ISBN: 978-1-948012-08-9

Contents List available at VOLKSON PRESS

International Symposium on Computer Science and Artificial Intelligence(ISCSAI)

Overdue Prediction of Bank Loans Based on Deep Neural Network

Li Xin[ab], Sun Guozi[abcd] and Li Huakang[abcd*]

[a]Jiangsu Key Lab of Big Data and Security and Intelligent Processing
[b]Nanjing University of Posts and Telecommunications, Nanjing, 210023, China
[c]Collaborative Innovation Center for Economics crime investigation and prevention technology, China
[d]State Key Laboratory of Mathematical Engineering and Advanced Computing, China NO.9, Wenyuan Road, Nanjing, China
[*]Corresponding Author: huakanglee@njupt.edu.cn

ARTICLE DETAILS

Article History:

Received 02 october 2017
Accepted 06 october 2017
Available online 11 october 2017

Keywords:

Overdue Prediction; Bank Loans; DNN;

ABSTRACT

With the development of information technology, the application of big data in financial aspects becomes more and more deepening. However, in the aspect of bank loans, the accuracy of traditional user loan risk prediction models, such as KNN, Bayesian, are not benefit from the data growth.This paper proposes to use DNN algorithm to forecast the risk of user loan based on the difficulties of current overdue prediction and the excellent learning ability of DNN. This article uses user basic information, bank records, user browsing behavior, credit card billing records, and loan time information to evaluate whether users are delinquent. Firstly, this paper record bank records according to the transaction type, respectively, to generate income and spending data. Secondly, to sum the user browsing behavior also, and to record the average of credit card bill. In addition, in order to reduce the effect of eigenvalue size on the result, all characteristics are standardized. Finally, users who lack user information are discarded and the above fields are spliced. The spliced fields are the basic input for DNN. From the experimental results, DNN algorithm increase over 6% prediction than kNN, Bayes algorithm.

1. Introduction

1.1. Background

From "Financial Statistics Report in 2016", it shows that annual RMB loans have increased 12.65 trillion yuan, a year-on-year increase of 925.7 billion yuan than last year, and as the end of the first quarter of 2016, the balance of non-performing loans of commercial banks was about 1.4 trillion yuan, an increase of 117.7 billion yuan from the previous quarter. Since there are many features in our dataset, kNN or Bayesian cannot behave well in learning these features. Thus, DNN is introduced to deal with this issue. With DNN's strong distributed storage and learning ability., each feature can be well connected.

1.2 Research Method

Further, strong ability of fault tolerance to noise neuron is also one of the most important advantages of DNN. DNN model helps the bank to judge whether the user has the qualification of the loan. Finally, the advantages and disadvantages of the algorithm are compared, then we can establish the relevant algorithm model to train and test the data, comparing the accuracy of each model, summarizing the advantages and disadvantages of the DNN model, and how to improve it.

1.3 Relative Work

In business fields, there are two main methods used: single classification method and combination classification method.

1.3.1 Single Classification Method

The Bayesian method was proposed by some researcher in 1992 [1]. This method is widely used in the industry, such as Bayesian statistics in genetics, Vehicle Classification in Video. A researcher also presents a Bayesian method of testing possibly non-nested restrictions in a multivariate linear mode [2]. This algorithm is also applied to sleep discrimination, spam filtering, context research, digital scanning and so on.

The kNN (k-NearestNeighbor) algorithm is also one of the typical classification algorithms. Pusou, GW Mineau proposed the use of kNN algorithm for text classification, the paper also compared the kNN algorithm with Bayesian algorithm, pointed out that kNN algorithm in this area has a higher accuracy. In addition, based on the kNN algorithm of the anomaly detection method by another researcher [3].

Neural network is another method which is widely used. A group researcher from different study have proposed credit card fraud detection with a neural network [4-7].

1.3.2 Combination Classification Method

In other study, the researcher also established the metacarage classification system, which was gradually improved on this basis, and they came up with the conclusion that combination classification did better in prediction than single classification [8,9].

1.4 Paper Structure

This paper is divided into the following parts. The Section 2 introduces the DNN model. Section 3 is the experimental result, which mainly presents the data of dataset, evaluation index and calculation platform. Section 4 is the summary of the article.

2 DNN MODEL

2.2 Principal of DNN

According to the location of different layers, the neural network layer within DNN can be divided into three types: input layer, hidden layer and output layer. As shown in Figure 1, generally the first layer is the output layer, and the last layer is the output layer, while the middle layer is the hidden layer.

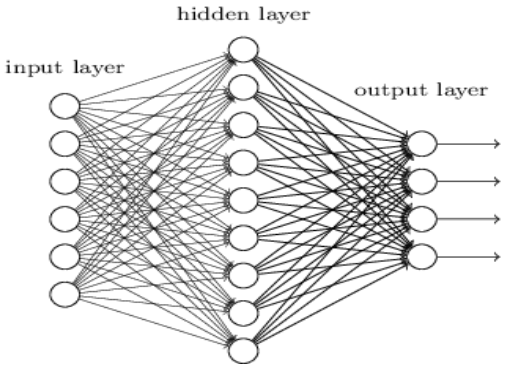

Figure 1: The Model of DNN

There is a total connection between the middle and the layers of DNN, which means that any neuron in the i layer must be connected to any one of the neurons in the i+1 layer. And the output layer neurons can also be more than one output, you can have multiple output, this model can be flexibly applied to classification of regression. DNN learns a linear relationship between output and input and gets intermediate outputs:

$$z = \sum_{i=1}^{m} w_i x_i + b$$

And then activation of the activation function, the activation function can use the Sigmoid function or the subsequent tanx, softmax, and ReLU [10].

2.3 DNN Algorithm for User Loan Risk Analysis

2.3.1 Data preprocessing

The sample data includes the user's basic attributes, bank records, user browsing behavior, credit card billing records, loan hours, and records of whether these customers have overdue behaviors [11]. In the user's basic attributes, there are 6 dimensions, which are user id, gender, occupation, education degree, marital status, household registration type. In the bank records there are 5 dimensions, which are user id, time stamp, transaction type, transaction amount and wage income mark.

We select the user id, the transaction type, the transaction amount of the three columns, group the transaction types according to the user id, and the transaction types are grouped and summed, and we get the 3-dimensional array, which is the user id, revenue and expenditure respectively. The user browsing behavior has 4 dimensions, which are user id, time stamp, browsing behavior data, and browse child behavior number [12]. We select user id and browse behavior data, group and sum by user id, and get the user id of 2-dimension array and browse behavior data.

Credit card bills recorded a total of 14 dimensions, which are user id, respectively, bills time stamp, bills of the previous period, payments of the previous period, credit cards, the current account balance, the current bill minimum payments, consumption number, the bill amount in current period, the amount of adjustment, circulation interest, available amount, borrow cash limit, payment condition, we to each column averaging, respectively. The loan time information is 2 dimensions, respectively for user id and loan time [13]. The customer has a 2-dimensional record of the delinquent behavior, which is the user id and the overdue label.
We then spliced all the data and eliminated duplicate columns, 24 dimensions. Because the data itself is not complete, there are still many missing data for the data that is stitched together [14]. For data missing from user base information, we chose to discard this batch of data using the dropna method. For the other missing data, there are two ways, one to fill in, another for the average filling, here we adopt average filling method, the calculation of each column of existing data, the average for the missing part of the fill in the average, as shown in Figure 2.

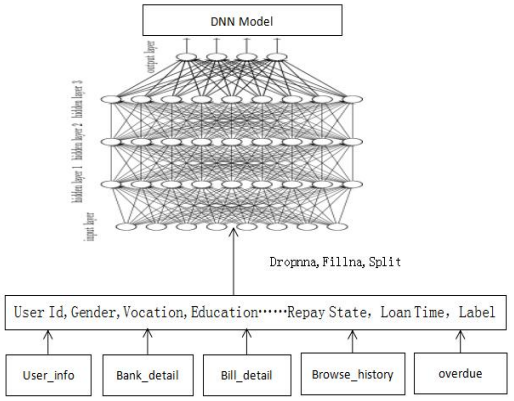

Figure 2: The Structure of DNN

2.3.2 DNN Algorithm Design

First, we need to partition the dataset. For these data, we divide the data set into training data and test data. We selected 80 percent of the data in a random selection of data as our training set, and we used the remaining 20 percent as our test set to evaluate our algorithm's accuracy.

Second, we need to set the parameters of the DNN algorithm. We will study the first 23 dimensions of the array, and set three layers of hidden layers, 10 stories, 20 stories, 10 layers, defining two output forms: 0 and 1 respectively. The DNN algorithm is designed as follows:

Step 1: read the data set
Step 2: divide the training set, test set
Step 3: put the training set into DNN model training 2000 times
Step 4: predict the test set and output the accuracy

3 EXPERIMENT AND RESULTS

3.2 Dataset

Rong360 works with financial institutions on the platform, provided nearly $70000 in loans the user's basic information, consumption behavior, bank payments, such as data, we use this batch of data as the training and test sets. Dataset is showed as Table 1.

Table 1: Dataset Used in Experiment

Dims	24
Precision	float
NO.of Raw data	55596
NO.of Processed data	11120

3.3 Evaluation Indicators

In this paper, the classification report function in sklearn package is used as the evaluation index. There are four scenarios for data test results:

(1) TP: the forecast is positive, and the label is positive.
(2) FP: the prediction is positive, and the label is negative.
(3) FN: the prediction is negative, and the label is positive.
(4) TN: the forecast is negative, and the label is negative.

The function displays the main classification indicators, returns the accuracy of each class tag, recall rate and F1 value. The following is the calculation formula of accuracy, recall rate and F1:

$$precision = \frac{TP}{TP+FP}, \quad recall = \frac{TP}{TP+FN} \quad F1 = \frac{2 * precision * recall}{precision + recall}$$

4 RESULT

After processing the data set and establishing the model, the following is the test of the accuracy of the algorithm. We have tested the algorithms mentioned above to compare the learning ability of the various algorithms in user loan and the accuracy of the test. The results are shown in the Table 2, Table 3, Table 4.

Table 2: Bayesian Model

	precision	recall	f1-score	support
Class 0	0.88	0.93	0.9	9703
Class 1	0.22	0.13	0.16	1417
AVG/ TOTAL	0.8	0.83	0.81	11120

Table 3: kNN Model

	precision	recall	f1-score	support
Class 0	0.88	0.98	0.92	9703
Class 1	0.28	0.07	0.11	1417
AVG/ TOTAL	0.8	0.86	0.82	11120

Table 4: DNN Model

	precision	recall	f1-score	support
Class 0	0.87	1	0.93	9703
Class 1	0.73	0.01	0.02	1417
AVG/ TOTAL	0.86	0.87	0.82	11120

In the table, 0, 1 represents the output of the prediction model, 0 represents users who are not overdue, and 1 represents users who are overdue, and the support is the number of occurrences of each class. From precision, we can know our model's ability to make a correct prediction, and from recall, we can learn our model's ability to recognize the class we want from the test set. The last line of the table is the weighted average of each indicator, it shows our model's ability more comprehensively.

4.2 Analysis and Comparison of Experimental Results

As shown in the table above, the Bayesian model and kNN model have similar results in predicting accuracy, but the kNN model performs better in the recall rate. Compared with the previous two algorithms, the DNN model showed better feature learning ability, especially the accuracy of 86%, which was satisfactory [15]. And it is worth noting that in the category for the prediction of 0, the accuracy of three kinds of algorithms are higher than the category of 0 accuracy, kNN model and the Bayesian model is significant difference among them, and do well in this respect within DNN models, specification for this part of the characteristics of the study, within DNN is much better than the above two algorithms [16].

5 CONCLUSION

As the number of overdue loans users increase at an alarming rate, bringing incalculable harm to Banks and society, but now for this research is still rare. In this context, this article puts forward the loan risk prediction within DNN algorithm as the core of the user. This paper mainly introduces the main application of DNN algorithm in user loan risk prediction, and elaborates the current economic background, traditional risk forecasting method. On this basis, the prediction model based on DNN algorithm is proposed, and the prediction results are compared with the traditional algorithm, and the feasibility of the model is verified.

However, the DNN model proposed in this paper still has some limitations and needs to be improved in future research. The main problem is the data set. Due to data involving personal privacy and confidentiality, dimensional data we get is not the most primitive data, and these data are in a certain extent, the lack of, these two causes within DNN algorithm in the process of learning can't better model. If future research can deal with these two problems, it is believed that the prediction accuracy can be greatly improved.

ACKNOWLEDGEMENTS

This work was supported by the NSFC (No.61502247, 11501302, 61502243,91646116), China Postdoctoral Science Foundation (No.2016M600434), Natural Science Foundation of Jiangsu Province (BK20140895, BK20150862), Scientific and Technological Support Project (Society) of Jiangsu Province (No.BE2016776), and Postdoctoral Science Foundation of Jiangsu Province(1601128B), Opening Project of Collaborative Innovation Center for Economics crime investigation and prevention technology(JXJZXTCX-015)

and Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing(2017A10).

REFERENCE

[1] Lavine, M., West, M., Praeger, A. 1992. Bayesian method for classification and discrimination. The Canadian Journal of Statistics, 20 (4), 451-461.

[2] Allenby, G.M. 1990. Hypothesis Testing with Scanner Data: The Advantage of Bayesian Methods. Journal of Marketing Research, 27 (4), 379-389.

[3] Wang, X.L., Ma, Y., Wilkes, D.M. 2015. A fast MST-inspired kNN-based outlier detection method. Information Systems, 48 (C), 89-112.

[4] Ghosh, S., Reilly, D. 1994. Credit Card Fraud Detection with a Neural Network [C]. Proceedings of 27th Hawaii International Conference on System Sciences, 621-630.

[5] Brause, R., Langsdorf, T., Hepp, M. 1999. Neural Data Mining for Credit Card Fraud Detection [C]. IEEE International Conference on Tools with Artificial Intelligence, Evanston, 103-106.

[6] Aleskerov, E., Freisleben, B., Rao, B. 1997. CARDWATCH: A Neural Network-Based Database Mining System for Credit Card Fraud Detect ion [C]. Proceedings of the IEEE ／ IAFE on Computational Intelligence for Financial Engineering, 220-226.

[7] Syeda, M., Zhang, Y.Q., Pan, `Y. 2002. Paral lel Granular Neural Networks for Fast Credit Card Fraud Detection [C]. Proceedings of the IEEE International Conferenceon Fuzzy Systems, 572-577.

[8] Stolfo, S.J., Fan, D.W., Lee, W. 1997. Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results [J]. Working notes of AAAI Workshop on AI Approaches to Fraud Detection and Risk Management, 97-07.

[9] Chan, P.K., Fan, W., Prodromidis, A.L. 1999. Distributed Data Mining in Credit Card Fraud Detection [J]. IEEE Intelligent Systems, 14 (6), 67-74.

[10] Kafai, M., Bhanu, B. 2016. Dynamic Bayesian Networks for Vehicle Classification in Video. IEEE Transactions on Industrial Informatics, 8 (1), 100-109.

[11] Fu, C., Zhang, P., Jiang, J., Yang, K., Lv, Z. 2017. A Bayesian approach for sleep and wake classification based on dynamic time warping method. Multimedia Tools and Applications, 76 (17), 1-20.

[12] Sahami, M., Dumais, S., Heckerman, D., Horvitz, E. 1998. A Bayesian Approach to Filtering Junk E-Mail. Papers from the Workshop Aaai, 1-8.

[13] Soucy, P., Mineau, G.W. 2001. A simple KNN algorithm for text categorization. IEEE International Conference on Data Mining, 647-648.

[14] Zitzmann, S., Lüdtke, O., Robitzsch, A. 2015. A Bayesian Approach to More Stable Estimates of Group-Level Effects in Contextual Studies. Multivariate Behavioral Research, 50 (6), 688.

[15] Stephens, M., Balding, D. J. 2009. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics, 10 (2), 681-690. Chuang, Y.Y., Curless, B., Salesin, D.H., Szeliski, R. 2001. A Bayesian Approach to Digital Matting. Cvpr, 2 (1), 264-271.