Risk Measurement of the Financial Credit Industry Driven by Data: Based on DAE-LSTM Deep Learning Algorithm

Guizhi Li, Yingkou Institute of Technology, China
Xuebiao Wang, Dongbei University of Finance and Economics, China
Datian Bi, Jilin University, China*

https://orcid.org/0000-0002-9198-3799
Jiayu Hou, Yingkou Institute of Technology, China

ABSTRACT

The risk measurement of the financial credit industry is an important research issue in the field of financial risk assessment. The design of a financial credit risk measurement algorithm can help investors avoid greater risks and obtain higher returns, so as to promote the benign development of financial credit industry. Based on the combined deep learning algorithm, this paper studies the risk measurement of financial and credit industry and proposes a fusion algorithm of deep auto-encoder (DAE) and long short-term memory (LSTM) network. The algorithm recombines the value of fixed features by using the unsupervised mechanism of DAE and extracts non-fixed features for measurement combined with the memory characteristics of LSTM network. The experimental results show that, compared with single generalized regression neural network and LSTM network, the average accuracy of DAE-LSTM algorithm is improved by about 6.49% and 3.25%, respectively, which has a better application effect in credit risk measurement.

KEYWORDS

Data Mining, Deep Auto-Encoder, Financial Credit Industry, LSTM

INTRODUCTION

Credit risk management is the core of finance. A perfect credit system plays an important role in the rapid development of the financial industry, especially the modern financial credit industry. The risk measurement of the modern financial credit industry has always been a key research issue in financial risk assessment. Financial credit risk forecasting helps to establish a data-centric financial risk early warning system to improve the accuracy of credit risk forecasting. In addition, it determines the key elements and paths that affect credit risk, helping investors reduce the losses of financial risks. In turn, financial credit risk helps the financial and credit industry improve and complete business risk management to better regulate and prevent related risks (Wickens, 2017).

The financial and credit industry has typical long-tail characteristics. The focus of traditional financial institutions is mainly on meeting the financial needs of 20% of high net-worth customers.
The other 80% of the small, scattered, personalized financing needs are distributed in the financial long tail. These needs cannot be satisfied for a long time because they require high cost and energy from traditional financial institutions. The modern financial and credit industry makes up for the deficiencies of traditional financial institutions and specializes in developing credit financing services for small, scattered, and personalized customers distributed in the tail of the financial long tail. While meeting the financing needs of many long-tail customers, it also makes the coverage of its industrial risks increasingly larger. Therefore, a financial risk early warning system for the financial credit industry is necessary.

There are two main methods for measuring and researching risk issues in the financial and credit industry: traditional statistical and econometric methods and modern artificial intelligence methods (Zhao & Jin, 2018). The statistical and measurement methods used in traditional financial credit risk prediction include discriminant analysis, logistic regression analysis, multivariate statistical regression, and probit models. By comparing and analyzing multiple regression models, West (2000) concluded that the multiexpert model and radial basis neural network are more effective in credit risk prediction and have higher accuracy of logistic regression in traditional methods. However, in recent years, the data have undergone changes like an increase in sample size, increase in characteristic indicators, and the multisourcing of data structures. Problems like the correlation between data characteristic variables, multicollinearity, imbalance between samples, and heterogeneity are also prominent. Therefore, it is more difficult to use the traditional econometric method to process the data model, resulting in a reduced prediction accuracy (Efron et al., 2004).

Scholars began to improve the algorithm based on the model design of dimension specification sample disequilibrium and traditional methods (Ozturk et al., 2016). For example, Hwang et al. (2010) introduced an ordered semiparametric function to replace the linear regression function, establishing an ordered semiparametric probit credit scoring model that achieved good results. Aburrous et al. (2010) used the fuzzy mathematics methods to point out security problems of electronic payment faced by traditional online banks. The study noted that virus attacks caused significant security risks to payment pages.

At present, the usage of improved logistic methods is widespread, including the use of multiple logit market share models to study peer-to-peer (P2P) loan evaluations (Lee & Lee, 2012), principal component analysis (PCA), and logistic regression models with bilevel selection within and between groups for grouping structural variables (Rimiru et al., 2017), credit models based on logistic regression and machine learning techniques (Paula et al., 2019), and lasso-logistic regression models under reduced demission (Chen & Xiang, 2017). These models deal with problematic data that have multiple sources, complex structures, missing labels, high dimensions, and uneven categories. In addition, these models have a better performance on credit risk measurement.

The above traditional models are widely used; however, most are based on strict assumptions. Testing and processing are required for obeying distribution, data stationarity, and correlation. Therefore, traditional statistical and measurement methods often require statistical assumptions based on certain circumstances. The time-series model of financial credit is only suitable for modeling stationary series. This restricts the widespread use and expansion of high-dimensional time series models, as well as reduces performance in processing complex structured big data.

With the rapid development of the financial technology industry, the data of the financial credit industry have shown the characteristics of huge volume, various types, low-value density, and high timeliness. Whether a model can be designed by extracting valuable indicators from complex data indicators is the key to evaluating the measurement effect of the model (Devi & Chezian, 2016). Obviously, it is difficult for traditional credit evaluation methods to extract more effective value information under a new situation. In this circumstance, scholars began to use data mining techniques like machines and artificial intelligence to solve the problem of credit risk prediction. These methods include neural networks, K-nearest neighbors, ensemble learning, support vector machines, Bayesian networks, and simulated annealing algorithms.
Ghodselahi (2011) proposed a new hybrid random forest support vector machine (RF-SVM) ensemble model for credit scoring that achieved good results compared with traditional models. Malekipirbazar and Aksakalli (2015) used a RF algorithm to predict the risk of borrowers of Lending Club, a foreign online lending platform. Mukid et al. (2018) applied the k-nearest neighbor algorithm for credit evaluation under the premise of considering the role of core points. Ermpinis et al. (2021) generated 50 multilayer perceptrons, 50 radial basis functions, 50 higher-order neural networks, and 50 recurrent neural networks for forecasting and trading economic index problems. Edmond and Girsang (2020) analyzed the feasibility of credit scoring using the iterative dichotomiser 3 (ID3) decision tree algorithm. In terms of deep learning applications, Shen et al. (2015) used an improved deep belief network to predict exchange rates.

Through verification and analysis, it was found that modeling LSTM for volatility containing noise has a significant prediction effect. Through in-depth analysis, LeCun et al. (2015) concluded that deep neural networks can ignore the influence of local extreme value problems, indicating that deep neural networks can be used as an effective method for frontier problems. Maldonado et al. (2017) empirically tested the performance of artificial intelligence classification models like SVM, Bayesian, and traditional econometric models in default judgment. The results showed that, when compared with the traditional econometric model, the artificial intelligence classification model has a lower false rate. In addition, deep learning has a more obvious application effect on complex nonlinear systems (Gopakumar et al., 2018). Chen and Ge (2019) tested the prediction effect of an LSTM neural network in dealing with financial time-series problems.

Some scholars turned their attention to the applicability of dynamic features, the ability to integrate textual information, and the initial exploration stage of models using deep learning for feature construction (Liu & Schumann, 2005). However, the research of deep learning in financial credit is still immature. Empirical research on the prediction model of dimension reduction and extraction of high-dimensional data features from the perspective of deep neural networks is rare (Wu & Huang, 2004).

From the perspective of large sample data, there are many limitations in using traditional statistical and machine learning methods to deal with financial credit risk prediction (Pan, 2011; Zhu et al., 2015). For example, the data need to be based on strict statistical assumptions or the data are likely to lose valuable information after direct dimensionality reduction. However, deep neural networks have the characteristics of self-learning data, low input requirements for data types and data structures (Heaton et al., 2017), and deep machine learning capabilities in the face of complex and changeable data. Compared with the traditional method of extracting features based on design, deep neural networks can use machines to learn features, more effectively integrate a large number of multisource data, and discover the changing laws of information (Bao et al., 2017; Gunnarsson et al., 2021). From the perspective of big data, the widespread application of machine learning and deep learning can provide more ideas for financial credit risk measurement. It plays an important role in promoting the development of the entire financial credit industry.

This article attempts to use a deep auto-encoder (DAE) network and a LSTM network for feature extraction. It then manipulates the screened valuable feature information into a fully connected neural network model for combined prediction. This combined deep-learning model applies to financial credit risk prediction. The effect is good. A method of fusion extraction using a DAE network and LSTM network for feature learning using a machine is adopted. The characteristic index is divided into fixed and non-fixed indices to dilute irrelevant factors and strengthen the role of effective factors. The algorithm can fully combine the valuable information of the feature and avoid subjective consistency. At the same time, the full connection method is used to predict the risks of the financial and credit industry based on extracted features. This can shorten the calculation time, improve the accuracy, enhance the interpretability and effectiveness of the model, and avoid the problem of gradient disappearance. By establishing a data processing model based on large sample data, the article investigates the inherent laws of financial credit data and provides suggestions for preventing financial credit risks.
The main contribution of the article is to use the DAE-LSTM combined deep learning algorithm to decompose, extract, and predict data features. It also provides solutions for research on credit risk measurement in the financial credit industry. The article explores data mining and risk prediction in the financial and credit industry.

Meanwhile, some implications from the loan interest rate, period, and amount can also be revealed. The article can judge if investors’ loan interest rate pricing is reasonable. By improving interest rate pricing, investment income can be improved without impacting credit risk. However, indicators like loan term and loan amount cannot be completely used as a single option to measure credit risk. The application of deep-learning combination algorithms for overall analysis will obtain more accurate prediction results.

RESEARCH METHOD

General Regression Neural Network

A general regression neural network (GRNN) is a variant of the radial basis neural network that can handle nonlinear, discontinuous, and high-frequency multidimensional data. Training of the network is more convenient and has good nonlinear approximation performance. The nonlinear relationship between variables can be mined by using nonlinear models. This can effectively improve the level of credit risk prediction. The credit risk prediction model has been transformed from a time-series model to a nonlinear model (Zhang et al., 2017).

The set input variable is \( X = [X_1, X_2, \ldots, X_n] \) and the output variable is \( Y = [Y_1, Y_2, \ldots, Y_k] \). \( X_i \) is the learning sample corresponding to the i-th neuron. \( Y_j \) is the j-th output. After inputting the sample, the mode layer is entered. The transfer function at this time is:

\[
P_i = \exp \left[ -\left( \frac{1}{2\sigma^2} (X - X_i)^T (X - X_i) \right) \right], \quad i = 1, 2, \ldots, n
\]

Next, it enters the summation layer to obtain the summation for these two categories:

\[
\sum_{i=1}^{n} P_i, \quad \sum_{i=1}^{n} Y_i P_i = \sum_{i=1}^{n} Y_j P_i
\]

Finally, the j-th output \( Y_j \) is obtained:

\[
Y_j = \frac{\sum_{i=1}^{n} P_i}{\sum_{i=1}^{n} Y_j P_i}
\]

The generalized regression neural network has a simple structure and a good prediction effect. However, it is prone to problems like overfitting, local extremum, and gradient disappearance or gradient explosion in the processing of complex data. Research shows that GRNN have fast convergence speed and high prediction accuracy when dealing with general sample data. This article attempts to apply it to predicting large sample financial credit data. It compares the prediction effect of GRNN and deep learning neural networks under the condition of large sample data.
DAE NEURAL NETWORKS

A DAE is an unsupervised network that can set the same input and output dimension vector to ensure that it can retain and obtain better information representation. Generally, it is time-consuming if the existing prediction model is adopted, as the prediction module needs to input and learn more than hundreds of thousands of pieces of data directly. However, when using a DAE network, it can extract the most typical information from the original data, retain the value of the original data, reduce the volume of data input, and put the processed data into a fully connected network for learning. This can shorten the time and reduce the complexity of computation (Pan et al., 2022).

The process can be divided into two steps. In the Encoding step, information A of the input layer is compressed (coded) into the hidden layer to obtain information H. The prediction error is obtained by comparing H and A before adjusting and improving the accuracy of the autoencoding through reverse transmission. Finally, essence data H is trained and extracted from the middle, hidden layer to implement the encoding.

In the Decoding step, the DAE network decompresses (decodes) the compressed data information H. This means that the essential data H in the encoding process is decompressed again to obtain a smaller volume of data with all valuable information. The data are simultaneously restored to the original dimension.

The encoding process and decoding process are expressed as:

\[ H_e(A) = Encoder(A; \Phi_E) \]  
\[ H_d(A) = Decoder(H_e(A); \Phi_D) \]

\(Encoder()\) is the encoding structure; \(Decoder()\) is the decoding structure. \(\Phi_E\) and \(\Phi_D\) are the network parameters of the encoding and decoding structures, respectively. \(H_e(A)\) and \(H_d(A)\) are the outputs of the encoding and decoding structures, respectively.

LSTM NEURAL NETWORK

LSTM neural networks, a variant of recurrent neural networks, can effectively solve the vanishing gradient problem. By introducing a set of memory cells, the LSTM model automatically learns and chooses when to forget historical information. It chooses to update the memory cells with new information.

The model controls changes in information by creating an input game, forget game, and output gate (see Figure 1). These affect the state and value of the LSTM neural network, determining which indicator information is forgotten and which should be retained. Finally, it outputs the useful data.

The number of hidden units is \(h\). A small batch input for a given time \(t\) is \(X_t \in \mathbb{R}^{n \times d}\) (\(n\) is the number of samples and \(d\) is the number of inputs). The hidden state of the previous time is \(H_{t-1} \in \mathbb{R}^{n \times h}\). The variables of time step \(t\) are as follows: input gate \(I_t \in \mathbb{R}^{n \times h}\), forget gate \(F_t \in \mathbb{R}^{n \times h}\), output gate \(O_t \in \mathbb{R}^{n \times h}\), candidate memory cell \(\hat{C}_t \in \mathbb{R}^{n \times h}\), memory cells \(C_t \in \mathbb{R}^{n \times h}\), and hidden state \(H_t \in \mathbb{R}^{n \times h}\).

\[ I_t = \sigma \left( X_t W_{xi} + H_{t-1} W_{hi} + b_i \right) \]
\begin{align*}
F_t &= \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f) \\
O_t &= \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \\
\hat{C}_t &= \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c) \\
C_t &= F_t \odot C_{t-1} + I_t \odot \hat{C}_t \\
H_t &= O_t \odot \tanh(C_t)
\end{align*}

Note that $W_{si}$, $W_{xf}$, $W_{xo} \in \mathbb{R}^{n_{od}}$, $W_{xc} \in \mathbb{R}^{d_{ch}}$, and $W_{hi}$, $W_{hf}$, $W_{ho} \in \mathbb{R}^{h_{ch}}$ are weight parameters. $b_f$, $b_o$, $b_c \in \mathbb{R}^{1\times h_{ch}}$ are bias functions.

**DAE-LSTM ALGORITHM DESIGN**

**Sequence Input**

The design of this algorithm is to decompose the input data, divide the indicators into fixed indicators and nonfixed indicators, and convert the indicators into a vector matrix. Among them, the indicators that define the borrower information and basic loan information of the j-th sample are marked as fixed indicators $M_{tj}$. The indicators that define the credit history, public information, and other credit transaction behaviors of the j-th sample are marked as nonfixed indicators $N_{tj}$.

The data input is as follows. Assuming time step $t$, the overall input data of the model for sample $j$ is $X_t$.
By dividing the input sequence data $X^T_i$ according to the fixed index and the nonfixed index, setting $X^T_i = [M^T_y, N^T_y]$, we obtain:

$$M^T_y = [x^{(1)}_{ij}, x^{(2)}_{ij}, \ldots, x^{(M)}_{ij}], N^T_y = [x^{(M+1)}_{ij}, \ldots, x^{(N)}_{ij}]$$

Note that sample $j = 1, 2, \ldots, k$ is a fixed index sample. The data are M-dimensional; the nonfixed index sample data are N-dimensional.

**ALGORITHM DESIGN**

Due to problems like high dimensions and insensitive supervision mechanisms, the design of the algorithm first considers decomposing and extracting the feature index (Astarabadi & Ebadzadeh, 2018; Lu & Bai, 2021). This decomposes the high-dimensional data index. The algorithm uses a DAE network for feature extraction for fixed indicators and a LSTM network for nonfixed indicators for feature extraction. Next, the algorithm fuses the indicators and inputs the extracted indicator combination into a fully connected network for training to complete the model prediction. The design of the model fully considers the problem that the fixed indicators are not sensitive enough to the supervision mechanism. It extracts the most representative information according to the characteristics of different types of indicators, retains the original data value, further compresses the calculation time, and reduces the computational complexity.

The fixed indicators include the borrower and general loan information. This necessary index of the loan cannot be changed within a period of time. The index, which reflects the inherent characteristics of the borrower, is relatively stable. Using the DAE network to train the data of the fixed indicator reduces the volume of the input data without losing the value of the original data. It also retains the features as much as possible (Kim et al., 2018; Peng et al., 2000). The nonfixed indicators include borrowers’ credit history. There will be changes within a period of time based on user behavior, period, environment, and other factors; therefore, the use of a LSTM network to extract nonfixed transaction indicators can improve the generalization performance of the model.

The DAE-LSTM algorithm design framework is shown in Figure 2.

*Figure 2. Framework of DAE-LSTM algorithm design*
Algorithm 1. DAE Network
Input: \( x \rightarrow x_{(1)}^{(1)}, x_{(2)}^{(2)}, \ldots, x_{(M)}^{(M)} \)
Output: \( Z \rightarrow Z_{(1)}^{(1)}, Z_{(2)}^{(2)}, \ldots, Z_{(M)}^{(M)} \)

a. Train the first autoencoder with the original input \( X \) and learn the first-order features of the original input.
b. First-order features are used as the input to the second autoencoder to learn the second-order features of the original input.
c. Take all second-order features as the input of the second autoencoder, learn the third-order features of the original input, etc. Obtain the fifth-order features.

Algorithm 2. LSTM Network
Input: \( x \rightarrow x_{(M+1)}^{(M+1)}, x_{(M+2)}^{(M+2)}, \ldots, x_{(N)}^{(N)} \)
Output: \( Z \rightarrow Z_{(1)}^{(N)}, Z_{(2)}^{(N)}, \ldots, Z_{(N)}^{(N)} \)

a. Concatenate the current input index with hidden state \( H \) of the previous node, multiply it by the weight matrix, and add the bias. Then, send it to the sigmoid activation function to obtain the output.// “forget gate.”
b. Activate the function to be the thin activation function. Join the output of the hidden layer of the previous node with the input index of the current node. Multiply by the weight matrix and add bias. Obtain the output through the sigmoid activation function.// “input gate.”
i. The output of the forget gate of the current node is a vector dotted with the cell state of the previous node.
ii. The two outputs of the input gate of the current node are dotted.
iii. The output of steps i and ii was added to obtain the cell state of the current node.// “update gate.”
c. Same operation as the forget gate. Concatenate the current input index with the hidden state \( H \) of the previous node, multiply by the weight matrix, and add the bias. Then, send it to the sigmoid activation function to obtain the output.
d. Obtain the output of the cell state of the current node from the tanh activation function.
e. The output of i and ii is dotted to obtain the hidden state of the current node.// “output gate.”

Algorithm 3. Fully Connected Network
Input: \( x \rightarrow Z_{(1)}^{(1)}, Z_{(2)}^{(2)}, \ldots, Z_{(M)}^{(M)}, Z_{(M+1)}^{(M+1)}, \ldots, Z_{(N)}^{(N)} \)
Output: \( Y \)

a. Join and output the features by algorithm 1 and algorithm 2.
b. Connect with features from each node in each layer of the network to extract the features extracted earlier.
c. Obtain the final predicted value \( Y \).
EMPIRICAL ANALYSIS

Data Processing

The test data of the experiment come from the database platform of the American Lending Club. The company is dedicated to personal consumption loans and SME loans. Listed in 2014, it became a leader in the financial and credit market industry. The analysis of its loan data will help tap the credit market risk control and early warning system. Its high research value can improve the enterprise loan risk control system.

The experimental data from Lending Club and data indicators are decomposed into fixed and nonfixed indicators. The fixed indicators include the borrower’s basic information (i.e., annual income, working years, housing nature, job information, and debt ratio) and loan basic information (i.e., loan interest rate, loan application amount, and borrowing cycle). This is a total of eight indicators. Nonfixed indicators include credit history transactions (i.e., number of overdues, time of first opening credit card, loan amount, number of inquiries, number of months since last default, months since last negative public record, outstanding amount, credit account usage rate, credit limit, total credit limit, total debt amount, borrowing status, and number of negative public records). This is a total of 13 indicators. The dependent variable is risk level (higher the level is lower the risk).

A total of 274,390 pieces of data were screened for analysis in the experiment. The names of the data indicators and descriptive statistics are listed in Table 1. The risk level is normalized and converted into a continuous value. After the risk level is normalized, the distribution of the number of sample values in different risk value intervals is shown in Figure 3.

Table 1. Indicator name and descriptive statistics

<table>
<thead>
<tr>
<th>Variable type</th>
<th>Indicator name</th>
<th>Mean</th>
<th>Std</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent v</td>
<td>credit risk level</td>
<td>5.3134</td>
<td>1.3611</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Fixed indicators</td>
<td>annual income ($)</td>
<td>70772</td>
<td>31.23</td>
<td>4000</td>
<td>6000000</td>
</tr>
<tr>
<td></td>
<td>working years</td>
<td>5.4864</td>
<td>3.5496</td>
<td>0.5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>debt ratio (%)</td>
<td>14.8280</td>
<td>7.2923</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>loan interest rate (%)</td>
<td>13.1149</td>
<td>4.3376</td>
<td>5.42</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>loan application amount ($)</td>
<td>14381</td>
<td>29.8500</td>
<td>500</td>
<td>35000</td>
</tr>
<tr>
<td></td>
<td>borrowing cycle (mth)</td>
<td>41.7430</td>
<td>10.2397</td>
<td>36</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 1 continued on next page
Due to incomplete and unbalanced data, this article fills in the missing data and transforms the date data. At the same time, according to the observation of the distribution map of sample values, it is shown that the risk value distribution of the original data is unbalanced. According to Figure 3, the number of samples with a credit risk value below 0.6 accounted for 33.53%, samples below 0.7 accounted for 49.21%, and samples below 0.8 accounted for 83.44%. It set the threshold and performed sampling according to the sample distribution to address the problem of unbalanced sample distribution. In the experiment, the threshold is set as the critical value of the credit level to judge and perform balanced sampling in the threshold interval. Then, the experimental dataset is divided into a training set and a test set by tenfold cross-validation.

Figure 3. Sample size distribution of credit risk measurement values
IMPLEMENTATION STEPS

The platform is implemented using PyCharm 3.9, Anaconda integrated library, and the TensorFlow 1.2.2 framework. Pycharm is an excellent Python IDE for international use due to its universal applicability in the field of deep learning development. Anaconda integrates common Python libraries for data science. It includes a conda package management system designed to solve software environment dependencies that can quickly and independently build different versions of the TensorFlow deep learning framework. After establishing the association between Pycharm and the Anaconda environment, the construction of a deep learning framework and data processing can be made more convenient and efficient. The following subsections describe the steps.

BUILD DAE COMPRESSION FIXED INDICATOR INFORMATION

The designed DAE has five layers. The fixed index data of the input layer are converted to the middle-hidden layer and then to the output layer after three hidden layers. The data dimension of the output layer is as close as possible to the data dimension of the input layer. The autoencoder network extracts and describes the input features. As the level increases, the features become increasingly abstract and global. The output of the hidden layer of the last layer has the overall characteristics of the input data.

DESIGN LSTM FOR FEATURE SCREENING OF NONFIXED INDICATORS

The LSTM network can remember indicators that have an impact on the results and forget indicators that do not help the results. Feature screening of nonfixed indicators is performed by designing three “gates” in the LSTM network. For loan indicator information, valid features like loan year and number of inquiries can be retained. Information that has less impact on the current can be discarded. It sets the number of input gate nodes of the LSTM memory network as $N - M - I = 13$. This corresponds to the nonfixed index information of 13 loans.

After filtering the features by combining the DAE-LSTM algorithm, the fixed and nonfixed indicators are fused to obtain the output data of the new index combination. This is reinput to the fully connected network for training. The risk of credit loans is measured through the three-layer fully connected network.

TRAINING METHOD AND OPTIMIZER SELECTION

Regarding the process of training the model, selecting the appropriate training method and optimizer can improve the prediction efficiency and accuracy. In this article, the minibatch method is selected to train the LSTM neural network. The goal is to predict the continuous value of credit risk. Therefore, the mean square error (MSE) is selected as the loss function. Regarding the optimizer, to make the LSTM neural network converge quickly, this article adopts the adaptive moment estimation (Adam) optimizer for optimization training. The Adam optimizer, proposed by Kingma and Ba (2014), is one of the most commonly used optimization algorithms. Unlike traditional optimization algorithms that usually set the learning rate to be constant, the Adam algorithm updates the learning rate during the optimization process. Therefore, the learning effect is more effective and the convergence speed is faster.

EXPERIMENTAL RESULTS

This section conducts an empirical analysis through the lending club data to compare and test the prediction effects of GRNN, basic LSTM, and combined DAE-LSTM algorithms. The proposed
DAE-LSTM algorithm is compared with the traditional GRNN and LSTM algorithms for five experiments (see Table 2). The MSEs of the three models are 0.0540, 0.0684, and 0.0818. The threshold is set to discretize the dependent variable. When \( f(x_i) > \text{thre} \), the selected sample is considered to have a good credit rating and a low credit risk \( (f(x_i) = 1) \). When \( f(x_i) < \text{thre} \), it is considered that the selected sample has a poor credit rating and a high risk \( (f(x_i) = 0) \). The average hit rates were 93.31%, 88.34%, and 85.04%. The average false alarm rates were 2.81%, 3.63%, and 3.45%. The average accuracy rates were 95.2%, 92.2%, and 89.4%. All four indicators showed that the DAE-LSTM algorithm has higher prediction accuracy, better model generalization performance, and better feature extraction than the single GRNN and LSTM models.

### Table 2. Comparison of experimental test results

<table>
<thead>
<tr>
<th>Group No.</th>
<th>ACC</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GRNN</td>
<td>LSTM</td>
</tr>
<tr>
<td>1</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>2</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>3</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>4</td>
<td>0.87</td>
<td>0.91</td>
</tr>
<tr>
<td>5</td>
<td>0.90</td>
<td>0.94</td>
</tr>
<tr>
<td>mean</td>
<td>0.8940</td>
<td>0.9220</td>
</tr>
</tbody>
</table>

Stratified sampling is adopted for the large sample data of the test to better display the graphical comparison effect. This means that the data are stratified to extract part of the sample to draw the graph. The threshold list of risk assessment is set as \( = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] \). According to the threshold value within the range \([0, 1]\), which is divided into 10 interval layers using the geometric distribution method, let the ratio of the number of samples drawn in each layer to the total number of this type be the same. A total of 1,800 interval samples are drawn in layers to calculate and visualize the average fitting value and MSE value of the sample. Refer to Figures 4-9 for a comparison of the effects.

1) GRNN effect observation
Figure 4. GRNN-Experimental prediction effect

Figure 5. GRNN-Comparison of experimental

Figure 6. LSTM-Experimental prediction effect
Figure 7. LSTM-Comparison of experimental

Figure 8. DAE-Experimental prediction effect

Figure 9. DAE-LSTM Comparison of experimental test results
DISCUSSION

Based on the historical data of the lending club, this article will use the DAE-LSTM algorithm to analyze the importance of the extracted features and calculate the information gain $Gain(D,A_i)$ by extracting the feature factors in the model.

\[
Gain(D,A_i) = H(D) - H_{A_i}(D)
\]

Notes that the entropy $H(D) = -\sum P_i \log_2 P_i$, condition entropy $H_{A_i}(D) = -\sum Q_{ij} \log_2 Q_{ij}$, for the sample set, the value of characteristics $A_i (i=1,2,...,n)$ is $X_j (j=1,2,...,k)$, and $Q_{ij}$ are the proportion of value $X_j$ under condition $A_i$.

The greater the information gain, the greater its importance in risk prediction (Zhu, 2021). Through calculation, three factors of fixed indicators (borrowing period and loan interest rate) and nonfixed indicators (loan amount) can obtain a larger information gain (the information gain values of the interest rate factor, term factor, and loan factor are 0.4320, 0.3578, and 0.3026, respectively). These three indicators have a strong correlation with the level of credit risk.

In the experiment, the large sample data are divided into 10 interval samples according to the credit risk level (see Table 3). In addition, the number of samples in each interval, average loan interest rate, average loan period, and average loan amount are obtained. When analyzing the relationship between the interest rate factor of the online lending platform and the borrower’s credit risk level, this study found that 43.85% of the loan projects’ interest rate pricing tends to be reasonable. However, the low-risk borrower loan programs (Groups 6 and 8) have lower loan amounts. The average interest rates paid by the optional loan programs are lower (the average interest rates paid are 0.0714 and 0.1855). The loan period is usually shorter. High-risk borrower loan projects (Groups 3 and 5) have higher loan amounts; however, the average interest rate paid by the loan projects that can be selected is higher. At the same time, they need to bear higher interest rates (the average interest rate is 0.8463 and 0.8124); the loan period is usually longer. The industries of borrowers with higher risk are mostly low-end service industries and other service industries with relatively high employee mobility. The industries of borrowers with lower risks are mostly labor-intensive industries like agriculture and manufacturing. The theoretical interest rate is not very different from the real interest rate.

<table>
<thead>
<tr>
<th>Group No.</th>
<th>Risk level</th>
<th>Interest rate</th>
<th>Time</th>
<th>Loan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gain</td>
<td>---</td>
<td>0.4320</td>
<td>0.3578</td>
<td>0.3026</td>
</tr>
<tr>
<td>1</td>
<td>0.2-0.3</td>
<td>0.7350</td>
<td>0.7527</td>
<td>0.5253</td>
</tr>
<tr>
<td>2</td>
<td>0.3-0.4</td>
<td>0.6367</td>
<td>0.6489</td>
<td>0.5004</td>
</tr>
<tr>
<td>3</td>
<td>0.0-0.1</td>
<td>0.8463</td>
<td>0.9103</td>
<td>0.6023</td>
</tr>
<tr>
<td>4</td>
<td>0.4-0.5</td>
<td>0.5621</td>
<td>0.4671</td>
<td>0.4479</td>
</tr>
<tr>
<td>5</td>
<td>0.1-0.2</td>
<td>0.8124</td>
<td>0.8571</td>
<td>0.5750</td>
</tr>
<tr>
<td>6</td>
<td>0.9-1.0</td>
<td>0.0741</td>
<td>0.0072</td>
<td>0.3316</td>
</tr>
<tr>
<td>7</td>
<td>0.5-0.6</td>
<td>0.4822</td>
<td>0.3006</td>
<td>0.3876</td>
</tr>
<tr>
<td>8</td>
<td>0.8-0.9</td>
<td>0.1855</td>
<td>0.0616</td>
<td>0.3508</td>
</tr>
<tr>
<td>9</td>
<td>0.7-0.8</td>
<td>0.3049</td>
<td>0.1153</td>
<td>0.3470</td>
</tr>
<tr>
<td>10</td>
<td>0.6-0.7</td>
<td>0.3927</td>
<td>0.2523</td>
<td>0.3744</td>
</tr>
</tbody>
</table>
Research shows that low-risk borrowers usually choose a shorter loan term and obtain a lower interest rate. High-risk borrowers usually choose a longer loan period and obtain a higher interest rate. The longer the investment project cycle, the higher the risk of uncertainty. This is different from previous research. As another option that the borrower can choose independently in addition to the interest rate, the loan term can reflect some implicit information so that the loan term is related to the credit default to some extent. For borrowers who need to meet capital turnover needs, the longer the loan period they choose, the poorer the borrower’s evaluation of their overall information (Klafft, 2008). However, this empirical analysis found that the loan period will also be extended due to the large loan amount. The expectation of future capital return will, thus, be reduced. Therefore, it cannot be used as a single option to measure the level of credit risk. The application of the deep learning combination algorithm for overall analysis will obtain a more accurate prediction effect. In summary, a deep study of the relevant factors of credit risk in the financial and credit industry, as well as a deep study of the relationship between loan project interest rates and credit risk levels, can provide a reference for the heterogeneity analysis of industry lending and interest rate pricing benchmarks.

CONCLUSION

Credit risk measurement of the financial credit industry often faces the problem of processing high-dimensional data. How to properly screen and compress characteristic indices and apply them to algorithms is an important basis for model design. Due to the poor screening effect and strong subjectivity of traditional algorithms (Ahn & Sura, 2020), this article used an unsupervised DAE encoder network for data compression and LSTM network for memory index screening. It made full use of the idea of discriminating according to different indices. The DAE-LSTM combined algorithm can be used to process high-dimensional data indices and process measure analysis. This can retain the inherent characteristics of indices, effectively utilize the advantages of information resources, and improve the computing speed and accuracy.

The results showed that the prediction accuracy of the DAE-LSTM algorithm is 95.2%. The average accuracy of the DAE-LSTM algorithm is 6.49% and 3.25% higher than that of a single generalized regression neural network and LSTM network, respectively. By applying the deep learning combination algorithm to the risk measurement of the financial credit industry, the behavior of the financial credit industry can be analyzed more scientifically and accurately, the possible financial credit risks can be forecast, and the occurrence of risks can be reduced.

The traditional financial industry often requires borrowers to provide detailed personal information and proof of repayment ability. Therefore, the assessment of personal credit is complicated. Compared with traditional finance, with the help of the internet, the modern financial and credit industry service model is convenient. In addition, the information processing is fast. Moreover, the modern financial credit industry has simple information requirements for borrowers and provides more simplified lending materials and procedures (Moreno-Enguix et al., 2019; Shang et al., 2014). It can absorb a large number of long-tail customers and divert the marginal profits of traditional finance, which requires a high-precision credit risk measurement algorithm to avoid misjudgment of high-risk users. However, it also increased its own credit risk. The application of data mining technologies like deep learning can solve the problem of traditional financial information asymmetry in some aspects, break the blind spot of credit risk judgment, and reduce the probability of financial credit risk.

In summary, the application of deep learning technology in financial credit will help the credit industry to improve information acquisition, processing, and application capabilities. It will alleviate the problem of information asymmetry and bring its credit services more in line with the needs of the development of the financial industry. At the same time, the combined application of deep learning technology has improved the risk management and control capabilities of the financial and credit industry while providing technical support for the flow of credit resources to credit loans and long-term loans with greater risks and higher technical requirements.
LIMITATIONS

In the risk measurement study of the financial credit industry, the level of economic development has a certain impact on the risk of credit default of borrowers. Therefore, it is appropriate to introduce variables reflecting the macroeconomic level, expand the number of samples, and consider the imbalance and heterogeneity of samples. The next step is to build a more dynamic and comprehensive credit risk identification model.

ACKNOWLEDGMENT

This work was supported by the National Social Science Foundation of China (21BTQ059) (Research on Information Behavior Preference Feature Mining and Recommendation Based on Users’ Cross-Social Media).
REFERENCES


Ermpinis (2021).


