Optimization of financial data processing strategy based on LSTM-GCN

Wenjia Wu[1,*], Minglei Han[2], Yandong Hu[3], Jiaqi Ma[4], Xiaolei Zhang[5]

[1]School of Finance, Harbin University of Commerce, Harbin, China

[2]Harbin University of Commerce Talent School, Harbin, China

[3]School of Computer and Information Engineering, Harbin University of Commerce, Harbin, China

[4]School of Finance and Public Relations Management, Harbin University of Commerce, Harbin, China

[5]School of Economics and Management, Shandong Jiaotong University, Jinan, China

* Corresponding Author Email: m18964799825_1@163.com

Abstract. In the process of training LSTM-GCN mixed model with labeled data, a common practice is to divide the data set into a training set and a validation set, and cross-validation techniques are frequently used to ensure the effective partitioning of data. The training phase mainly involves the selection of optimization methods such as gradient descent to update the model parameters, so that the model can correctly identify the feature patterns in the data. In this process, we first define a loss function to quantify the difference between the model's predicted value and the actual value. For example, when dealing with classification problems, the cross-entropy loss function is often used; Whereas in regression problems, the mean square error loss function or the absolute error loss function is more common. Then, the loss gradient of each parameter is calculated by backpropagation algorithm, and optimization strategies such as gradient descent are applied to adjust the parameter values. Through repeated iterations, these algorithms aim to minimize the loss function as a core goal for training the model. In addition, the details of model training, such as setting the appropriate learning rate, taking the appropriate regularization measures, and choosing the appropriate weight initialization strategy, are also key steps that cannot be ignored. This paper conforms to the research and development trend of the industry, focusing on how to use the LSTM-GCN model to capture temporal and spatial characteristics, process multi-modal data, process nonlinear data and dynamic weight allocation, etc., providing new ideas for Internet finance companies in data processing. This model combines the advantages of LSTM and GCN, and has prominent advantages such as timeliness, strong representation learning ability, associability, and interpretability. It can process both time series data and graphic data, and can analyze and mine the correlation and regularity in large-scale time series data, so as to process financial data more accurately. It greatly reduces the risk of investment decision.

Keywords: LSTM-GCN model, financial investment decision, financial data processing

1. Existing financial data processing methods

1.1. Based on regression model and time series model

The financial data processing method based on regression model and time series model is to establish a mathematical model through curve fitting and parameter estimation based on the time series data obtained from systematic observations.

1.2. Financial data processing model based on deep learning technology

Deep learning can process a large amount of data quickly, provide accurate prediction and decision support, and complete automated tasks with high efficiency. Deep learning technology can be roughly divided into data preparation, model construction, model training, model evaluation, model application, optimization and improvement, and other steps, and these steps are not strictly linear relationship, but a continuous iterative process.

Deep learning technology has a wide range of applications in the field of financial data processing. For example, a stock prediction model based on deep learning and convolutional neural network proposed by researchers at the University of California, Berkeley is used to predict the trend and volatility of future stock prices. The model processes a large number of historical stock price data, news information and other information. To establish a deep neural network model, which can accurately predict the stock trend and volatility and reduce investment risk and improve decision-making efficiency; Researchers from the University of Tokyo in Japan developed a high-frequency trading system based on deep learning and reinforcement learning technology for predicting changes in stock prices and trading volumes, the system can process market data and information in real time, and use deep neural network algorithms for stock prediction and trading decisions, this system improves trading efficiency and accuracy, but also reduces trading costs and risks.

In financial data processing, the following kinds of deep learning technologies are often used:

a) Recurrent neural networks (RNN)
b) Convolutional neural networks (CNN)
c) Reinforcement Learning (RL)
d) Long Short-term Memory Network (LSTM)

2. Architectural models

This paper proposes a new financial data processing and model training system and device, which belongs to the technical field of data processing. The system utilizes three main deep learning models -- Long short-term memory network (LSTM), Graph Convolutional Network (GCN), and other deep learning strategies to enhance the efficiency and intuitiveness of financial data processing. The LSTM model is able to effectively manage the flow of information through its unique three-gate control mechanism (forget gate, input gate and output gate) and the design of the cell state. The GCN model, on the other hand, captures complex data relationships through deep learning of graphical data. Combining the advantages of LSTM and GCN, this system provides an efficient and intuitive solution for financial data processing, while reducing the complexity of deep learning technology implementation.

2.1. Model Design

a) LSTM model

Three gating mechanisms are used inside LSTM to delete and increase the information of neurons, which are respectively: forgetting gate f, input gate i and output gate o (t indicates that the LSTM unit is at time t). Among them, the forgetting gate is used to control which information needs to be deleted by the current neuron. The specific calculation process of the forgetting gate is as follows:

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right)$$

The input gate is used to control how much information received by the current neuron can be retained in the current cell state. The state update calculation process of the input gate is as follows:

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right)$$

$$\tilde{C}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t$$

The output gate controls how much information the current neuron can output to the next moment, and the output gate status update calculation process is as follows:

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t \times \tanh\left(C_t\right)$$

The above formula represents the activation function of neural network σ sigmoid function, which maps variables into vectors between [0,1] according to the input; $\tilde{C}_t$ Represents the information in the candidate neurons; W_f, W_i, W_o, W_c Represents the weight in the calculation process of LSTM neuron state update; b_f, b_i, b_o, b_c Represents the bias in the calculation process of LSTM neuron state update. The introduction of these gates allows the LSTM network to process sequence data more accurately and retain information across time.

Since the RNN will pass the extracted information to the next unit during training, the information will disappear when the length of the chain accumulates to a certain extent. The appearance of LSTM is to solve the problem of gradient disappearing and gradient explosion in the process of long sequence training. It has a gating mechanism (three gates can selectively filter the old information, the new information and all the information of the current unit state), which can effectively control the propagation of gradient in time and space, so as to alleviate the problem of gradient disappearance and make the model easier to train and converge.

In addition, compared with conventional RNNS, LSTMS can perform better in longer sequences. The LSTM requires four linear layers (MLP layers) per cell to run in each sequence-time step. Support for long-term memory is achieved by maintaining an internal state called the "cell state", which allows long sequence data to be processed efficiently and has long-term memory capability. The linear layer requires a lot of storage bandwidth to compute, but the cyclic structure of the LSTM can be unrolled into a feedforward neural network, which allows efficient parallel computation on hardware such as Gpus, speeding up the training and reasoning of the model.

In addition, the gating mechanism in the LSTM can also adaptively choose to retain or discard the input data, so it has a good effect on problems that need to deal with long-distance dependencies.

To sum up, this project uses LSTM basic technology to process highly complex and high-dimensional financial data, which can more efficiently process historical financial data with huge data volume and make time series prediction, thus improving model performance and efficiency and accuracy of financial risk prediction.

b) GCN model

The purpose of GCN is to use convolution to extract the spatial information and attribute information of non-European structured data, and to dig deeply into the feature laws in the graph model. According to the definition of Kipf et al., given an undirected graph G=(V,E,A) is composed of A set of nodes and an edge set E,A is an adjacency matrix, where $A \in R$, $input^{NxN}$ changes

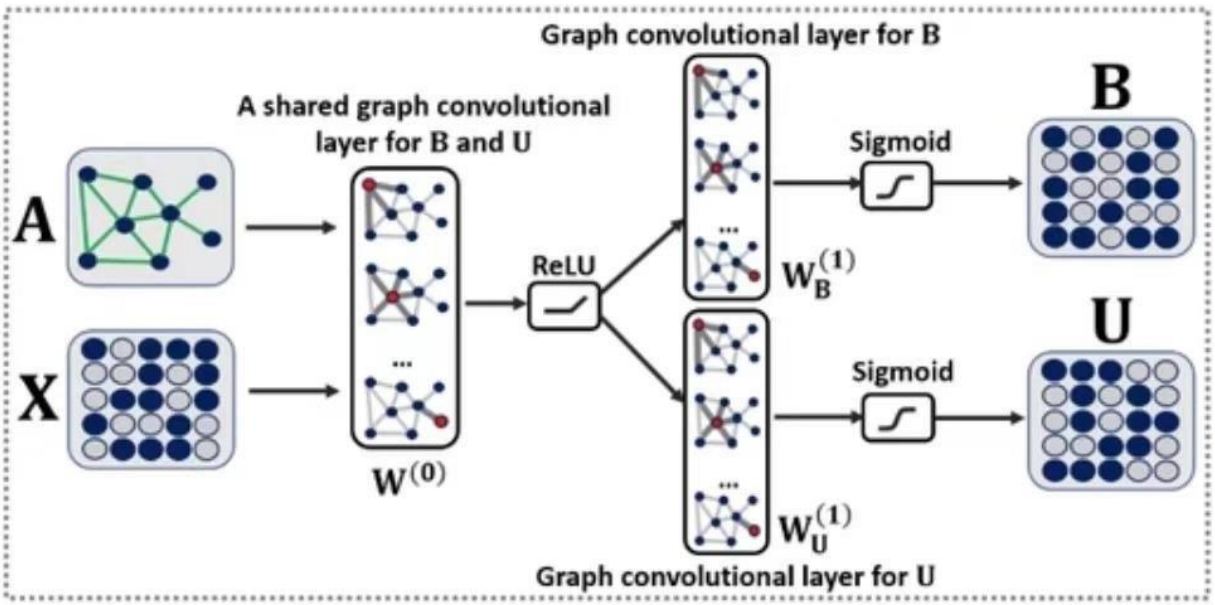

Figure 1. The processing method taken by the graph convolutional neural network

Quantity x and output variable y, the processing method taken by the graph convolutional neural network is shown in equation (1), and the forward propagation formula of the graph convolutional neural network is shown in equation (2), where = $\tilde{A}$A+i, i is the identity matrix of size N*N: is the degree matrix of undirected graph; $\tilde{D}$h $^{(L)}$ $\in$R^{NxD} represents the output value of the L-th layer: W $^{(L)}$represents the parameter value of the L-th layer; σIs the activation function.

$$f(X, A) = Y \tag{1}$$

$$(H^{(l+1)} = f(H^{(l)}, A) = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})2)$$

In GCN, each node is regarded as a feature vector, and the feature vector of the node is updated through the information exchange between neighboring nodes. Specifically, GCN represents the input graph data as an adjacency matrix, which records the relationship between each node and its directly connected neighbors. The nodes are then fused and updated by means of the convolution operation, which is defined based on the adjacency matrix.

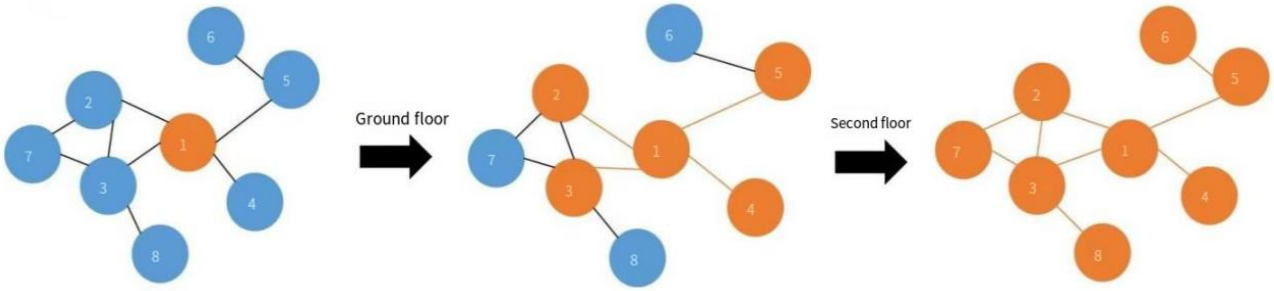

Figure 2. Schematic diagram of GCN spatial feature extraction

As shown in the figure, the essence of the graph convolutional neural network is to spread the features of each node to the next layer after weighted average of the feature information of its connected nodes, and with the deepening of the number of layers, the node information that can be aggregated to each node is further, so as to represent the structural features of the entire graph model and carry out the next operation.

c) LSTM-GCN combinatorial modeling

Data preparation: First of all, sequence data and graph data need to be prepared. Sequence data is usually time series or serialized data, such as text or audio. Graph data is a graph structure made up of nodes and edges that can represent various types of relationships. In this stage, the historical financial data and real-time financial data corresponding to the business parameters of real-time events in the financial industry will be obtained, and the historical financial data and real-time financial data will be correlated and processed to get the resulting financial data. This stage mainly includes the following specific tasks:

① Data collection: Obtain raw data, which can be done from databases, files, apis, etc.

② Data cleaning: dealing with errors, missing values, outliers and other problems in the data to ensure the quality of the data.

③ Data conversion: Format conversion, normalization and other operations are carried out on the data to facilitate subsequent exploration and analysis.

④ Feature extraction: Machine learning algorithms are used to extract useful features from the data.

⑤ Feature selection: Select the feature with the most predictive performance from all the features.

⑥ Data set partitioning: Dividing the data set into training sets, verification sets, test sets, etc., according to a certain proportion.

⑦ Data set sampling: For particularly large data sets, random sampling, layered sampling and other methods can be used to sample the data.

⑧ Data visualization: visualization of data through charts, interactive graphics, etc., so that people can better understand the data.

In the process of data preparation, it is necessary to combine the actual financial data situation, improve and adjust through practice to ensure that the final data set can meet the requirements of use.

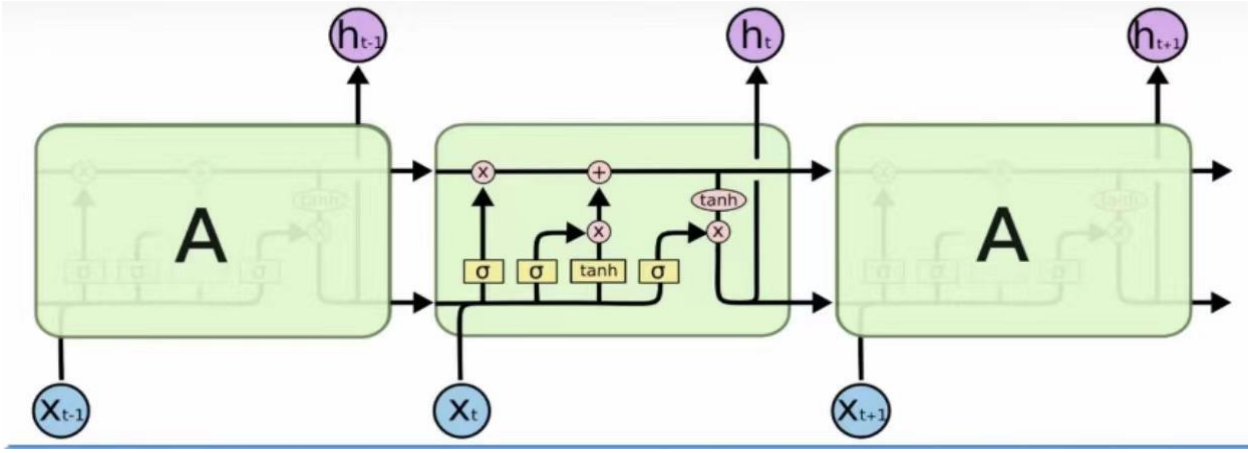

Figure 3. Schematic diagram of LSTM internal modules

2.2. Sequence coding and diagram coding

In the modeling process, the LSTM network is used to encode the sequence data, map the sequence data to a vector representation of fixed dimensions, and the GCN network is used to encode the graph data, fuse the results of sequence coding and graph coding, and pass the fused feature vector to the subsequent model layer for processing, that is, the source mode data is encoded as the implicit feature z. Which is then used by the decoder to generate the target mode.

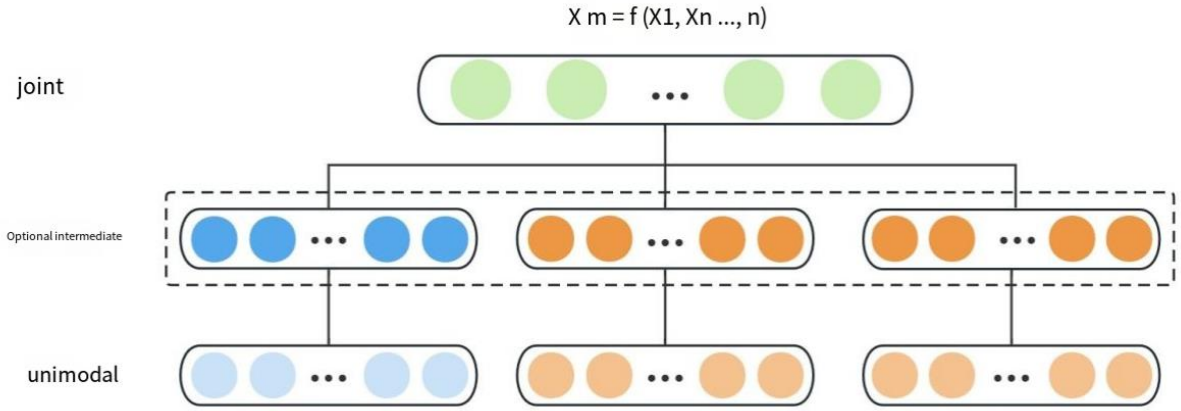

Figure 4. Schematic diagram of encoding

2.2.1. Feature passing

The fused feature vector is passed to the subsequent model layer, and the full connection layer, Softmax layer, etc., is used for classification, regression or other tasks.

When performing the classification task, the model will input the comprehensive feature vector into the fully connected layer, and perform a series of linear transformations and nonlinear activation operations in the fully connected layer to obtain the output result of the model. The output results of the model are then fed into the Softmax layer for prediction of the target category. Among them, the Softmax layer is usually used to map the model output results to a probability distribution, with the sum of the probability values of the various categories being 1, to facilitate the prediction of the target category.

When the regression task is performed, the comprehensive eigenvector is input into the fully connected layer, where a series of linear transformations and nonlinear activation operations are performed to obtain the output results of the model. Then, according to the specific task requirements, different loss functions can be selected to compare the model output and the target value, so as to obtain the training loss of the model and carry out backpropagation. Finally, the trained model is used to make predictions on the new data.

2.2.2 Model training

Use labeled data to train LSTM-GCN model. By using optimization algorithms such as gradient descent to update the parameters, the model can learn the appropriate feature representation.

When using labeled data to train LSTM-GCN model, the data set is divided into training set and validation set, which are usually divided by cross-validation method. Then, the parameters are updated using optimization algorithms such as gradient descent so that the model can learn a suitable feature representation. Specifically, the error between the model's predictions and the true values will be evaluated by constructing a loss function. For example, a cross entropy loss function may be used for a classification task, and a mean square error loss function or an absolute error loss function may be used for a regression task. Then, a backpropagation algorithm is used to calculate the gradient of the loss function for each parameter, and optimization algorithms such as gradient descent are utilized to update the parameters of the model. Among them, the gradient descent algorithm minimizes the loss function by iteratively updating the model parameters, thus achieving the purpose of training the model. In addition, in the training process, it is also necessary to pay attention to some details, such as setting the learning rate, regularization, and initializing the weights.

2.2.3. Output prediction

In this model, LSTM neural network is used to extract the historical information of time series, and GCN neural network is used to extract the topological relationship between different assets in the financial market, so as to extract the temporal and spatial characteristics. In the training process, the model will predict the future data according to the patterns and rules learned from the historical data, so as to obtain the prediction result of financial risk. The prediction results can be obtained after the model is trained and the model is applied to make the prediction. And the model will need to interpret the predicted results to understand the situation of financial risk prediction. Finally, according to the forecast results, sort out the corresponding report or document, and output the result to the user. These output results include the predicted value, confidence and other related information, which helps users to understand the actual situation of financial risks, and formulate corresponding risk control and management strategies.

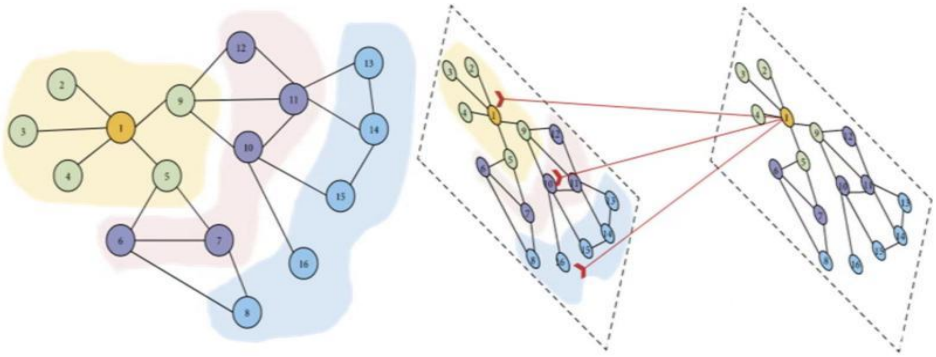

Figure 5. Schematic diagram of LSTM-GCN model

2.3. Model workflow

LSTM-GCN model can comprehensively consider a variety of factors (such as financial data, news public opinion, policies and regulations, etc.), so as to predict possible risk events in the future. The model obtains data related to risk events by collecting, pre-processing and constructing time series graph, then constructs GCN according to the time series graph, and then takes the node features of

GCN output as the input of LSTM to learn the dynamic characteristics of data on time series. Finally, the LSTM-GCN model that has been trained by the model is used to predict the possible risk events in the future, and the corresponding risk avoidance and control decisions are made according to the prediction results.

3. Model testing and analysis

Measurement of model capability test data based on Accuracy, Precision, Recall and AUC equivalence

In-depth analysis of the causes of financial risks plays an important role in understanding, identifying, evaluating and managing risks. Based on the above cognition of the causes of financial risks, we can identify financial investment risks from two aspects: systemic risk and non-systemic risk.

(1) Systemic risk. It mainly includes policy and regulation risk, capital and financial market risk and economic operation risk.

(2) Non-systemic risk. It is mainly related to the invested enterprise and the invested project itself, including credit risk, operational risk and moral risk.

Based on the causes and identification of financial investment risk, this paper selects multiple indicators involving market risk, liquidity risk, management risk, moral risk and joint risk as alternative risk indicators from both qualitative and quantitative levels

This paper uses the actual case data to conduct empirical evaluation and analysis of the LSTM-GCN financial risk prediction model. The data set adopts the financial data disclosed by Orient Wealth and investing, etc., and tests the model through the four financial risk events of Evergrande incident, Credit Suisse crisis, bankruptcy of Silicon Valley Bank, collapse of Credit Suisse and Lehman Brothers.

4. Model comprehensive evaluation

The experimental results show that the proposed GCN-LSTM-CoPoD anomaly detection method is superior to the above methods in terms of Accuracy, Precision, Recall and AUC value on the MBD data set. In addition to the improved accuracy rate, precision rate and recall rate to some extent, compared with other methods, the proposed method can obtain better AUC value, which is about 27%~40% higher than other methods, indicating that the anomaly oriented effectiveness and detection value of the proposed method have been improved. As shown in FIG. 9, ROC curve can well demonstrate the anomaly detection ability of various methods, and the cloud service timing anomaly detection algorithm proposed in this paper is superior to other methods, with high test accuracy.

Compared with LSTM model, the accuracy of LSTM-GCN model is 0.12 higher, which indicates that the number of samples correctly predicted by this model accounts for a high proportion of the total number of samples; The accuracy rate was higher by 0.0162, indicating that the number of samples correctly predicted as positive by the model accounted for a higher proportion of all the samples predicted as positive, and the model had a stronger ability to judge positive samples; The recall rate is higher by 0.0307, which indicates that the number of samples correctly predicted as positive by this model accounts for a higher proportion of all samples that are actually positive, and the model has a stronger ability to identify positive samples; A higher AUC value of 0.0202 indicates that the LSTM-GCN model has better classification performance and stronger ability to correctly distinguish positive and negative samples. In summary, compared with LSTM model, this model has better performance, higher accuracy in financial risk prediction, and can provide investors and relevant institutions with more accurate prediction results. LSTM-GCN model has stronger processing power and expression ability. Compared with only using the LSTM model, the LSTM-GCN model can consider the relationship between nodes in the sequence data, and use the graph

convolutional neural network for data preprocessing and feature extraction, which enhances the model's ability to learn information such as connectivity and structure between nodes. At the same time, the LSTM-GCN model can also use the learnable weights in the GCN model to update the relationships between nodes, so as to further improve the accuracy and generalization ability of the model.

Similarly, compared with the GCN model, the LSTM-GCN model has a higher accuracy of 0.11, which indicates that the model has better classification effect; The accuracy rate is higher by 0.0251, which indicates that the model has stronger ability to judge positive samples; The higher recall rate is 0.1018, indicating that the model has stronger ability to identify positive samples; A higher AUC value of 0.0951 indicates that the model has better classification performance and stronger ability to correctly distinguish positive and negative samples. In summary, compared with the GCN model, the LSTM-GCN model has better performance, can consider the time sequence information, and can adapt to more complex dynamic changes; The performance is more stable under the condition of data missing and noise interference; And better performance with smaller training sets.

Then LSTM-GCN model is compared with ARIMA model and random forest model.

LSTM-GCN model and ARIMA model are both commonly used time series prediction algorithms. LSTM-GCN model can consider both time series information and graph structure information, and is suitable for dynamic change modeling. The ARIMA model is a classical linear model based on time series difference and autoregressive moving average, which is suitable for stable time series. According to the above experiments, the accuracy rate of LSTM-GCN model is 0.16 higher than ARIMA, the accuracy rate is 0.0152 higher, the recall rate is 0.033 higher, and the AUC value is 0.0786 higher. It can be seen from the analysis that LSTM-GCN model can adapt to more complex dynamic changes, and its performance is more stable under the condition of data missing and noise interference. Therefore, it performs better in application scenarios such as time series prediction and dynamic change modeling. At the same time, accuracy rate and recall rate represent the accuracy and coverage of model prediction results, while AUC can reflect the overall performance of model classification prediction. Therefore, it is more accurate to use LSTN-GCN model for financial risk prediction. Through research and analysis, it can be seen that the reasons are as follows: ARIMA model can only process simple linear time series data, and completely rely on the past time series data to predict the future trend of change, unable to take into account the correlation between different assets. On the other hand, LSTM-GCN model can use LSTM network to extract features from time series data, and use GCN network to model the relationship between nodes, which can better process nonlinear and multi-variable time series data. Meanwhile, The LSTM-GCN model can extract more accurate feature representations by constructing graph models and using topological structures between nodes. Moreover, the model is more adaptable to new data, more robust in dealing with outliers and other problems, and can better predict financial risks in unknown environments.

Compared with the random forest model, the accuracy rate of the LSTM-GCN model is 0.14 higher than that of the random forest model, the accuracy rate is 0.0258 higher, the recall rate is 0.0337 higher, and the AUC value is 0.0951 higher. According to the above data analysis, compared with the random forest model with overfitting problems, LSTM-GCN model has higher accuracy rate, accuracy rate, recall rate and AUC, and performs better in time series prediction for specific data sets. The LSTM-GCN model can combine the characteristics of LSTM network and GCN network to better deal with non-linear and multivariate time series data. It can also extract more accurate feature representation through the topological structure between the modeling nodes of graph neural network, so as to accurately predict financial risks.

5. Conclusions

Through the above analysis, we can see that the field of financial risk prediction is faced with many challenges and difficulties, and the traditional financial data processing and risk prediction methods have obvious disadvantages. This paper aims to combine professional knowledge and

technical means such as machine learning, cloud computing and data mining to study a new method and new idea of financial data processing and risk prediction based on LSTM-GCN model, so as to improve the accuracy and effectiveness of risk prediction results and bring great technological progress and value enhancement to the financial investment field. To provide better investment decision reference for the majority of investors, and promote the digital transformation and upgrading of the financial industry.

References

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[2] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. *arXiv preprint arXiv:1706.02216*.

[3] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.

[4] Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*.

[5] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.