mean square error is used for the model loss function, and the root-mean-square error is used to predict the model accuracy. Based on the above parameter Settings, getting the LSTM model based on prediction set. Then, based on the actual data set, we take the individual attributes and data processing requirements of users in the actual data set as the input set, so as to predict the dilemma data of users' missing data processing requirements. [5]

c.Result analysis

Finally, the case test proves that the reliability of the model is high and the operation level is in line with the expectation. The following table2 shows some data of the strength test:

Table 2 Strength test data

	User data matching degree				
User	Time-consuming &labour-intensive	Lack of expertise	Excessive outlier	Tedious task	The collected data is not true
User A	0.8564	0.0032	0.0024	0.8426	0.0025
User B	0.0034	0.8962	0.0056	0.0009	0.0012
User C	0.0026	0.0103	0.9024	0.0203	0.2304
User D	0.5654	0.6958	0.0016	0.5963	0.0015

As shown in the figure, through the input of the user's individual attributes and the user's data processing needs, and then according to the previous data memory set, we can judge: 1. User A has the difficulty of time-consuming and labor-intensive data processing and cumbersome tasks, so he can seek the help of partners to jointly complete the data processing task. 2. User B has the dilemma of lack of professional knowledge in data processing, so improving the learning of professional knowledge is the primary solution for user B; 3. [6] User C has the problem that the outlier value is too high in the process of data processing, so he can conduct a new data collection and compare the data obtained last time. 4. User D has the same level of time and effort in data processing, lack of relevant knowledge and other problems, which indicates that the model memory set still needs to be improved.

2.3Based on K-means clustering and negative binomial model, the user platform usage intention prediction model is constructed

According to the survey results, the user's intention to use the data service platform will be affected by three factors: user's individual attributes, data processing needs and data processing needs dilemma. [7] Therefore, this paper takes these three indicators as the first-

level indicators that affect the willingness to use data service platforms, and studies their impact on people's willingness to use data service platforms. Among them, 4 second-level indicators representing individual attributes, 5 second-level indicators representing data processing needs, and 6 second-level indicators representing the dilemma of data processing needs are selected, as shown in the following table3:

Table 3 Primary and secondary indicators

Primary index	Secondary index
Individual attribute	Gender, identity, major, degree
Data processing requirement	Data acquisition, data preprocessing, data analysis, data visualization, data mining
Data processing requirements dilemma	Time-consuming and labor-intensive, lack of professional knowledge, excessive data outliers..

a.K-means clustering

In the process of applying this method, University teachers and students' data set X, each sample is composed of feature vectors of m attributes, that is,X={x1 ,x2 , … ,xn }. When the intention level of data service platform is k, n samples can be divided into k subsets C,C={C1,C2,…,CK}, the calculation steps of K-means clustering method are as follows: 1. Initialization. k sample data h（0） in the dataset X university teachers and students were randomly selected as the initial clustering center of k class intention level. h(0)={h1(0),h2(0),…,hk(0)|hi(0)∈X,i=1,2,…,k}, h(0) is the cluster center that uniquely applies the intention level to a certain user.

Cluster the samples. Calculate the distance of (n-k) samples to the cluster center:

$$\mathrm{d}(x_j - \mathrm{h}_l^{(0)}) = \sum_{w=1}^{m} (x_{wj} - h_{wl}^{(0)})^2 \quad \text{.....(11)}$$

Where, Xwj is the WTH attribute value of sample xj; hwl（0） is the WTH attribute value of sample hl(0). xj∈X, hl(0)∈h(0); Each sample is divided into the intention level category of the cluster center closest to it, and the clustering result is obtained: C={C1 （0） ,C2 （0） ,…,CK （0） }.

Calculate new clustering centers. For the initial clustering result C（0）, the mean value of the samples contained in the current use intention level is calculated, and the new clustering center h(1) ={h1(1),h2(1),…,hk(1)}.

$$h_i^{(1)} = \frac{1}{|C_i^{(0)}|} \sum_{x \in C_i^{(0)}} x \quad \text{.....(12)}$$

Where, |Ci(0)| is the number of samples contained in the initial clustering results of Class i using intention level, i= 1,2,... , k.

Iterative optimization. K-means algorithm uses the sum of squared error criterion function to evaluate the clustering performance. The sum of squared error E of the final clustering result can be calculated by the following formula: $E = \sum_{i=1}^{k} \sum_{x \in C_i} || x - \mu_i ||^2$,

where $\mu_i = \frac{1}{|C_i|}\sum_{c \in C_i} x$ is the Ci mean vector of the final clustering result. To a certain extent, E reflects the tightness of samples under different categories of use intention level around its mean vector, The smaller the value of E, the higher the similarity of samples under each intention level category.

In order to minimize the square error, the iterative method is usually adopted: repeat the above steps, after iteration round t, if the corresponding willingness level of each sample in the dataset X is the same as the iteration result of round (t-1), then the iteration is stopped, so that:

$$C = C^{(t)} \text{.....(13)}$$

Where, $C^{(t)}$ is the clustering result obtained by the t iteration.

b.Model establishment

Next, a multi-factor influence model is constructed to study the relationship between the intention to use data service platform and each influence factor. [8] Since the dependent variable (that is, the intention to use the data service platform) is the discrete counting data, the negative binomial regression model is considered to construct the influencing factor model. First, the negative binomial regression model is used to construct the functional relationship between the intention to use data service platforms and different influencing factors, as shown in the equation:

$$\ln(y_{i,t}) = \alpha + \beta_1 x_{i,t,1} + \beta_2 x_{i,t,2} + \beta_3 x_{i,t,3} + \cdots + \beta_{38} x_{i,t,38} + \varepsilon_{i,t} \text{.....(14)}$$

Where, yi,t represents user i's intention to use the data service platform under the t attribute, i represents the user, and t represents the user attribute; xi,t,j indicates the JTH factor affecting the demand for shared bicycles; βj represents the regression coefficient of the JTH independent variable in the regression equation. α represents the intercept term; εi,t denotes the residual term.

c.Result analysis

Based on the questionnaire data and calculation results, it is stipulated that the user's intention to use the data service platform k is divided into 6 levels, which are represented by numbers 1-6, in which level 1 is the strongest intention to use, which decreases in turn. Grades 1-2 are defined as strong use intention, grades 3-4 as medium use intention, and grades 5-6 as low use intention.

As can be seen , among individual attributes, the user's degree is the main factor affecting the level of use intention. Users with data analysis and data visualization requirements are more willing to use data service platforms; When data processing tasks are too time-consuming and labor-intensive, users are willing to use the platform's website to solve existing problems.

3 Conclusions

A. The lack of unified data standards, no unified data standards have been formed, resulting in poor data quality, and the lack of effective data acquisition methods.

B. The experience of data service platform and the security and stability of the platform are the main factors that affect the participation of teachers and students in the data service platform.

C. Based on variance filtering and Logstic regression, we analyzed the demand characteristics of different groups, designed a personalized participation platform plan for different groups, and combined with the actual situation of individuals to achieve accurate promotion of activities.

D. It is a good way to improve the experience of the data service platform to integrate the ordinary big data service platform with task publishing and order receiving.

E. Big data service platform market development potential is huge.

References

[1] Chen S, Zhao J. The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication[J]. IEEE communications magazine, 2014, 52(5): 36-43.

[2] Ghavimi F, Chen H H. M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications[J]. IEEE Communications Surveys & Tutorials, 2014, 17(2): 525-549.

[3] Bender P, Black P, Grob M, et al. CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users[J]. IEEE Communications magazine, 2000, 38(7): 70-77.

[4] Wu Gang, Chen Guixiang. Operational mechanism of Big Data governance in universities: Functions, problems and Countermeasures [J]. University Education Science,2018,0(6):34-3866. (in Chinese

[5] Yang Yang. Planning, Design and Implementation of University Big Data Platform [J]. Journal of Shenzhen University: Science and Technology Edition, 2019,37(S01):146-149. (in Chinese

[6] Hu Shui-wen, Jing Zhou, WANG Huijun. Research on Key elements and optimization path of Big Data governance System in Chinese universities -- Based on the research perspective of DEMATEL-ISM [J]. Electrochemical education research, 2022 lancet (11) : 38-44 + 52. DOI: 10.13811 / j.carol carroll nki. Investigate. 2022.11.005.

[7] Tian Yukun. Research on the Development of University Education Management Empowered by Big Data -- Review of Research on University Education Management Based on Big Data [J]. Chinese Journal of Science and Technology, 2019,17(11):1318.

[8] Jones S, Ball A, Ekmekcioglu C. The data audit framework: a first step in the data management challenge［J］. International Journal of Digital Curation, 2008, 3 (2): 112－120.