The application of artificial intelligence in stock investment

Zhao Gao
Department of Industrial Engineering and Management, School of Mechanical Engineering, Shanghai Jiao Tong University, Dongchuan Road 800, Shanghai 200240, China

Corresponding author’s e-mail: gaozhao@sjtu.edu.cn

Abstract. This paper mainly studies the application of artificial intelligence and machine learning in the field of stock investment. The principles and characteristics of KNN, k-Means, bisecting k-Means, and ANN algorithms are studied to compare the effects, similarities and differences of different algorithms. The algorithms are implemented through Python programs for stock analysis. According to the P/E ratio, dividend rate, fixed asset turnover rate, gross profit margin and other indicators of each stock, the stocks are classified and clustered to predict the stock development prospects and provide reference for selecting appropriate investment strategies.

1. Introduction
In recent years, the development of artificial intelligence technology has advanced by leaps and bounds. In particular, AlphaGo artificial intelligence developed by Google Inc. in the United States announced the challenge of human Go players in 2016 and finally won the victory, pushing artificial intelligence to the top. After that, neural networks, machine learning, etc. are well known, and artificial intelligence is widely used in various fields, including the stock and securities industry [1].

In the investment field, the stock market is complex and changeable. It is very difficult to choose a suitable investment strategy. How to make accurate predictions is the core issue. Stocks are influenced by politics, economy, and market. They are also affected by investor behaviors. The fluctuation of individual factors may cause stock prices to fluctuate drastically. Strict analysis makes it difficult to establish a straightforward linear function expression between stock prices and various factors. Therefore, although many traditional investment models can provide some suggestions for investment, they are often unsatisfactory for the prediction of financial products that lack regularity. The artificial intelligence method has many characteristics such as nonlinearity, learning, self-organization, and self-adaptation [2]. In complex scenarios, it can effectively make up for the shortcomings of traditional financial measurement models [3]. Therefore, quantitative investment based on artificial intelligence and machine learning technology has developed rapidly, providing investors with effective support [4].

In machine learning, the development of some algorithms has been quite mature, such as the KNN algorithm, k-Means algorithm and so on. Such algorithms can be applied to stock investments and can achieve very good results [5]. For example, there are two types of companies on the market. The first category is mature companies, also known as “cash cows”. They generally have low P/E ratios and high dividend rates. Their stocks are called blue chips. The second category is a company with broad development prospects but also uncertainties in the future, the so-called “growth companies”, which generally have high P/E ratios and low dividend rates [6]. If you can accurately distinguish between blue-chip stocks and high growth stocks in the market, you can provide a good reference for investors.
For another example, the machine learning algorithm can be used to identify the business strategy adopted by the company behind the stock. According to the fixed asset turnover rate and the sales gross profit margin, the company implements a low-cost strategy or a differentiation strategy, in which the former reduces its production costs and makes its products have a greater price advantage, small profits but quick turnover, often with high turnover and low gross margin [7]. The latter pays for efficiency by sacrificing efficiency. At the same time, the price of the product increases, and the corresponding stock often has a low turnover rate and a high gross profit margin. With the machine learning algorithm, the above can be accurately achieved.

Of course, the application scenarios and effects of different algorithms have some differences, and each algorithm has its local adaptability. While applying machine learning to stock forecasting, by comparing the prediction effects of different algorithms, it is also good for investors to adopt the best algorithm when formulating investment strategies to ensure the best accuracy and avoid the mistakes caused by improper selection of methods.

Next, Chapter 2 introduces two classification algorithms, KNN and ANN, and classifies the stocks according to financial indicators and known classified stocks, then compares the application effects and differences between the two. Chapter 3 introduces two similar clustering algorithms including k-Means and bisecting k-Means. The stocks are clustered according to financial indicators to observe whether the expected clustering effect can be achieved. The effects of the two algorithms are also compared. The final chapter summarizes the conclusions of the previous article and points out future research directions.

2. Application of classification algorithm in the stock market
The classification algorithm in machine learning is used to classify the sample data of unknown classification according to the data set of known classification, which belongs to supervised learning. Here we introduce the principles and characteristics of KNN and ANN and use the financial indicators of stocks to study the effect of the algorithms by classifying stocks into blue chips and high-growth stocks and dividing corporate strategies into low-cost strategy and differentiated strategy.

2.1. KNN algorithm
The KNN algorithm, the k nearest neighbor algorithm, belongs to supervised learning. The basic idea of this method is to treat each sample as a point in the P-dimensional space. When given a test data, calculate the distance between each training sample and the test data. Select the N training sample points closest to the test data and get the category of each training sample. Judging the category of the test data according to the principle of minority-majority.

The KNN algorithm can effectively solve the problem of sample imbalance when classifying, that is, it is suitable for classifying rare events. Besides, it has high accuracy because it is not sensitive to outliers. Of course, KNN also has some drawbacks. For example, since it is necessary to calculate the distances, the process is computationally intensive and the space overhead is large. In general, the KNN algorithm is highly adaptable, especially for the automatic classification problem with a very large sample size.

The implementation of the traditional KNN algorithm is understandable. The program flow is as follows.

![Figure 1. The algorithm processing flow chart of KNN.](image)
Next, we use KNN for stock classification. First, identify the type of stock. P/E ratio, dividend rate and classification label for each stock in the training sample are known. According to the price-earnings ratio and dividend rate, all stocks are divided into two categories. Category 1 corresponds to blue-chip stocks and 2 corresponds to high-growth stocks. For a new stock, we can use the training results of KNN to classify it reasonably. For example, the test data is (P/E ratio 6.12, dividend rate 0.0399), and the classification label is 1 with KNN’s calculating.

![Figure 2. The identification result of the type of the stock with KNN which shows that the label of test data is 1.](image)

Next, carry out corporate strategy identification. Mainly according to the fixed asset turnover rate and gross profit margins, divide the corporate strategies behind the stocks into three categories. Category 1 is the low-cost strategy, category 2 is the differentiation strategy, and category 3 represents other types, including some companies with low turnover and gross margins, which may suggest strategy failures. For example, the data point to be classified is (fixed asset turnover rate 21.5043, sales gross profit margin 69.8744), and KNN is used to determine that the classification label is 2, that is, the differentiation strategy is implemented.

![Figure 3. The result of the corporate strategy identification with KNN which shows that the label of test data is 2.](image)

2.2. ANN algorithm
Artificial Neural Networks (ANN) simulates the information processing process of the human brain, with the characteristics of large-scale parallel processing, distributed information storage, and good self-organizing self-learning ability [8]. BP algorithm, also known as the error backpropagation
algorithm. The BP neural network algorithm can theoretically approximate arbitrary functions and has strong nonlinear mapping ability. Moreover, it has great flexibility since the number of middle layers of the network, the number of processing units of each layer, and the learning rate of the network can be set according to specific conditions. It has a wide range of applications in many fields such as optimization, signal processing, pattern recognition, intelligent control, and fault diagnosis.

In this paper, we build a simple fully connected neural network framework. The basic idea is to use the matrix to simulate the operation process between weight coefficients and the output result of the previous layer, and the processing of the activation function.

The framework consists of three parts: initialization function, training function, and running function. The initialization function requires us to set the number of nodes in the input layer, output layer, hidden layer and set the learning rate. In the training function, we accept the data of the input layer and implement the algorithm (such as BP algorithm) to optimize the weight. In the run function, we run the trained neural network and output the classification results.

The network trains with a given training set. For the test data, the probabilities that the data belongs to each classification are finally output, and the most probable probability is selected as its prediction classification.

Similarly, using the data in 2.1, ANN is used for stock type identification and enterprise strategy identification respectively, and finally, the results are similar to figure 2 and figure 3.

3. Application of clustering algorithm in the stock market

Different from the classification algorithm, the clustering algorithm is used to cluster the sample into several categories according to the intrinsic characteristics of the dataset, which belongs to unsupervised learning. Here we introduce k-Means and bisecting k-Means and cluster the stocks according to financial indicators.

3.1. K-Means algorithm

The k-Means algorithm is a common clustering algorithm. The main idea of k-Means is to randomly select k sample points from the sample set as the initial centroids according to the specified number of divisions k. Then, each data point is traversed, and each point is assigned to the cluster at the closest centroid as the initial division. After that, the previous center point is replaced by the mean center point of each cluster. Iterate according to the above steps until convergence or a termination condition, and the final division is obtained [9].

The k-Means algorithm can maintain good scalability when performing clustering, especially when dealing with large datasets. At the same time, the efficiency is high with good clustering effects. Since the number of divisions k is specified by the user, different k values result in different results. The algorithm is very sensitive to the initial cluster center point, and different initial partitions may lead to the final different partitioning rules. A large deviation may occur if there is an outlier in the sample since we choose the cluster mean as the new center point. Of course, some flaws can also be optimized by some methods. For example, we can use the median instead of the mean to perform the clustering, that’s the k-Mediods algorithm.

In the k-Means program, several functions are encapsulated, including data reading function, distance calculation function, creation of initial centroid list function, and k-Means main function. The flow diagram of creating the initial centroid list function and the k-Means main function are as follows, wherein the main function mainly utilizes the idea of iteration.
Figure 4. The algorithm processing flow chart of k-Means including (a) The function of creating the initial centroid list, (b) The main part of the k-Means algorithm function.

The dataset of Chapter 2 is re-clustered with k-Means, and the final classification result and the centroid position of each cluster can be obtained.

Figure 5. Clustering results and the positions of the centroids with k-Means (a) With P/E ratio and dividend rate, (b) With turnover rate and gross profit margin.

3.2. Bisecting k-Means algorithm
The main idea of bisecting k-Means is to first treat all points as one cluster and then split the cluster into two. Then split the cluster that minimizes the cluster cost function (that is, the sum of squared errors) into two clusters. This is done until the number of clusters is equal to the number k given by the user [10]. One of the implied principles is that because the SSE of the clustering can measure the clustering performance, the smaller the value, the closer the data points are to their centroids, and the better the clustering effect. Therefore, we need to divide the cluster with the largest SSE again, because it is likely that multiple clusters are treated as one.
Bisecting k-Means is derived from k-Means. Because the clustering result of the traditional k-Means algorithm is susceptible to the selection of the initial cluster centroids, so the algorithm is improved by selecting the initial center point more strictly. Make the distance between the centroids farther and avoid the initial cluster center belongs to one cluster, which overcomes the problem that the algorithm falls into local optimum to some extent.

Bisecting k-Means program includes the traditional k-Means but can avoid the classification error caused by the randomness of the initial cluster center.

Figure 6. The algorithm processing flow chart of bisecting k-Means.

It can be seen that the bisecting k-Means algorithm can also obtain similar clustering results.

4. Conclusions

It can be seen that both the classification algorithm and the clustering algorithm can achieve good results when applied to the actual stock market. However, in comparison, there are still some differences between different algorithms.

In terms of classification algorithms, for a simple dataset, KNN is sufficient to obtain accurate results, and it may be difficult to obtain stable classification results with ANN. However, for large datasets with fuzzy classification boundaries, the neural network can perform the classification more accurately, provided that the appropriate number of hidden layer nodes and learning rates need to be selected.

For the clustering algorithm, if the general range of k is known for a simple dataset, k-Means can accurately complete the work of clustering but the clustering result of bisecting k-Means may be
unstable. However, the advantage of bisecting k-Means is that it can avoid falling into the local optimal solution with an unreasonable k-value, which is the deficiency of traditional k-Means. Therefore, if the number of specific clusters is unknown to a large amount of data, it is more appropriate to use the binary k-Means algorithm.

The real stock market has complex data and so many indicators, and the stock characteristics and market trends are not often completely determined by the data we see. Therefore, when using quantitative methods such as machine learning, it is necessary to choose appropriate methods based on actual conditions and specific problems.

References
[1] Dunis, C.L., Middleton, P.W., Karathanasopolous, A., & Theofilatos, K. (2016) Artificial intelligence in financial markets. Palgrave Macmillan, London.
[2] Moghaddam, A.H., Moghaddam, M.H., & Esfandyari, M. (2016) Stock market index prediction using artificial neural network. Journal of Economics Finance & Administrative Science, 21(41): 89-93.
[3] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(1): 259-268.
[4] Sun, Z., Zhao, S. (2019) A review of the application of machine learning in stock investment. In: the 4th International Conference on Humanities Science, Management and Education Technology (HSMET 2019). Singapore. 151-159.
[5] Zemke, S., (1999) Nonlinear index prediction. Physica A, 269(1): 177-183.
[6] Barker, R.G., (1999) Survey and market-based evidence of industry-dependence in analysts’ preferences between the dividend yield and price-earnings ratio valuation models. Journal of Business Finance & Accounting, 26(3-4): 393-418.
[7] Ding, B., Li, L., Zhu, Y., Liu, H., Bao, J., Yang, Z. (2019) Research on Comprehensive Analysis Method of Stock KDJ Index based on K-means Clustering. In: the 3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019). Dalian. 484-491.
[8] Takeuchi, L., & Lee, Y., (2013) Applying Deep Learning to Enhance Momentum Trading Strategies in Stocks. The Journal of Finance, 2013, 41(16): 5-16.
[9] Chen, Y., & Hao, Y. (2017) A feature weighted support vector machine and k-nearest neighbor algorithm for stock market indices prediction. Expert Systems with Applications, 80: 340-355.
[10] Banerjee, S., Choudhary, A., & Pal, S. (2015) Empirical Evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means Clustering Algorithms. In: IEEE International WIE Conference on Electrical and Computer Engineering. Dhaka. 168-172.