Retraction

Retraction: Application of Keras neural network in the era of big data (J. Phys.: Conf. Ser. 2083 042090)

Published 9 September 2022

This article has been retracted by IOP Publishing following an allegation that raises concerns this article may have been created, manipulated, and/or sold by a commercial entity. In addition, IOP Publishing has seen no evidence that reliable peer review was conducted on this article, despite the clear standards expected of and communicated to conference organisers.

The authors of the article have been given opportunity to present evidence that they were the original and genuine creators of the work, however at the time of publication of this notice, IOP Publishing has not received any response. IOP Publishing has analysed the article and agrees there are enough indicators to cause serious doubts over the legitimacy of the work and agree this article should be retracted. The authors are encouraged to contact IOP Publishing Limited if they have any comments on this retraction.

Retraction published: 9 September 2022
Application of Keras neural network in the era of big data

Junqi Wang*
Tianjin University of Technology, Tianjin, China

*Corresponding author e-mail: wangjunqi@tjut.edu.cn

Abstract. The experiment uses crawler tools to obtain data, and the data is preprocessed to find missing values and eliminate invalid data, meanwhile, the model is constructed by information entropy and ID3 algorithm so as to select the desired amount of features, and then basic modeling and data filtering is performed to train and evaluate the model for the first time, finally, in order to get a more ideal model, this experiment The optimal model is obtained by changing the number of hidden layers and neurons of the neural network to build a high-level neural network API neural network model written by pure python - Keras neural network model. The results show that when the model defines a 2-layer neural network and the number of neurons in the hidden layer is fourteen, the accuracy of the model is the highest, and the accuracy of the test set is as high as ninety-one percent.

1. Introduction

Big data is available for computing very large data sets, and we can discover useful patterns and trends in the data by examining a process called data mining of big data sets [1]. Data mining is a step in the knowledge discovery process that includes important techniques for examining big data [2], big data sets are used to identify relationships, patterns and trends by sifting through large amounts of complex data with the ability to accumulate deeper insights from multiple variables and values rather than relying on small studies using descriptive, predictive or experimental designs, using advanced data modeling techniques, algorithms, statistical methods and mathematical techniques, much can be learned about the income classification of big data practitioners, and we can place data in user-friendly but powerful pivot tables and multidimensional charts. The data can be checked for anomalies and correlations. It can be classified, clustered, sliced, described, and subjected to iterative analysis and predictive model development [3].

The shortage of big data talents has become one of the factors affecting the rapid development and application of big data. It is necessary to cultivate big data talents, and many colleges and universities have opened big data courses, big data majors, etc. to deliver big data talents to the society. This article is based on the employment statistics of big data in 2020, I studied the employment data of big data practitioners in four first-tier cities, such as Beijing, Shanghai, Guangzhou and Shenzhen, and four second-tier cities, such as Nanjing, Xi’an, Zhengzhou and Hangzhou. Hopefully, a good revenue classification model will be obtained from it.

Data mining techniques provide a wide range of successful applications in many fields. Among the available machine learning classification models [4], consider first the decision tree classification model, where we first use data mining techniques to extract previously unknown and potentially useful implicit information from a set of data in order to find novel and useful patterns that might otherwise remain
unknown [5]. Important data mining techniques include: classification, clustering, regression, decision trees, outlier detection, neural networks, and statistical techniques. The definition of decision tree approach suggests that the construction of decision trees by decision tree algorithms can reveal the main factors that influence high or low income [6] [7].

Because the accuracy of traditional machine learning algorithms is relatively lower than that of neural network algorithm models, a Keras-based neural network algorithm model is constructed through existing data to achieve accurate classification.

2. Purpose and framework of the study

2.1. Purpose of the Study

Income status is a hot topic that never gets old, and the income of people engaged in big data business is no exception. Among many influencing factors, this paper obtains the main factors affecting employment income by building a decision tree classification model, and to make the model more accurate in classification, a more intelligent neural network model is used to build the classifier.

2.2. Schematic diagram of the research design

Schematic diagram of the research and experimental design framework as shown in Figure 1.

![Figure 1. Schematic diagram of the research and experimental design framework](image)

3. Design and implementation of the study

3.1. Crawling data and data cleaning

The author browsed the relevant employment websites and recruitment website to crawl the employment information and job information of big data talents in 2020, and the obtained data to find the missing values as well as to perform the cleaning of these data.

3.2. Data Analysis

Based on the employment and recruitment data of big data positions in 2020, the employment and recruitment data of big data positions in four first-tier cities, such as Beijing, Shanghai, Guangzhou and Shenzhen, and four second-tier cities, such as Nanjing, Xi'an, Zhengzhou and Hangzhou, a total of eight cities, were studied, and according to statistics, the number of big data position recruitment companies in first-tier cities is much higher than that in second-tier cities, as shown in Figure 2.
3.3. Feature selection

3.3.1. Overview of the method of feature extraction. By analyzing the employment situation of big data talents in 2020, the features that influence employment income are crawled: ① job position; ② job seniority; ③ education; ④ company size; ⑤ work city; ⑥ city one or two lines; ⑦ south and north regions, among which the feature ④ company size is excluded because the feature is difficult to examine. The main features are selected among the remaining 6 features to build the model, and the screening method is to construct a decision tree model by the ID3 algorithm. The ID3 algorithm involves the following definitions of concepts:

Information entropy: \[ H(x) = - \sum x p(x) \log p(x) \]

Information entropy is a measure of uncertainty, which reflects the degree of uncertainty through the amount of information [8].
Conditional Entropy: \( H(Y|X) = -\sum_x p(x) \log p(Y|x) \)

Information Gain: \( IG(Y|X) = H(Y) - H(Y|X) \)

The ID3 algorithm selects the tree nodes based on the information gain. The greater the information gain, the better the selectivity of this feature, i.e., the greater the contribution of this feature to the overall event.

3.3.2. Model construction and analysis. A. Model construction.

Data set: ① job position; ② job qualification; ③ education; ⑤ working city; ⑥ city first and second line; ⑦ south and north region, etc. are used as feature data.

Data set labels: income is used as the label, and the label has 3 values, ["Low", "Medium", "High"], and the criteria for classification are annual income > 200,000 as "High", "Medium" for 100-200,000, and "Low" for <100,000.

B. Model Building

model=tree.DecisionTreeClassifier(criterion='entropy')

C. Parameter setting, looking for key features

When setting the depth of the tree max_depth=4 and max_depth=6, the features that appear in the decision tree are: ② ③ ⑤ ⑥ ⑦.

The depth of the tree max_depth is not limited, and no feature ① job appears in the decision tree, therefore, this feature is discarded in this experiment. In several experiments, the feature appearing in the root node is ③ education, so education is the key factor affecting income.

3.4. Model building and scoring

New dataset: ② ③ ⑤ ⑥ ⑦ five features as feature data. The model is trained using the training set and the labels of the training set, and the model still uses the decision tree model that comes with Sklearn:

model=tree.DecisionTreeClassifier(criterion='entropy')

Using 20% of the data as the test set, the scores of the model were obtained:

Test set score 0.8493975903614458.

The training set scores 0.8874622356495468.

From the results, it can be seen that the present decision tree model results are more ideal with an accuracy of more than 80%. In order to get a more ideal model, a neural network model is established.

3.5. Model improvement - based on Keras neural network model

Keras is a pure python-based deep learning framework based on theano/tensorflow. keras is a high-level neural network API that can quickly translate your "ideas" into results.

The experiments were conducted using the neural network model shown in Figure 4.

![Figure 4. Neural network diagram](image-url)
In this experiment, the better model is obtained by changing the number of hidden layers and neurons of the neural network.

A. Model Construction

Import tensorflow as tf

model.add(tf.keras.layers.Dense(i,input_shape=(j,),
activation='relu'))
model.add(tf.keras.layers.Dense(3,activation='softmax'))

B. Find the optimal number of network layers and neurons

A 2-layer neural network is built. In Figure 5 shows the accuracy of the practice set and test set for model training.

![Accuracy diagram](image)

Figure 5. Accuracy diagram

From Figure 5, it can be seen that the model has the highest accuracy when the number of neurons in the hidden layer is 14, and the accuracy of the test set is as high as 91%. Next, the number of layers of the neural network is changed, and after the experimental study, it is found that the difference between the experimental results of multi-layer hidden layer and single-layer hidden layer is not much, so the final model is defined as a 2-layer neural network with the number of neurons in the hidden layer of 14.

4. Conclusion

In this paper, a new neural network model is built to capture the key factors affecting the income of the employed people, and it is found that improving education is the key to improve the income of big data practitioners with the help of data mining technology. By analyzing the big data recruitment information data in 2020, it is found that there are far more big data recruitment companies in first-tier cities than in second-tier cities, while the education required for big data positions is mainly bachelor's degree. Among them, a Keras neural network model was established for the processed data to achieve an accuracy rate of more than 90% for income classification, which provides great help for the majority of practitioners and potential employment. However, the model is not yet able to account for the relationship between employment income and characteristics such as company size and age of the employed, which deserves further research.

References

[1] Larose & Larose, 2015 D.T. Larose, C.D. Larose Data mining and predictive analytics (2nd ed.), Wiley & Sons, Inc, Hoboken, NJ (2015).
[2] Jothi et al., 2015 N. Jothi, N.A. Rashid, W. Husain Data mining in healthcare: A review Procedia Computer Science, 72 (2015), pp. 306-313.
[3] Caryn Scheinberg Andrews, Julie Vosit Steller, Abigail Kra Friedman, Development of collaborative family nurse practitioner education using physician preceptors: A focused critical ethnographic study, Nurse Education Today, Volume 107, 2021, 105110, ISSN 0260-
[4] Andreas Muller. Python machine learning fundamentals tutorial [M]. People's Post and Telecommunications Publishing House, 2018.

[5] I.H. Witten, E. Frank, H. Ma. Data Mining: Practical Machine Learning Tools and Techniques (Third ed.), Morgan Kaufmann, San Francisco (2011).

[6] Yue Genxia. Simulation of filling and classifying medical big data based on decision tree algorithm [J]. Computer Simulation 2021. 38(1): 451-454.

[7] Miriam Andrejiova, Anna Grincova, Daniela Marasova, Identification with machine learning techniques of a classification model for the degree of damage to rubber-textile conveyor belts with the aim to achieve sustainability, Engineering Failure Analysis, Volume 127, 2021,

[8] Mao Yimin. A parallel association rule incremental mining algorithm based on information entropy and genetic algorithm [J]. Journal of Communication, 2021. 3.