
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
 Volume 25, Issues 4, pp. 2567–2574, DOI 10.12694/scpe.v25i4.2846

STOCK QUANTITATIVE INTELLIGENT INVESTMENT MODEL BASED ON MACHINE LEARNING ALGORITHMS

KANGYI WANG*

Abstract. In order to cope with sudden changes in market style, the author designed and implemented an investment system. In response to the inability to cope with sudden changes in market style, the author proposed an improved scheme when using a classifier for training, using a hidden Markov model to select time points in the same market state as the training samples for the classifier. The results of backtesting and real market testing show that the improved investment system can capture changes in market style, and during the testing period, the test results obtained higher returns compared to the pre improved classifier. At present, the investment system implemented by the author has become the core backend module of the officially launched intelligent investment advisory product Zhiyu Liangtou APP, providing strong investment decision-making suggestions for small and medium-sized investors. In the future, we can provide backend support for more user end products.

Key words: Machine learning algorithms, Stocks, Quantification, Intelligent investment model

1. Introduction. In recent years, with the continuous growth of people's wealth and the improvement of financial awareness, stock investment is receiving more and more attention. People investing in stocks can not only achieve diversified asset allocation, but also share the dividends of the development of the real economy. In the investment process, people try to obtain the highest possible returns while taking the smallest possible risks [1]. The use of manual methods for stock trading has many drawbacks, which can greatly discount the investment results. For example, human emotions during manual operations, the speed and ability of the human brain to process information, and the ability to evaluate risks and returns all limit the level of investment.

With the development of computer tools, people are trying to use computers to automatically discover investment opportunities, calculate risk indicators, and place orders. Quantifying the use of trading rules in investments, allowing the system to execute without emotion, overcomes human weaknesses, and this is called quantitative investment. Quantitative investment is widely welcomed due to its advantages such as discipline, systematicity, accuracy, and timeliness [2]. The most widely circulated quantitative investment case is Renai ssance Technologies in the United States, a fund under the company called Medallion. The industry speculates that it used hidden Markov models to guide investment, with an annualized return of 35.6% between 1988 and 2008. The Chinese stock market started relatively late and has characteristics such as a large number of retail investors and being greatly influenced by policies. According to the efficient market hypothesis, the Chinese market is far from reaching an efficient state, but because of this, the profit margin of the Chinese market is very large, which also attracts people to continuously make quantitative investment attempts.

In order to achieve success in quantitative investment, the key lies in establishing effective mathematical models or directly predicting stock prices, choosing to buy stocks when they are about to rise or sell stocks when they are about to fall, or indirectly predicting which stocks will perform relatively well in the future, in order to buy these stocks and form an investment portfolio to obtain higher returns than market benchmarks.

A major test criterion for the success of quantitative investment is that the return on investment portfolio should be higher than the benchmark return, which is generally represented by market indices such as the Shanghai and Shenzhen 300 Index, representing the weighted average return of the market [3,4].The reason for requiring the return of an investment portfolio to be higher than the benchmark return is because investors can directly purchase index funds in the fund market and obtain the benchmark return. If the return on the

*This work was supported by Research on Teaching Reform of Machine Learning and Data Mining Based on Curriculum Ideological and Political Education, Project number JC202321. Department of Computer Science Changzhi University Changzhi 046011 Shanxi China (cy12017625@czc.edu.cn)

investment portfolio is lower than the benchmark return, investors can directly give up active investment and purchase index funds. In response to these issues, the author starts from domain knowledge, places equal emphasis on financial background and statistical significance, filters a large amount of data, and establishes a feature library. When using these factors as features, the author uses hidden Markov models to improve, identify market styles, and adjust the samples used for training, so that the classifier can adapt to the new style as soon as possible.

2. Methods.

2.1. Market Style Recognition Method Based on Hidden Markov Model. The author uses Hidden Markov Model (HM) to improve the Adaboost multi factor model. The main method is to divide the implicit states of the past period of time through some indicators, select historical data that is in the same state as the current time as the training set, and then use a classifier to screen for stocks with high future returns [5].

There are several main reasons for choosing HMM:

1. There are fewer HMM parameters. When HMM learns parameters, the only parameter that needs to be manually selected is the number of implicit states.
2. The number of hidden states in HMM has some explanatory power in the application scenarios of financial time series data. Although the implicit state is assumed, after analysis, a posterior explanation can be made based on experience. For example, assuming there are three hidden states, after training, obtain parameters, and then use these parameters to predict features, it is likely to find that the three hidden states correspond exactly to the three states of rise, fall, and consolidation in the stock market. This is different from some other black box models. With explanatory power, the model becomes more convincing and safer to use. The number of states needs to be determined by oneself. For example, the market can be divided into two states: up and down, or three states: up, down, and consolidation. It can also be subdivided into five states: Big up, small up, consolidation, small down, and big down, etc. [6,7]. The setting of the number of states needs to be weighed: if the number of states is small, then the average number of samples per state is larger, containing more information, and the model has higher stability. However, the description of different window conditions may become weaker, especially in some extreme market situations, which may not be covered, but are only included in ordinary states. If the number of states is large, then the average information per state is less, there may be a situation where a certain state has only been present for a few days in the past few years, and the model's stability may decrease, but the model can differentiate between extreme market states.
3. The meaning of HMM's implicit state is somewhat consistent with the real stock market. In the minds of subjective investors or mainstream investment methodologies, the stock market itself has some implicit states. Names such as "bull market" and "bear market" refer to people's classification of stock market conditions. Although different people may have different definitions of bull and bear, most investors acknowledge the existence of a state in the stock market.
4. HMM is different from other stochastic process or time series models. Models such as Black Scholes assume that stock prices follow a lognormal distribution, but use historical data to fit mean p and volatility unable to characterize short-term characteristics, because the fitting result will bring the p-value closer to the long-term mean, it is easy to overlook the short-term trend. HMM can use different states to represent different distributions, and the mean returns and volatility of different states are different[8]. For example, in a bull market, the mean returns of all stocks are larger, in a bear market, the mean returns of all stocks are smaller, and in a volatile market, the absolute logarithm of the mean returns of all stocks is very small, this way, the explanatory power of the actual situation is stronger, that is, as long as different states are given, different distributions can be used for different states.
5. Hidden Markov models can consider a relatively moderate time span. Other time series models, such as the exponential mean EMA, typically set parameters of a few days or more, and give more weight to recent data, while the impact of long-term data gradually converges over time. That is to say, if the parameter is set to 20, the data over the past 30 days has almost no impact. If there is a long-term trend in the stock market, the EMA moving average may not be considered. But HMM is a probability graph model that assumes that the state transition matrix remains unchanged throughout the entire period and is calculated entirely based on historical data using maximum likelihood. It also uses equal weights

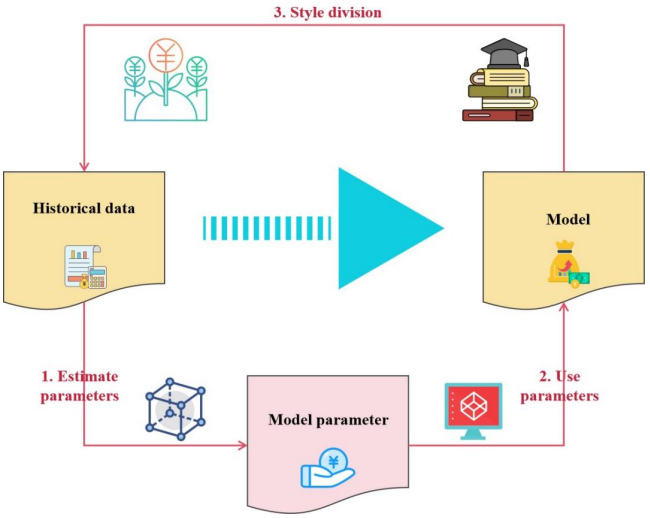

Fig. 2.1: Flow Chart of Hidden Markov Model Style Identification

for long-term data. The parameters of the state transition matrix in the stock market are generally believed to be determined by many exogenous variables, and their changes will not be sudden, and the time span of the change process will not be in the tens of days. Therefore, using a hidden Markov model here is appropriate.

HNM also has some drawbacks in its application. The first-order Markov property may not be satisfied. If there is indeed a clear state division in the stock market, then the true state of a certain day is likely to depend not only on the state of the previous day, but also on the state of the previous time. The model is still a black box. After conducting style recognition, a state sequence is obtained, and even if corresponding with relevant background knowledge, it is likely that it cannot be explained. The main process of style recognition is to first estimate the model's parameter $\lambda = (A, B, \pi)$ using historical data (features) after establishing the model, by using these parameters again, the most likely implicit state sequence to generate historical data is calculated, resulting in a style sequence. This is equivalent to clustering past time based on some indicators. The entire style recognition process is shown in Figure 2.1 [9].

After completing the style division of the market, it is possible to sample time points under the same style, instead of using a systematic sampling method every 20 days to form a training set for training. This style recognition method also needs to be based on the assumption that the market style at the current time point can find the same situation in the past period of time. If there is a fundamental and permanent change in the market, starting from the current time and completely entering a new market style that cannot be found in the past, then the model cannot be effective [10].

However, on the one hand, the possibility of fundamental and permanent changes in the market is extremely low, and China's stock market policies are gradually beginning to stabilize. On the other hand, even if fundamental changes occur, it may still be possible to identify past styles that are similar to the new style to some extent, and may still have a certain effect. Therefore, the use of this model is interpretable.

2.2. Specific steps for style recognition. Taking the warehouse adjustment date on January 3, 2022 as an example, explain the specific operation of using HBM for style recognition. Firstly, determine which features to use as observation variables. HM assumes that the distribution of observed variables is determined by implicit states, which correspond to some states of the stock market. Therefore, the features used for style recognition should correspond to the entire market, not to individual stocks. The feature library constructed above is not applicable [11].

Because the purpose of the author's use of this model is to distinguish which type of feature is more effective, the features used should also be able to demonstrate the strength of a certain type of feature in the

market, and the features should be as orthogonal as possible to each other. Drawing inspiration from the five factor model, this section also uses market factor RM, market value factor SMB, book to market ratio factor HML, profit level factor RMW, and investment level factor CMA as features. As these five factors have already described the market from different dimensions and have low correlation with each other, therefore, it can meet the requirements. The market factor RM is used to measure the systemic risk of the entire market, that is, whether the current market situation is good or not. The author will directly use the Shanghai and Shenzhen 300 index return as the daily market factor RM value. The market value factor SMB represents the size of the risk premium given by the market to a company's size [12,13]. The theory of corporate finance holds that the smaller the asset size of a company, the greater the risk, and this view is also confirmed by the behavior of investors. Therefore, stocks with smaller market capitalization should have higher returns. The market value factor SMB is used to determine how strong the small market value effect is in the entire market. The construction method of the factor is to select the stocks with the lowest market value on the previous day, calculate their average return on that day, and then subtract the average return on that day of the stocks with the highest market value on the previous day, in order to obtain the value of the market value factor SMB for that day.

The book to market ratio factor HML describes the estimated risk premium of the market on the company's additional financial risks. The book to market ratio B/M of a company is the owner's equity divided by the total market value. The higher the book to market ratio, the lower the market's valuation of the company, and the market is not optimistic. Therefore, such a company needs higher stock returns to compensate, in order for investors to be willing to invest. The book to market ratio factor HML can be used to measure how much compensation a company with a high market to market ratio should have compared to a company with a low market to book ratio. The specific calculation method for the factor value is to calculate the average daily return for companies with the lowest book to market ratio of B/M in the entire market, subtracting the average daily return for companies with the lowest book to market ratio of B/M in the entire market, to obtain the value of the HML factor [14].

The profit level factor RMW measures the market's premium on a company's profitability. The author directly uses the return on equity (ROE) to measure a company's profitability. The calculation method for the RMW factor is to calculate the average daily return for companies with the highest one-third of the total market ROE, minus the average daily return for companies with the lowest one-third of the total market ROE, to obtain the value of the RMW factor.

The investment level factor CMA represents the market's premium on investment risk. The level of investment can be measured by the reinvestment rate. Generally speaking, companies with lower investment rates have higher risks, and their stocks should have higher returns. Otherwise, they cannot compensate for the risks they bring, and vice versa. In practical operation, the reinvestment ratio can generally be calculated using the annual growth rate of total assets. The calculation method for the investment level factor CMA is also similar, using companies with the lowest one-third of the annual growth rate of all market assets to calculate the average daily return, subtracting the average daily return of companies with the highest one-third of the annual growth rate of all market assets, in order to obtain the value of the investment level factor CMA. Next, we need to make the factors reflect the situation of the next 20 days, so for each factor, we will take the geometric average of the values of the next 20 days at each time point. In the last 20 days before January 3, 2022, due to the inability to obtain data on January 3 at that time point, if the data is less than 20 days in the future, the geometric average of the factor values until the last day will be taken. After this operation, each factor contains information about how a certain style will last for 20 days.

Due to the fact that the state switching of the stock market is not very frequent, if measured in days, it is likely that there will be frequent switching of states in one or two days, and switching back and forth, which is also contrary to the starting point of this article. Therefore, the author takes 5 days as the dimension, calculates the geometric mean of the 5 factors every 5 days, divides 240 trading days into 48 5 days for training, and obtains an implicit state sequence. Then, each 5 day is the same state [15]. How many implicit states are self assumed. According to prior knowledge, there are often classification methods for styles such as large/small cap, value/growth, and the five selected features are also good indicators to reflect these styles. If the time span is too long, it may make the style transformation too complex, such as the transformation probability matrix

and other parameters may also change. Therefore, the author chose a short time span and a small sample size. When the sample size is not large, selecting too many categories can result in too few days for some categories. For the above reasons, the author directly assumes that there are four categories of implicit states [16].

Before the opening of the market on January 3, 2022, select the data of the above 5 features from the previous 240 trading days, assuming there are 4 hidden states, and use the Baum Welch algorithm to estimate the parameters of the model. The estimated state transition matrix is as follows:

$$\begin{bmatrix} 0.42 & 0.35 & 0.23 & 0 \\ 0.10 & 0.40 & 0.50 & 0 \\ 0.16 & 0.39 & 0.40 & 0.05 \\ 0 & 0.53 & 0.47 & 0 \end{bmatrix} \tag{2.1}$$

2.3. Sampling Method Based on Style Identification. December 30, 2019 belongs to implicit state 2, so what needs to be done is to count the number of time points in the entire sequence that are not the last 20 days and belong to state 0. The reason for removing the last 20 days is because the "future 20 day yield" cannot be obtained from the last 20 days and cannot be added as a sample to the training set of the classifier. The statistical result is that a total of 95 days belong to implicit state 2. Next, they will be sampled and some time points will be selected as training samples. Next, conduct sampling over these 95 days. The specific method is to first remove the time points less than 20 days before January 3, 2022, as from January 3 onwards, these time points cannot obtain the stock returns for the next 20 days. Then filter the duration of each entry into implicit state 2. If the duration is less than 10 days, skip this period and only sample in stages with a duration of at least 10 days. The sampling method is to sample on the 1st, 6th, and 11th days (if any) at each stage that lasts at least 10 days. If the total number of sampling days is less than 6 days, it is considered that the data volume is too small, and the system sampling method in the previous chapter is still followed, which is to extract a total of 12 samples from the 20th day before the current time, the 40th day before the current time, and the 240th day before the current time. If it is greater than 6 days, it indicates that the data volume is sufficient, and the filtered date data is used as the training sample for the classifier.

This sampling has two benefits:

1. Sampling should be conducted as much as possible at each stage of entering the corresponding hidden state, and periods where the hidden state does not last long enough or is not stable are excluded [17]. Guaranteed sample size. If the number of samples is too small, it will affect the performance of the classifier. After sampling using the above algorithm, a time series is obtained, at which the stock market belongs to the same state. Use the data at these time points as the training set, and use the Adaboost algorithm based on Decision Stamp classifier from the previous chapter to train the model. At the time point of January 3, 2022, using this sampling method, the training set data was obtained from 6 time points, including February 5, February 19, October 17, October 24, November 7, and November 14, 2019. From the distribution of data, it can be seen that under the new sampling method, the data from February 19th to October 17th was discarded because the style during this period was not very close to the latest style, and only data consistent with the latest time point style was used. The sample size for this sampling is 6328, with 3164 positive and negative samples each.
2. Using 16 weak classifiers, cross validation was conducted multiple times on January 3, 2022, with an average accuracy of 61.93%, which is 58 The 12% cross validation results increased by 3 81%, significantly improving the classifier performance. If we adhere to this investment method, the advantages will continue to accumulate, and ultimately we can obtain very considerable excess returns.

3. Retrospective testing and production environment testing. Adjust positions every 20 days. On each adjustment day, HBM is used for style recognition, and then data samples from the same style are used as the training set. The Adaboost method based on Decision Stump classifier is used for classification training. The 50 stocks that are most likely to hit the top 25% of all stocks in the next 20 days are selected for equal weight buying, and this process is repeated for backtesting. From March 16, 2022 to October 31, 2022, production environment testing was conducted using its own funds, and the net worth curve was concatenated after the backtesting curve. The yield curve for backtesting and production environment testing is shown in Figure 3.1, with the benchmark being the Shanghai and Shenzhen 300 Index [18].

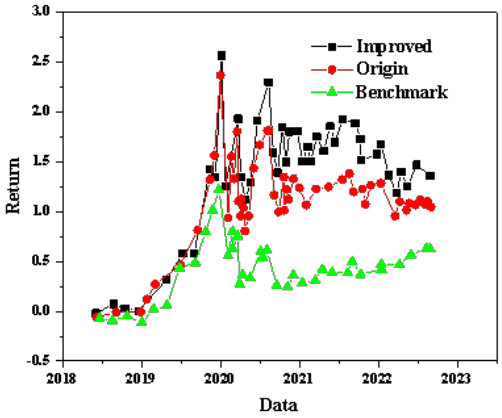

Fig. 3.1: Comparison of Net Worth Curve before and after Improvement

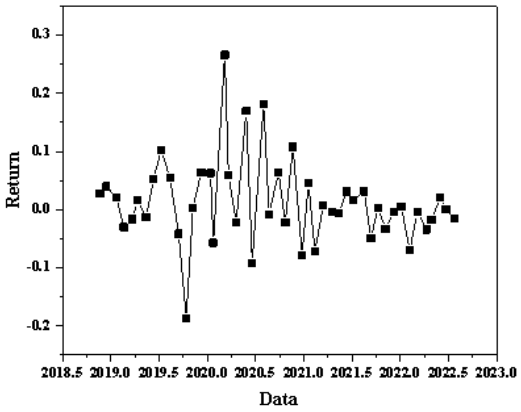

Fig. 3.2: Improved strategy with excess returns for each period

In Figure 3.1, the vertical line represents March 16, 2022, followed by the actual test curve. The improved investment strategy has defeated the benchmark in 27 out of 47 adjustment cycles (including 39 backtesting cycles and 8 real cycles), and overall remains the same as before. The performance in each specific cycle is shown in Figure 3.2.

The period with the maximum excess return has increased excess return, while the period with the lowest excess return has increased negative excess return. This phenomenon to some extent indicates that the improved strategy may amplify the effect of a certain style while increasing the overall effect. In addition, the overall style switching in 2022 is relatively frequent, and the improved investment strategy performs better than the previous strategy, indicating that the ability to follow style switching has become stronger after the improvement.

During the period from March 16 to October 31, 2022, only 2 out of 8 issues defeated the benchmark, and in the improved production environment testing, it was increased to 4 issues. And it can be clearly seen from the curve that although there are unfavorable factors in the production environment such as sliding points and market shocks that cannot be truly simulated in the backtesting environment, the improved strategy

Table 3.1: Comparison of Investment Strategy Evaluation Indicators before and after Improvement

Index	Before improvement	Improved
Total return rate	114.33%	144.42%
Benchmark return rate	71.99%	71.99%
Annualized rate of return	22.72%	27.11%
α	0.071	0.135
ρ	0.995	0.820
sharpe ratio	59.9%	79.6%
Maximum withdrawal rate	50.636%	39.99%
Information ratio	41.4%	50.4 %
Strategic volatility	31.3%	29.1%

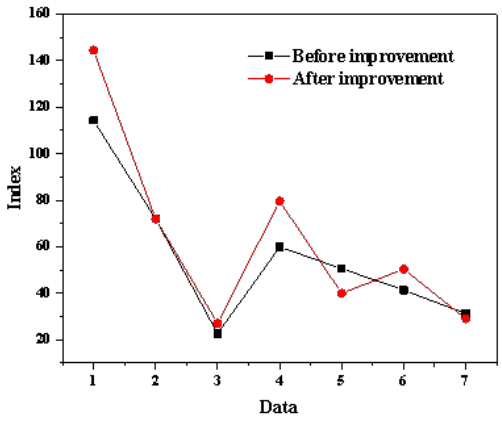

Fig. 3.3: Comparison of investment strategy evaluation indicators before and after improvement

still performs strongly. The comparison between the improved strategy and the evaluation indicators before improvement is shown in Table 3.1 and Figure 3.3 [19].

It can be seen that after improvement, a small increase in prediction accuracy will significantly improve the effectiveness of the entire strategy, with an annualized yield increase of nearly 3%, while other indicators also become better [20].

4. Conclusion. The author improved the investment strategy of the Adaboost multi factor model and improved the performance of the classifier. Firstly, a hidden Markov model is used to partition the stock market style, and then a sample is taken from dates that are consistent with the current date style. The factor values of the sampled dates are used as features, and then the Adaboost algorithm based on Decision Stump is trained and predicted. From the results of backtesting and real market testing, it can be seen that the method proposed by the author has improved the return on investment strategies. The author's work also comes with some other achievements. If a feature library is established, the features can be used in machine learning algorithms. When using other methods to research new topics, new features can also be selected from the library and directly used. In addition, the market style segmentation method based on hidden Markov also has the function of monitoring market style. The author's work has improved some of the shortcomings and innovations in the application of machine learning in quantitative investment in the past, mainly including the following points: For feature selection, the author starts from domain knowledge and uses a large amount of fundamental data to select the most predictive features. The author did not predict the financial time series, but instead classified

the relative quality of stocks at each time point, and predicted whether the future rise and fall of stocks were within the highest range. The advantage of this is that it can avoid the influence of many other factors, such as macro factors, policy factors, etc., by simply identifying relatively strong stocks.

REFERENCES

[1] Lin, N. (2021). Analysis of the impact of inflation expectations based on machine learning intelligent models. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology777(4), 40.
[2] Han, Y. (2022). Intelligent fluid identification based on the adaboost machine learning algorithm for reservoirs in daniudi gas field. Petroleum Drilling Techniques, 50(1), 112-118.
[3] Liu, J., Lin, L., & Liang, X. (2021). Intelligent system of english composition scoring model based on improved machine learning algorithm. Journal of Intelligent and Fuzzy Systems, 40(2), 2397-2407.
[4] Zhou, D., & Dong, D. (2023). An intelligent model for evaluating college students' mental health based on deep features and a multiview fuzzy clustering algorithm. Journal of Mechanics in Medicine and Biology, 23(08),5677.
[5] Qing, Y., & Zejun, W. (2021). Research on the impact of entrepreneurship policy on employment based on improved machine learning algorithms. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology67(4), 40.
[6] Sun, Z., Guo, X., Zhang, X., Han, J., & Hou, J. (2021). Research on robot target recognition based on deep learning. Journal of Physics: Conference Series, 1948(1), 012056 (6pp).
[7] Lee, H., Kim, J., Kim, E. K., & Kim, S. (2021). A novel convective storm location prediction model based on machine learning methods. Atmosphere, 12(3), 343.
[8] Hwang, J., Lee, P., Mun, S., Karathanassis, I. K., & Gavaises, M. (2021). Machine-learning enabled prediction of 3d spray under engine combustion network spray g conditions. Fuel, 293(123), 120444.
[9] Geurkink, Y., Boone, J., Verstockt, S., & Bourgois, J. G. (2021). Machine learning-based identification of the strongest predictive variables of winning and losing in belgian professional soccer. Applied Sciences, 11(5), 2378.
[10] Liang, H., Liu, G., Zou, J., Bai, J., & Jiang, Y. (2021). Research on calculation model of bottom of the well pressure based on machine learning. Future Generation Computer Systems, 124(6),324.
[11] Falahati, A., & Shafiee, E. (2022). Improve safety and security of intelligent railway transportation system based on balise using machine learning algorithm and fuzzy system. International journal of intelligent transportation systems research99(1), 20.
[12] Ma, G., & Pan, X. (2021). Research on a visual comfort model based on individual preference in china through machine learning algorithm. Sustainability, 13(14), 7602.
[13] Yuan, Z. (2021). Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm. Journal of Intelligent and Fuzzy Systems, 40(2), 2069-2081.
[14] A, Z. L., A, D. C., A, R. L., & C, Q. W. B. (2021). Intelligent edge computing based on machine learning for smart city. Future Generation Computer Systems, 115(46), 90-99.
[15] Zhao, Y. (2021). Research and design of automatic scoring algorithm for english composition based on machine learning. Scientific programming88(Pt.14), 2021.
[16] Kong, P., Peng, X., Zhang, W., & Lian, Z. (2021). Optimization of helicopter rotor airfoil wind tunnel test model based on intelligent algorithm. Journal of Physics: Conference Series, 1995(1), 012036 (6pp).
[17] Shi, L., Ding, X., Li, M., & Liu, Y. (2021). Research on the capability maturity evaluation of intelligent manufacturing based on firefly algorithm, sparrow search algorithm, and bp neural network. Complexity, 2021(3), 1-26.
[18] Chen, H. (2021). Research on innovation and entrepreneurship based on artificial intelligence system and neural network algorithm. Journal of Intelligent and Fuzzy Systems, 40(2), 2517-2528.
[19] Liu, Z., Tian, W., Cui, Z., Wei, H., & Li, C. (2021). An intelligent quantitative risk assessment method for ammonia synthesis process. Chemical Engineering Journal, 420(5), 129893-.
[20] Wang, Y. B., Wei, M. G., Liu, X. T., Chen, C., Liu, J. X., & Wu, Y. J., et al. (2022). Quantitative multi-phase-field modeling of non-isothermal solidification in hexagonal multicomponent alloys. China Foundry, 19(3), 263-274.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 14, 2023
Accepted: Nov 24, 2023