Economic Management & Global Business Studies
World Scientific Reports
2023, Volume1, Issue1, 2023, Type: Articles

Research article

An Intelligent System for Default Identification in Commercial Bank

Mia Anderson[1] **Oliver King**[2*] **Tsai-Qin**[2]

1.Colorado State University, College of Health and Human Sciences, United States

2.Queensland University of Technology, Faculty of Health, Australia

*oliver.king@qut.edu.au

Credit default identification technology usually requires a higher classification accuracy, and low rate of false positives, in view of the traditional credit default identification model based on machine learning, in dealing with high dimensional imbalance of financial transaction data on the problem of lower overall classification accuracy, along with the development of the large data, the credit transaction data of large-scale and high dimension feature, It is necessary to ensure the efficient and timely construction of the credit default recognition model. Semi-supervised learning can reduce the cost of manual labeling, and train the model in time by making full use of a large amount of unlabeled data and a small amount of labeled data, which has important research significance in the field of credit default recognition. At the same time, the traditional supervised learning model can only recognize the default behaviors that have happened before, but the default behaviors are variable, and the semi-supervised learning algorithm can recognize the unknown default behaviors. From the perspective of semi-supervised learning, DBN, as a semi-supervised deep learning framework, utilizes its feature learning ability and iForest unsupervised learning algorithm's ability to identify abnormal behaviors to propose a default identification method by use of the DBN-Iforest. This paper first introduces the experimental data, including the data source and the pretreatment process of the original data. Then the performance evaluation criteria applied to all experimental models in this paper are introduced in detail. Then the steps of DBN-iForest algorithm are described in detail. Next, algorithms for iForest optimization are introduced, including particle swarm optimization algorithm and simulated annealing optimization algorithm. Finally, through comparative analysis of experimental results at different levels, the advantages of the DBN-Iforest model are proved, the classification performance is improved, and the difficulty of data resource shortage in credit default recognition is overcome. The semi-supervised default identification method by use of the

DBN-IForest fully improves the classification performance of DBN and iForest by using less label data.

Keywords: Default Identification; Semi-supervised learning; Particle Swarm Optimization; Simulated Annealing

Statement: The data in this study can be provided without reservation by the corresponding author, also, the authors have no potential conflicts of interest.

1. Introduction

Nowadays, in the research on risk control in the financial field, the data sources are usually the credit investigation data of central banks, or the customer data collected internally by banks and other financial institutions, as well as some data purchased externally, such as communication operators and other external institutions. The data collected through the above means usually have certain problems, such as lack of diversity, lack of integrity, accuracy and authenticity to be considered, and the above information cannot be obtained in time, which cannot meet the real-time requirements. From the point of view of business development mode, fraud is mostly concealed and changeable, and it is difficult to identify fraud through a single data information. Conventional financial foundations frequently construct credit scoring models in view of rule base or factual investigation of authentic information to accomplish risk control. Notwithstanding, these techniques by and large have restrictions regarding idealness and information uprightness, and can't accomplish extensive and risk control. From the perspective of data analysis technology, the commonly used credit default recognition and throwing algorithms mainly include Artificial Neural Network (ANN), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Support Vector Machines (SVM), etc. Financial market has a series of complex and nonlinear characteristics. These classical measurement models containing parameters are not suitable for analyzing complex, high-dimensional and noisy massive transaction data. For example, The performance of classical machine learning models depends on

artificial feature selection to a large extent, which is not enough to deal with the risks from all channels all the time.

Represented by deep learning of the artificial intelligence technology, combined with the massive trading data of big data environment, able to take care of the issue of difficult high-dimensional complex data analysis, significantly reduce the cost for human, and strengthen the risk control and business processing ability, and achieve good application in finance data risk control. Compared to artificial description, construct characteristics of technology, deep learning could make well use of large data under the background of massive data resources to train the model, can describe the original data hidden rich intrinsic information, obtain the most effective feature expression, and through the construction of relatively complex network structure to fully excavate the correlation between the data. How to identify quickly and accurately fraudulent transactions is a huge challenge in the financial sector. Reasonably control fraud risks in the financial field, organize illegal financial behaviors, avoid economic losses, and prevent and control user excessive consumption and excessive liabilities are of great theoretical significance and practical application value to the competitiveness of excessive consumption and risk management level of domestic financial institutions through technical ability. The extension of the application scope of DL has become a research hotspot in many different fields. In this paper, it is mainly employed on default identification.

2. Related Works

2.1 Related Works on prediction problems

In prediction models, input data are generally high-dimensional and sparse, and such input data is the most common in network scale recommendation systems [1]. For example, for the prediction of click-through rate, the input is mainly classified features such as "country= China". The unique thermal coding of these features is usually "[0,1,0]". However, this tends to result in excessive high-dimensional feature Spaces, and to reduce the dimensionality these binary features are generally converted into dense vectors with real values (or called embedding vectors), a process often

called embedding [2].

Prediction models based on deep learning are of great importance for dealing with large-scale data prediction in the future [3], for example, credit prediction and click rate prediction are both large-scale problems. Identification of complex prediction features and exploration of implicit or rare crossover features is the key to do a good job in forecasting. Nevertheless, most of the data in the web-scale recommendation system are discrete and classified [4], which leads to a large number of sparse feature Spaces, making feature exploration more challenging and leading to the generation of most linear methods, such as logistic regression model [5]. Linear methods are simple in structure, explicable and easy to expand, but they limit the expressive ability of the model. On the other hand, cross features are very important in improving the expressive ability of the model. The disadvantage is that it often requires manual feature engineering or ergodic search to identify these features [6]. Furthermore, it is difficult to generalize these features to implicit feature crossover. In order to avoid task-specific feature engineering, prediction models based on deep learning are the key to solving this problem. Neural networks can start learning from a single feature without human intervention [7], and this potential has been demonstrated in natural language processing and image recognition [8]. Features extracted from specific task towel by convolutional neural network have replaced those extracted manually by SIFT algorithm and become the latest technology in the field of image recognition [9]. Similar models have also been applied to natural language processing to build a language processing model from scratch without a lot of feature engineering [10]. In order to liberate people from tedious feature engineering, neural network must be used to complete automatic feature extraction.

2.2 Characteristics of prediction problems

Whether credit or hit the prediction, I don't like images, voice, and other fields with continuous, dense data, time and space of good local correlation, in most of the input of prediction model are discrete and high dimension, scattered characteristics on different category, to solve the prediction model of high dimensional discrete data input, The key is to use embedding to reduce the dimension of data into dense and

continuous vectors. For example, FM is used for pre-training [11], or joint training with the method, or concat with the embedding feature vectors extracted from other data sources. Secondly, it is always a debate whether the prediction model should be wide or deep. Wide&Deep model framework proposed by Google [12] once became the basic framework in the industry after it was proposed. The methods in the field of prediction are either modified in the Wide part or the Deep part, and the essence is the integration of common methods.

2.3 Related Works on credit default

Foreign research on the recognition and management of credit default started earlier, mainly focused on the judgment of default behavior based on rule base and traditional machine learning algorithm. In 1980, Wiginton et al. conducted comparative tests on individual credit models based on linear discrimination and logistic regression, and the results showed that the latter had higher prediction accuracy [13]. In 1992, Nath et al. used commonly used mathematical programming method and regression analysis to conduct credit modeling in literature [14], and experiments showed that the performance of the two algorithms was relatively close. In the same year, Davis et al. compared and analyzed top-down inductive learning algorithms (G&T and ID3) with multi-layer perceptron and back-propagation neural network algorithm, which was verified in the Bank of Scotland credit dataset [15]. In fact, after 1980, data mining research began to rise, and ANN and SVM were applied in the field of credit default recognition [16]. Later, the application of artificial neural network in the financial field attracted the attention of many scholars, who applied it to credit scoring [17,18]. Fantazzini et al. used random forest algorithm to study the credit risk violation of small and medium state-owned enterprises, and the results showed that its performance was better than the classical logistic regression model [19]. Since then, ensemble learning has become an significant study direction on finance. Based on Bagging and Boosting, West et al. constructed an artificial neural network as a meta-learner in 2005, integrated and modeled it, and finally verified it, indicating that the classification performance of this model has been improved [20]. In 2007, Kirkos et al. conducted fraud identification on Greek manufacturing companies based on

neural network algorithm and comparison with statistical methods. In 2011, Bhattacharyya et al. compared the identification of fraudulent credit card transactions by logistic regression, SVMs and RF. The results showed that logistic regression algorithm was inferior to the other two algorithms in terms of specificity and accuracy. There are many researches in the financial field based on integrated learning [21]. As the data machines in the financial field often show unbalanced characteristics, abnormal detection methods are also commonly used. Yamanishi et al. used unsupervised SmartSifter algorithm and Hellinger distance to predict fraudulent behaviors for medical insurance fraud [22]. In 2015, AlaRaj et al. proposed LR, ANN and SVM as base classifiers to build a credit scoring model integrating the same base classifier and different base classifiers. The integration of different base classifiers showed better classification prediction performance in the comparison of experimental results, and was also better than the performance of a single classifier [23].

Nowadays, although deep learning has many successful cases and increasingly mature technologies in CV, speech identification and NLP, there are few related applications and researches in the field of credit default recognition. This paper combines traditional machine learning algorithm with deep learning algorithm to make up for the defects of traditional methods in high-dimensional transaction data and improve the detection rate of fraudulent transaction behavior prediction. The research objective of our manuscript aim to construct a credit default recognition method based on deep learning from supervised learning and semi-supervised learning. In a word, our paper mainly has the following two aspects of research contributions and highlights:

(1) Proposed a semi-supervised credit default recognition model combining DBN and isolated forest algorithm. In view of the problem that annotation based on massive financial transaction data needs a lot of time, material resources and manpower, and iForest algorithm randomly selects dimensions under high-dimensional data every time, wastes a lot of dimension information and reduces the reliability of the algorithm, DBN model is used to realize the data mapping between different layers. In

addition, the complex mapping between original data and features learned autonomously through the network can be completed through a small amount of label data. Combined with the advantages of unsupervised learning iForest algorithm in detecting abnormal data, the anomaly recognition ability and classification performance of the model can be improved.

(2) A different optimization algorithm was proposed to improve the parameters of the DBN-iForest model. The main parameters affecting the performance of the isolated forest algorithm were optimized, and the average accuracy of the DBN-iForest model was taken as the objective function of the optimization algorithm. Two optimization algorithms, particle swarm optimization and simulated annealing, were mainly used. Finally, the relative optimal combination of parameters is found to further enhance the performance of the method classifier.

3. Methodology

3.1 Deep belief network

With the rapid development of DL related technology research, many NNs based models with deep structure have emerged, among which the proposal of deep belief network has become a landmark breakthrough in the development of deep learning field. Before 2006, the training of artificial neural network usually adopts back propagation algorithm, which has problems of gradient disappearance and difficulty in feature extraction in multi-hidden layer neural network, and DBN has solved the above problems to some extent.

Deep belief network is a deep neural network model [24], in which Restricted Boltzmann Machine (RBM) is the most important basic constituent unit of DBN. Its essence is to maximize the probability of the model producing qualified samples. The essence of DBN is to add probabilistic generation model of classifier combination based on dimension reduction of multi-layer RBM stack. Deep belief network is a semi-supervised learning method, by unsupervised greedy stratified by training multilayer matter and supervised learning method to the whole network reverse fine-tuning two stages, fusion without supervision, and have the advantage of

supervised learning, the powerful features of deep structure of the network expression ability, step by step, the characteristics of the inductive extract more representative Thus, it has powerful classification and prediction functions of high-dimensional feature vectors.

3.1.1Semi-supervised learning

DBN is a semi-supervised learning method consisting of multi-layer RBM unsupervised pre-training and supervised reverse fine-tuning. This method mainly considers to take full advantage of an enormous number of unlabeled information and few marked information to train and classify the model, and its theoretical basis is as follows: labeled dataset L $L = \{(x_1, y_1), (x_2, y_2), \cdots (x_{|L|}, y_{|L|})\}$, unlabeled dataset= $\{x_1^{'}, x_2^{'}, \cdots, x_{|U|}^{'}\}$, learner $f: X \to Y$. Where, $x_i^{'}, x_i^{'}$ $\lfloor X, y_i \lfloor Y$ represents the corresponding label and $|L|, |U|$ represent the size of the L, U dataset respectively.

D.j.Miller and H.S.Uyar investigated the achievability and soundness of utilizing unlabeled information to enhance the performance of learner according to the viewpoint of data distribution estimation [25]. If all data meet the distribution composed of L Gaussian distribution, its specific description is shown in formula (1).

$$f(x|\theta) = a_{l=1}^{\circ L} \alpha_l f(x|\theta_l) \tag{1}$$

Where, $a_{l=1}^{\circ L}$ is the mixing coefficient and $\theta = \{\theta_l\}$ is the parameter. The random variable determined by the probability $P(c_i|x_i, m_i)$ of the selected mixed component m_i and the eigenvector x_i is the label. Therefore, based on the assumption of maximum posterior probability, the optimal taxonomy is shown in Formula (2).

$$m_i \, \mathrm{h(x)} = \mathrm{argmax} a_k^{\circ j} P\left(c_i = k \middle| m_{i=j}, x_i\right) P(m_i = j|x_i) \tag{2}$$

Where, $P(m_i = j|x_i) = \frac{\alpha_j f\left(x_i \middle| \theta_j\right)}{a_{l-1}^{\circ L} \alpha f\left(x_i \middle| \theta_j\right)}$. As indicated by the above depiction, the reason for semi-administered learning can be deciphered as utilizing preparing information to gauge $P\left(c_i = k \middle| m_{i=j}, x_i\right)$ and $P(m_i = j|x_i)$. It can be seen that the former is related to labels, while the latter is related, so a large number of unlabeled data can improve the generalization ability of the classifier.

The investigation consequences of T.Chang and F.J.Oles showed that when a model was deteriorated into the type of $P(x, y|\theta) = P(y|x, \theta)P(x|\theta)$, unlabeled data could improve the performance of the classifier [26].

3.2 Isolated forest algorithm

Isolation Forest (iForest) [27,28] algorithm is usually used to monitor abnormal data or mine outliers. At present, due to its linear time intricacy and high accuracy, the isolated forest algorithm has been broadly applied in the fields of intrusion detection, finance, intelligent transportation and medical treatment [29,30], and is suitable for processing high-dimensional data and large-scale data, with high practicability. The essential idea of iForest algorithm is about the strategy of dividing data space. Firstly, it randomly selects dimensions and segmentation values, and then obtains a convergence value by constructing multiple Isolation trees (iTree) through integration method (Monte Carlo method) and these isolation trees build a forest. The stability and accuracy of the algorithm are improved. In fact, the essence of iTree structure can be regarded as a random binary tree, or similar to its mechanism. Each node on iTree either contains two nodes of left child and right child, or only contains a single leaf node. The construction of iForest usually involves the following important parameters, that is, the number of iTree and the number of iTree random samples. Generally speaking, the stability of the algorithm is related to the number of isolated trees, and the more iTree the algorithm has, the better its stability. The purpose of random sampling is to separate abnormal samples from normal samples. When the two samples are too close to each other, the number of abnormal data segmentation will increase, leading to greater difficulty in distinguishing abnormal samples. Therefore, reasonable parameter setting can enhance the classification performance of the method. IFroest algorithm, moreover, you also need to give all isolated tree to limit its cannot exceed the maximum depth, can to a certain extent, guarantee the algorithm's time complexity is lower, because of the abnormal data samples is relatively far less than the normal sample, the number of at the same time, the characteristics of the two significant differences, due to the abnormal data path length compared with small, So limiting the height of isolated trees can improve the efficiency of isolated forest

algorithm to a certain extent.

After the completion of the construction of n iTrees, the abnormal score of each sample needs to be calculated. The tested sample is traversed through all iTrees in iForest, and the path length of each traversal is retained. The path length of the data in all iTree is summarized, and the abnormal score of the final sample is obtained according to the algorithm's own anomaly score calculation formula.

4.An Intelligent System for Default Identification in Commercial Bank

From the perspective of semi-supervised learning, DBN, as a semi-supervised deep learning framework, utilizes its feature learning ability and iForest unsupervised learning algorithm's ability to identify abnormal behaviors to propose a default identification system based on DBN-Iforest.

4.1 Default identification system based on DBN-iForest

4.1.1 Algorithm idea

Traditional credit default identification methods are usually based on supervised machine learning methods, which not only require a large amount of labeled training data to construct models, but also perform better for known fraud types, while financial data lack a large amount of labeled data, and the forms of financial fraud are constantly changing. In the training of models, it is an important study problem in the field of finance to reduce the model's demand for labeling samples and improve the recognition rate of fraudulent transactions. DBN, as a semi-supervised deep learning model, can make full use of a large number of unlabeled samples. Therefore, this paper proposes to construct a deep belief network model to extract features through training a large number of unlabeled data, and finally complete feature dimension reduction of original high-dimensional data. When the iForest unsupervised learning algorithm identifies the default behavior, it randomly selects a certain attribute to divide the data set each time, and builds an isolated forest on the high-dimensional dataset, which wastes a lot of dimensional information and reduces the reliability of the algorithm. Therefore, we proposes a semi-supervised credit default recognition

model based on DBN-iForest, which can enhance the classification performance of the credit default recognition model on the premise of using a few label data.

4.2 Optimization of default identification system based on DBN-iForest

In order to further improve the default identification system based on DBN-iForest, this section mainly adopts optimization measures for the important parameters of iForest. As the number of iTree continues to promote, the difference between iTree in iForest will become smaller and smaller, ultimately leading to the reduction of generalization performance. The increase in the number of iTree samples means that the scale of a single iTree increases accordingly, which eventually leads to the increase of memory space consumed by iTree construction and the complexity of computation. The selection of parameters needs to consider the specific actual situation to decide, often setting reasonable values for variables can enhance the classification accuracy of the model to a certain extent. Grid-based search and heuristic search are common parameter optimization strategies.

For further improvement of credit default recognition based on semi-supervised DBN-iForest, this section proposes to optimize important iForest parameters with PSO algorithm.

5.Data processing

The experiment in this paper adopts the public data loan-default-prediction provided by Kaggle competition. This dataset is used in a loan default prediction contest and is a list of financial transactions related to individuals. The total number of samples in the training set is 105471, the number of positive samples is 95688, and the number of negative samples is 1145,771. The labels are composed of two types: default and non-default. Due to the extreme imbalance of the original data, this experiment deleted most of the class instances through the undersampling technology, and improved the sampling ratio of negative samples. 11,100 positive samples and 900 negative samples were randomly selected, and finally divided in the ratio of 8:2, that is, 9600 transactions were training samples and 2,400 transaction records were test samples.

5.1 Experimental evaluation indicators

(1) Basic evaluation indicators

In this paper, credit default recognition is divided into two categories, including normal and default. Default means that the user fails to repay the loan within the specified time, and normal means that the user pays the loan within the specified time. Therefore, the confusion matrix of credit default recognition is introduced in the Table 1.

Table 1 Confusion matrix based on credit default identification

	The prediction is normal	The prediction is default
Reality is normal	TP	FN
Reality is default	FP	TN

The confusion matrix for credit default recognition is described as follows: Where, TP means correctly identified normal transactions, FP means incorrectly identified default transactions, FN means incorrectly identified normal transactions, TN means the correctly identified default transactions, and $N = TP + FN + FP + TN$ is the total amount of samples.

Classification accuracy ($Accuracy$), its meaning for the total accuracy of all samples correct classification, as shown in formula (3).

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} = \frac{TP+TN}{N} \tag{3}$$

Accuracy rate, take negative samples as an example, that is, for this class, the ratio of the number of samples predicted to be of this class and correct to the total number of samples actually classified into this class, as shown in Formula (4).

$$Precision_n = \frac{TN}{TN+FN} \tag{4}$$

Recall rate, taking negative samples as an example, represents the ratio of the number of correctly classified classes to the actual number of negative samples, as introduced in Formula (5).

$$Recall_n = \frac{TN}{TN+FP} \tag{5}$$

F1-score represents two comprehensive evaluation indexes of accuracy rate and

recall rate, as shown in Formula (6). The higher the F1-score, the better the performance of the model.

$$F_1 = 2 \times \frac{Precision \times Recall}{recision + Recal} \tag{6}$$

6.Ablation experiments

6.1 Comparative analysis of default identification system based on different unsupervised learning algorithms

The default identification system based on OneClassSVM and LOF is based on the operating system macOS 10.12.6, processor 2.0GHz Intel Core I7, memory 16GB, programming language Python 3.5 environment. The above two algorithms are implemented using the sklearn library encapsulated in Python.

The iForest algorithm is compared with LOF and OneClassSVM algorithm with rows. The nearest neighbor number k that influences the performance of LOF method is determined through experiments, and the final value is 700. The parameter of OneClassSVM is the default value, the final results are introduced in Table 2.

Table 2 experimental results of iForest and other unsupervised algorithms

Algorithm	AUC	Run time
iForest	0.6628	2.4674
LOF	0.558559	23.853765964508
OneClassSVU	0.557748	6.460512161254883

From the Table 2, we can found that, the AUC value of iForest-based credit default recognition model is 10.4% higher than that of LOF model and 10.5% higher than that of OneClassSVM model. Therefore, relatively speaking, iForest can achieve higher classification effect in credit default transaction recognition. Secondly, the running time of iForest algorithm is also less than that of LOF and OneClassSVM, so iForest has higher efficiency in credit default transaction identification, which proves the advantages of iForest model in credit default identification that cannot be ignored.

6.2 Comparison between DBN-iForest and other supervised models

The credit default recognition model based on SVM, KNN, LR and RF is constructed in the environment of operating system macOS 10.12.6, processor 1.8

GHz Intel Core i5, memory 8GB and programming language Python 3.5. The above four algorithms are implemented using python's sklearn library.

The DBN-iForest model is compared with SVM, KNN, LR and NN classifiers. LR, KNN and SVM use default parameters. Among them, SVM algorithm has many problems such as low efficiency, long running time, high memory requirement and manual adjustment of many parameters in the training process. The neural network model is a three-layer network structure [740,100,1]. Adam algorithm is used for training. The batch size is 100, the number of iterations is 100, and the training time is long. RF Sets the number of decision trees to 200. Table 3 indicates the experimental results.

Table 3 Comparison of experimental results between DBN-iForest algorithm and other

Algorithm	AUC	Classification accuracy
DBN-iForest	0.7605	0.8934
SVM	0.5	0.9248
LR	0.6022	0.9269
KNN	0.5277	0.9201
RF	0.5409	0.9235
NN	0.6432	0.9177

From the Table 3, we can found that, although the classification accuracy of the DBN-iForest proposed in this paper is not as good as SVM, LR, KNN, RF and NN, the AUC value is significantly higher than other models. For unbalanced credit data, the meaning represented by AUC value has more practical reference significance, and the higher the AUC, the better the model classification performance.

6.3 Comparison between DBN-iForest and the model before and after combination

Experimental comparison was made between the proposed credit default model based on DBN-iForest and PCA- iForest, DBN and iForest. The ablation experiment results are introduced in Table 4:

Table 4 Ablation experiment results of various model experiments

Algorithm	AUC	Classification accuracy(%)
PCA- iForest	0.5418	0.8688
iForest	0.6396	0.8755
DBN	0.58085	0.9091
DBN-iForest	0.7605	0.8934

From the Table 4, we can found that, first of all, the AUC value of iForest model reduced by 9.78% and accuracy decreased by 0.67% after PCA dimensionality reduction, indicating that the classification effect of model was not ideal by PCA dimensionality reduction, which fully indicates that PCA linear dimensionality reduction method is not applicable. Compared with iForest model, the accuracy of DBN- iForest model proposed in our article is 1.02% higher and the AUC value is 12.09% higher than that of iForest model, which fully proves that DBN has better advantages than PCA in feature extraction. It also shows that one of the main reasons affecting iForest performance in high dimensional data is dimension reduction. Secondly, although the classification accuracy of the DBN- iForest model proposed in our article is slightly lower than that of DBN model, its AUC value is improved by 17.965%, which fully demonstrates that the combined model has better classification performance. Moreover, for the field of credit default recognition, it is more important to improve the identification of abnormal fraudulent transactions. The classification accuracy is similar, but the AUC is quite different, which fully indicates that the default identification system based on DBN- iForest has a stronger ability to identify default behaviors.

6.4 Comparison of optimized models

PSO algorithm and SA algorithm are employed to find the optimal solution of iForest. The DBN-iForest model pairs before and after the final optimization are shown in Table 5.

Table5 comparison of models before and after optimization

Algorithm	AUC	Classification	iForset main

		accuracy(%)	parameters
DBN- iForest	0.7605	0.8934	(100,250)
DBN-iForest-SA	0.7753	0.8923	(85,62)
DBN-iForest	0.7709	0.8937	(53,54)

From the Table 5, we can found that, the important parameters of the iForest algorithm are optimized by the optimization algorithm. The SA optimization algorithm proposed in our article enhances the AUC value of the DBN- iForest model by about 1.5%, while the POS optimization algorithm improves by about 1.1%. In a word, the experiment accuracy and performance of the method could be further improved by using optimization algorithm in practical application.

7.Conclusions

Through in-depth analysis of the current problems existing in the credit default identification field, this article mainly from the two layers of a semi-supervised learning and supervised learning, further study of the default identification system based on the DL, mainly using the characteristics of deep learning algorithm for high-dimensional data learning ability strong characteristic, combine with traditional machine learning algorithm, enhance the classification performance of classical ML-based method on high dimensional data. The main work is as follows:

(1) Literature research was conducted on the research issues in this paper, and the related works was summarized. The current problems of credit default were analyzed, and the application of DL in this field was deeply understood, which laid a foundation for the default identification system based on deep learning proposed later.

(2) in view of the insufficiency of tag data, and unsupervised learning isolated forests algorithm in high-dimensional data fraud recognition performance of default on the limited problem, put forward the combination of the depth of the belief network self-learning ability, modeling ability and the advantage of strong robustness, build a semi-supervised credit default recognition based on DBN - iForest model. The results indicate that this model can fully improve the classification performance of DBN and iForest by using less label data.

(3) In order to further optimize the number of isolated forest at the top of the default identification system based on DBN-Iforest, two commonly used optimization algorithms (PSO and SA) are selected in this paper to find the optimal parameter combination for iForest respectively. Due to different optimization methods of each optimization algorithm, different optimal parameter combinations are finally obtained. Through comparative experiments, it is verified that the two algorithms can achieve the best combination in the problem of credit default recognition and further improve the classification performance of the model. However, by comprehensive comparison, SA has better relative effect.

Author Statement

Fangze Cheng: Conceptualization, Method, Validation, Writing-original draft. **Qianning Tang:** Method, Validation, Investigation, Resources, Data processing, Writing-review & editing. **Jie Liu:** Software,Writing - review & editing, Supervision. Furthermore, the authors state that there is no conflict of interest.

Funding

There is no specific funding to support this study.

Reference

[1] Hinton G，Deng L，Yu D et a1．Deep neural networks for acoustic modeling in speech recognition[J]．IEEE Signal processing magazine，20 1 2，29.

[2] Zheng Y—T, Zhao M， Song Y et a1. Tour the world：building a web-scale landmark recognition engine[A]. 2009 IEEE Conference on Computer Vision and Pattern Recognition[C]. IEEE，2009：1 085—1 092.

[3] Lu Y'Duan Y Kang W et a1. Traffic flow prediction with big data: a deep learning approach[J]_IEEE Transactions on Intelligent Transportation Systems，20 1 5，1 6(2)：865-873．

[4] Shah Y Hoens T R，Jiao J et a1. Deep Crossing：Web-Scale Modeling Without Manually Crafted Combinatorial Features[A]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM

[C]. 2016：255-262.

[5] Andrewcucchiara. Applied Logistic Regression[J]. Technometrics，20 1 2，34(3)：2.

[6] Bobadilla J, Ortega F, Hemando A et al. Recommender systems survey[J]. Knowledge—based systems, 20 1 3, 46: 1 09—1 32.

[7] Lecun Y Bengio Y'Hinton G. Deep learning[J]. nature, 2015, 521(7553)：436.

[8] Schmidhuber J. Deep learning in neural networks：An overview[J]. Neural Netw，2015. 61：85-117.

[9] Li H，Zhe L，Shen X et a1. A convolutional neural network cascade for face detection [A]. Computer Vision&Pattern Recognition[C]. 2015.

[10] Wang R，Fu B，Gang F et a1. Deep&Cross Network for Ad Click Predictions[J]. 2017.

[11] Rendle S. Factorization Machines[A]. IEEE International Conference on Data Mining[C]. 2011.

[12] Cheng H T, Koc L, Harmsen J et a1. Wide&Deep Learning for Recommender Systems[A]. 2016.

[13] Wiginton, J. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. Journal of Financial and Quantitative Analysis, 15(3), 757-770. doi:10.2307/2330408

[14] Nath R， Jackson W M， Jones T W . A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis[J]. Journal of Statistical Computation & Simulation, 1992, 41(1-2):73-93.

[15] RH Davis, DB Edelman, AJ Gammerman. Machine-learning algorithms for credit-card applications[J]. IMA J Management Math, 1992.

[16] Lahsasna A， Ainon R N， Wah T Y . Credit Scoring Models Using Soft Computing Methods: A Survey[J]. International Arab Journal of Information Technology, 2010, 7(2):115-123.

[17] Gleit A . Quantitative methods in credit management. INFORMS, 1994.

[18] Desai V S， Crook J N， Overstreet G A . A comparison of neural networks and linear scoring models in the credit union environment[J]. European Journal of

Operational Research, 1996, 95(1):24–37.

[19] Fantazzini D, Figini S. Random Survival Forests Models for SME Credit Risk Measurement[J]. Methodology and Computing in Applied Probability, 2009, 11(1):29-45.

[20] West D , Dellana S , J Qian. Neural network ensemble strategies for financial decision applications[J]. Computers & operations research, 2005, 32(10):p.2543-2559.

[21] Finlay S . Multiple classifier architectures and their application to credit risk assessment[J]. European Journal of Operational Research, 2011, 210(2):368-378.

[22] Yamanishi K , Takeuchi J I , Williams G , et al. On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms[J]. Data Mining & Knowledge Discovery, 2000, 8(3):275-300.

[23] Ala"Raj M , Abbod M . A systematic credit scoring model based on heterogeneous classifier ensembles[C]// International Symposium on Innovations in Intelligent Systems & Applications. IEEE, 2015:1-7.

[24] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput. 2006 Jul;18(7):1527-54. doi: 10.1162/neco.2006.18.7.1527. PMID: 16764513.

[25] Kong Y Q , Wang S T . Feature selection and semisupervised fuzzy clustering[J]. Fuzzy Information and Engineering, 2009, 1(2):179-190.

[26] Zhou D , Schlkopf B , Rasmussen C E , et al. Learning from Labeled and Unlabeled Data Using Random Walks[C]// Pattern Recognition, 26th DAGM Symposium, August 30 - September 1, 2004, Tübingen, Germany, Proceedings. Springer Berlin Heidelberg, 2004.

[27] Fei T L , Kai M T , Zhou Z H . Isolation Forest[C]// IEEE International Conference on Data Mining. IEEE, 2008.

[28] Liu F T , Ting K M , Zhou Z H . Isolation-Based Anomaly Detection[J]. Acm Transactions on Knowledge Discovery from Data, 2012, 6(1):1-39.

[29] Sun L , Versteeg S , Boztas S , et al. Detecting Anomalous User Behavior Using an Extended Isolation Forest Algorithm: An Enterprise Case Study[J]. 2016.

[30] Ding Z , Fei M . An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window[J]. IFAC Proceedings Volumes, 2013, 46(20):12-17.