Research on the Application of Machine Learning in Quantitative Investment

He Jiawen[1], Wei Ziyi[2], Zhu Xuanbing[3]

[1]College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002

[2]School of Finance, Dongbei University Financi & Economics, Dalian, Liaoning, 116025

[3]Dalian Maritime University, Dalian, Liaoning, 116000

Keywords: machine learning; quantitative investment; finance; big data

Abstract: With the development of the economy in recent years, the field of financial investment is growing stronger. Quantitative investment is a type of investment strategy that uses mathematical methods to analyze and model financial markets. Machine learning requires computer programs to pass data sets learning on the Internet to improve its performance when dealing with specific tasks. Both of them need to extract information from the data, so combining them into research has become a current hotspot. Based on this, this article analyzes the application of machine learning in quantitative investment. First of all, an overview of machine learning is introduced, and its basic concepts and classification are introduced. Then the technology and advantages of quantitative investment and the current status of application of machine learning in quantitative investment are explained. Finally, two aspects of the specific application including big data processing and financial investment models are analyzed. Only continuous innovation research can maximize the value of machine learning in financial investment.

1. Introduction

The most important application of machine learning methods to quantitative investment is that it can make China's financial market more rational. As far as China's stock market is concerned, most retail investors are still in the stage of intuitive investment or original rule investment, and the market is vulnerable to emotions. The stock price is distorted, and the distortion of the stock market is not conducive to the capital market to play its functions of financial communication and information dissemination [1]. The strategic design and computing resource thresholds of machine learning are far beyond the scope of ordinary investors. With clear advantages in risk, market trading volume will be concentrated in institutions using machine learning methods. Based on the expertise of institutional investors, we have reasons to believe that the market will become more rational in the long run.

2. Machine Learning Overview

Machine learning is another important research area of artificial intelligence applications after expert systems, and it is also one of the core research topics of artificial intelligence and neural computing. Machine learning is a relatively young branch in the field of artificial intelligence, and its development process can be divided into 4 period: 1) The mid 1950s to mid 1960s, which is a warm period; 2) The mid 1960s to mid 1970s, known as the cool period of machine learning; 3) the mid 1970s to mid 80s, called revival period. 4) Beginning in 1986 is the latest stage of machine learning. Machine learning in this period has the following characteristics: machine learning has become a new marginal subject and has become an independent course in colleges and universities; it integrates various learning methods and has various forms. Research on integrated learning systems is on the rise; a unified view of various basic issues of machine learning and artificial intelligence is being formed; the scope of application of various learning methods is expanding, and some of the results of applied research have been transformed into commodities.

2.1 Basic Concepts of Machine Learning

The main purpose of machine learning research is to use computers to simulate human learning activities. It is a method for studying computers to identify existing knowledge, acquire new knowledge, continuously improve performance and achieve self-improvement [2]. There are three research goals for machine learning: 1) Cognitive model of human learning process; 2) Universal learning algorithm; 3) Method of constructing task-oriented dedicated learning system. In the basic model of the learning system, there are 4 basic constituent links. The environment and the knowledge base are somehow a collection of information expressed in the form of knowledge representation, which respectively represents external information sources and the knowledge of the system; the environment provides certain information to the learning link of the system, and the learning link uses this information to improve the system's knowledge base to improve the effectiveness of the execution link to complete the task. The "execution link" completes certain tasks based on the knowledge in the knowledge base, and at the same time feedbacks the obtained information to the learning link.

2.2 Classification of Machine Learning

2.2.1 Symbol-based machine learning

Symbol-based machine learning is based on the collection of symbols that represent entities and relationships in the problem domain. The symbol learning algorithm uses these symbols to introduce new and effective general rules, and the rules are also expressed by these symbols.

1) Variant space search. Candidate elimination algorithms rely on the concept of variable space, which is a collection of all concept descriptions consistent with training examples. These algorithms have more examples to reduce the size of the variant space.

2) ID3 decision tree induction algorithm. ID3 is the same as the candidate solution exclusion algorithm, which inducts concepts from examples. The algorithm has the following advantages: the representation of learned knowledge; the method of controlling computational complexity; the inspiration of selecting candidate concepts information; has the potential to process noisy data.

3) Inductive bias and learning ability. Inductive bias refers to the learning program used to limit the concept space or select concepts in this space.

4) Knowledge and learning. Traditional knowledge learning methods mainly include mechanical learning, guided learning, inductive learning, analogical learning and interpretation-based learning.

5) Unsupervised learning. The clustering problem is to compare the similarity between a group of unclassified objects and measure objects. The goal is to classify objects into categories that meet certain quality standards.

6) Reinforcement learning. Reinforcement learning is the design of algorithms that transform the external environment into a way to maximize the amount of rewards.

2.2.2 Connected machine learning

The connectionist approach is to express knowledge as a network of small individual processing units to activate or inhibit state patterns. Inspired by the structure of animal brains, connectionist network learning is achieved by modifying network structure and connection weights through training data [3]. In the connected system, the processing is parallel and distributed, there is no symbol processing in the symbol system. The patterns in the domain are encoded as digital vectors; the connections between neurons are also replaced by digital values; the conversion of the patterns is also the result of digital operations-usually using matrix multiplication. The designer's choice of connecting system structure constitutes the inductive bias of the system. Algorithms and system structures that apply these techniques generally use training methods instead of direct programming. This is also the most advantage of this method. The machine learning methods of connectionism mainly include the following: the basis of connected networks, perceptual learning, back propagation learning, competitive learning, Hebbian consensus learning, attractor network or memory.

2.2.3 Emerging learning model

Emergence models are inspired by genetics and evolution. Genetic algorithms begin with a set of candidate solutions to a problem, and candidate solutions evolve according to their ability to solve the problem: only the fittest survive, and they exchange each other to produce the next generation of solutions. In this way, the solution is continuously enhanced, just like the evolution of the real world described by Darwin. Emerging learning models mimic the most beautiful and powerful forms of life evolution of plants and animals in nature. It is mainly used in genetic algorithms, classifier systems, and genetics program design, artificial life, and society-based learning.

3. Quantitative Investment And Research Status

3.1 Quantification Technology

In general, financial quantification technology can be divided into two categories, one is P Quant and the other is Q Quant. Although they are the same asset pricing mechanism, their principles and audiences are quite different, and their respective trends have disappeared.

Q Quant refers to a risk-neutral measure. Under the "risk-neutral" theoretical assumption, historical data is only a record of past numbers. They have nothing to do with the future, so they cannot directly help predict the future trend of financial products. The pricing mechanism is still mainly used based on mathematical models, such as stochastic processes and partial differential equations, the pricing models derived from them are mostly theoretical and obscure. P Quant refers to the true probability measure, which is different from "risk neutral" and "real probability" under the theoretical assumptions, the probability distribution required to build a pricing model should be estimated based on historical data, not based on mathematical models. In other words, the future trends predicted by this pricing model are mainly based on statistical data. Therefore, it is "real", and the larger the amount of data, the more likely its prediction effect will be close to the actual effect in the future, which is the so-called "big data" [4]. In order to deal with the huge historical data, product assistance is often indispensable, so the product technologies related to P Quant are mainly time series, Bayesian algorithms, machine learning, etc., which are closely related to computer molding process technology.

From this, it can be seen that according to the difference of historical data, the difference between Q Quant and P Quant is actually obvious, the former is based on the assumption of the future to calculate the present, and the latter is based on the reference to history to infer the present. Although both need to be applied to historical data, but the former usually builds a model first, and then continuously refines the parameter performance of the model through historical data, so the role of historical data is mainly to optimize the model's sharpening stone; the latter usually builds several backups first. Select the model, and then apply the historical data to different candidate models, and select the model that performs best based on the resulting calculation results, so the role of historical data in P Quant is upgraded to the touchstone of a selection model [5]. It is difficult to judge which theory is more scientific, because history is worthy of reference, but history will not repeat itself. Historical data may not only help us make scientific predictions, but may also lead us astray.

However, this difference is of great significance to financial market participants with different roles. Specifically, because Q Quant mainly relies on mathematical models and does not rely on historical data, this means that even in the case of relatively scarce data. We can still develop some new products out of thin air based on this theory, which is undoubtedly good for sellers in the financial market. Sellers represented by investment banks and securities firms are mostly engaged in derivative pricing, that is, through development and selling new financial derivatives to achieve profitability. The competitors compete with each other based on the technical advantages and disadvantages of the original product and the actual needs of the subscription market. Therefore, they rely more on the manufacturing characteristics of "from scratch" capabilities of Q Quant [6].

3.2 Advantages of Quantitative Investment

Compared with traditional investment methods, quantitative investment mainly has the following advantages: (1) Data-based decision-making: That is, quantitative investment is based on the principle of probability statistics. Through the mining of a large amount of historical data, it finds the inherent laws of data generation to guide decision-making. (2) Overcoming the limited information processing capabilities of the human brain: In traditional investment methods, fund managers rely mainly on fast and real-time analysis and decision-making of data generated by financial markets, but the human brain's ability and speed of information processing are limited. When the data of tens of thousands of stocks in the market is changing rapidly, the energy of the human brain becomes insufficient. The quantitative investment system can track changes in market conditions in time, analyze a large amount of market data in real time, and continually find profit opportunities. (3) Discipline: After the model sets the corresponding investment strategy, the computer will strictly execute the trading signals it generates, without excessive manual intervention, and it can avoid the impact of human subjective emotions as much as possible. □

3.3 Application Status of Machine Learning in Quantitative Investment

The earliest machine learning methods used abroad include support vector machines (SVMs), decision trees, etc. The initial exploration at this stage shows the attractive prospects of machine learning in the field of quantitative investment. At the same time, the long-established deep learning community has spread exciting news: At the 2012 image recognition competition, Alex Net won the championship by far surpassing the second place. It is no exaggeration to say that the success of Alex Net has once again inspired the academic and industry enthusiasm for deep learning. Deep learning tools have gradually become an important technical means in quantitative investment research.

Today, the application of machine learning in quantitative investment is not limited to the early simple strategy construction combining financial rules, but has developed into a system-level project using a series of new technologies such as integrated learning frameworks and deep learning technologies [7]. Quantitative investment research using machine learning methods is a financial-computer interdisciplinary research topic, and its research value and significance have gradually received recognition and attention from the computer industry: the 2017 International Machine Learning Conference (ICML) included the use of Fourier Ye analyzes the research of improved LSTM neural network, which proposes a new network structure based entirely on financial market applications.

With the gradual deepening of research on related topics in the financial and computer fields, a lot of research results have been produced in the academic world in recent years, and research on core network design, system architecture, and integration with traditional research methods has made significant progress. China's financial market has not been established for a long time, and there are constraints such as lack of hedging tools, trading system restrictions, and (stock) market sentimentality. Objectively, there are indeed some unfavorable conditions for the development of the quantitative investment industry. The objective facts are: China's quantitative investment industry started late, but this does not prevent academic and industry research on related topics.

Domestic scholars have also experienced early exploration of BP neural networks and SVR simple models. Since then, research results based on ensemble learning and deep learning have gradually appeared, such as random forests [8].

The industry's focus on machine learning started roughly after 2015, at the beginning of the second high-speed development period of China's quantitative investment industry. Guoxin Securities' 2016 research report based on Adaboost is an earlier literature. The machine learning method it uses is a framework-level boosting method. After 2017, powerful securities companies have published a large number of research reports on machine learning and deep learning applications, and some of them have more in-depth integration of securities companies in the financial industry. Deep accumulation in research, experimental exploration of classic integrated learning methods (boosting, stacking) and deep learning methods (LSTM, GRU).

Due to the special nature of the financial industry, the public research reports are mostly seller research reports, and the buyer materials that actually conduct market transactions are scarce. The effective strategies are also strictly confidential. We have every reason to believe that the research reports and academic papers published today are just the tip of the iceberg of machine learning applications in financial data analysis applications.

4. Application of Machine Learning in Quantitative Investment

Quantitative investment technology, which relies on mathematical models and computer programs as the theoretical and applied foundation is booming. Quantitative investment can be traced back to the first quantitative investment fund created by Thorpe in 1969, and it has a history of fifty years; Due to the rigorous trading logic, strict program settings and relatively stable return on investment of this trading method, it has been booming so far, and its market share has continued to expand. The medallion fund helmed by Simmons is one of them. The average annual rate of return has exceeded 30%, and can continue to earn excess returns even during the financial crisis. As of 2017, the size of funds using quantitative investment concepts in the global market has exceeded US $ 3 trillion, accounting for 30% of the total number of funds and it has become the most important participant in the global financial market. China's quantitative investment appeared relatively late compared to western countries. With the introduction of stock index futures, with the emergence of feasible hedging instruments, quantitative investors have gradually appeared in our market. China's capital market and investment concepts are still immature. Research by most scholars believes that market effectiveness is low, and it is the characteristics of precisely immature and large fluctuations that have made quantitative investment a huge room for growth. This year, the quantitative interface of brokers will be reopened to private equity funds, meaning that a large amount of funds will enter the market. Domestic institutional investors are actively making arrangements. According to the statistics of Fortune.net, as of the end of 2017, the number of funds based on the concept of quantitative investment was clearly stated in the filing of the Securities and Futures Commission. With the advent of the big data era, and machine learning and artificial intelligence technologies have attracted more attention from fintech companies. The application of machine learning in the field of quantitative investment will also be one of the future research directions [9]. Machine learning has achieved impressive achievements in many fields; financial data relationships are more complex, and there are often shortcomings of traditional structured data analysis method. The linear and parameterized equations often cannot fit the data relationship well, and the homogeneity in the market is more serious, which leads to the gradual disappearance of alpha. Many machine learning algorithms are non-parametric, non-linear, theoretically can better capture the data relationship, and at the same time, machine learning algorithms have relatively few applications in the stock market. What Rose said "don't be a crowded market", relatively small algorithms are currently more likely to bring excess returns. Many domestic securities firms, funds, etc. have also started relevant research.

4.1 Big Data Processing

At present, the big data is usually above the PB level. This is a large amount of data that people cannot directly perform statistics and analysis, which must be processed with the help of cloud computing. Modeling on big data can not be separated from the machine learning and deep learning. Big data research common in the investment field includes:

1) Big data news public opinion analysis

Foreign Raven Pack companies and some domestic companies provide financial field news and other media public opinion modeling and analysis services. There are also many related studies in the academic world. For example, Ding Xiao et al. (Published in 2015) described a paper based on event-driven strategies for deep learning.

2) Image identification

Orbital Insight uses machine learning models to identify satellite images and predict retail profit, oil inventory, mine output, and more.

3) Financial knowledge graph

The construction of the financial knowledge map is also an important application direction for the investment field.

4.2 Financial Investment Model

Because machine learning is only used as a tool for big data processing, strictly speaking, the above applications do not belong to the application of machine learning in the economic and financial fields. Generally, deep learning is good at end-to-end optimization. In the field of quantitative investment in finance, investment signals are obtained from processing raw data, online deep learning models, and then strategy optimization and risk management through technologies such as reinforcement learning. There should be relevant research and practice in the industry. However, I do not quite agree with this approach. The following two points:

(1) The data boundaries of the issues in the investment field cannot be effectively defined.

(2) The future sample space is different from the historical sample space.

Using deep learning as a predictive model directly for asset price prediction is also one of the ways that can usually be thought of Batres-Estrada (2015) uses deep learning models to predict the positive and negative returns of stocks, and select stocks to build portfolios based on the prediction results. The distribution characteristics of thick tail spikes of financial assets are not suitable for the prediction target. I need to study in depth how to effectively define investment problems, model selection, and data boundaries, such as the following exploratory studies:

1) Application on high-frequency data sets. Dixon (2017) used a recursive neural network (RNN) model to predict price reversal and build a strategy on S & P S & P500 stock index futures minute-level transaction quotation order data, and finally achieved good results. High-frequency data can reach more than TB level (1TB = 1000GB), and the sample distribution is also more stable and should be more suitable than the daily level data.

2) Zhu et al. (2016) used a deep learning model to predict the boundaries of the stock cabinet. The theory of the stock cabinet believes that during the operation of the stock, a certain price region will be formed, that is, the stock price will fluctuate within a certain range. If the price goes out of range, it will enter a new box. The trading strategy is to buy the stock when the stock price breaks the top border, or sell the stock when the stock price breaks the bottom border.

Although big data provides microeconomic data, there are multiple levels from the record of human microbehavior to macroeconomic and financial phenomena. For example, in the behavior of social groups, when a certain critical point is reached, the scale and group distribution will be a macro-explosive growth. The financial investment model needed in the era of big data is not just a simple statistic, but needs to learn more effectively those structures and relationships that were neglected in previous statistics on large data sets through machine learning and deep learning models. Therefore, it is possible to obtain better model prediction results [10]. At present, there have not been any related research results on the effective fusion of influential financial investment domain models and machine learning models.

In those areas where machine learning has mature applications, such as financial models such as anti-fraud and micro-credit, various behavioral features are used to discriminate financial behaviors of users, such as the time between two mouse clicks, and time-series features such as browsing time were often ignored by statisticians. Fortunately, in mature Internet companies, these data can be effectively obtained through buried point technology. Behavioral economics can provide guidance for feature selection. More importantly, machine learning models are used in improving the effectiveness of existing business processes, but when faced with process and product optimization, related discipline collaboration is even more important. In the field of recommendation, professor Zhang Yongfeng and others maximize the interpretability of the recommendation model and maximize the benefits in economics of the combination of such theories and achieved good results. In the future, effective integration of complex economics, behavioral economics, and machine learning-related theories is worthy of in-depth research. For example, complex networks and social networks have been the social sciences and behavioral economics in recent years. One of the hot topics in learning, the application of probabilistic graph models and deep learning on social network data are also non-one of the hot topics, and then combine the results of multiple areas to be explored further. In the area of investment, combined with game theory, strengthen the depth of learning need to continue to build investment strategies are all explored in depth.

In order to meet the challenges of the era of big data, will machine learning and deep learning accompanied by big data have a revolutionary impact on the economic and financial fields? Well-known hedge funds such as Renaissance Technology Corporation and Two Sigma have been recruiting machine learning-related experts in large numbers. In the 1970s, machine learning began to be applied to the economic and financial fields. At that time, it was more commonly referred to as statistical learning and nonlinear optimization. The Renaissance company was founded by the famous mathematician James Simons. Although the Renaissance company has always kept a low profile, its internal use of machine learning technology has long been well known in the industry. The company has maintained it for decades. There is a Nova fund within the well-known Medal Fund that has sustained high returns. This fund was responsible for the core staff of the team that originally developed the artificial intelligence language translation system at IBM around the 1990s.

5. Conclusion

In short, machine learning has not seen a disruptive effect in the field of financial investment, because the use of machine learning in the field of financial investment has already begun, and now it is only more famous. In the future, with the explosion of financial-related data, under the improvement of new algorithms and the continuous optimization of new deep learning algorithms, the application of machine learning and deep learning in the field of financial investment will become increasingly important.

References

[1] Xiong Haifang. Application of quantitative investment analysis in securities investment teaching [J]. Finance Teaching and Research. 2014 (02): 125-128.

[2] Yang Junqi, Guo Hongze, Yang Pengcheng, Li Qi. On quantitative stock selection and qualitative stock selection [J]. Times Finance. 2017 (06): 100-102.

[3] Wang Kai, Long Weijiang. Research on high transfer stocks based on ensemble learning[J]. Times Finance. 2016 (36): 78-80.

[4] Wang Shuyan, Cao Zhengfeng, Chen Mingxuan. Research on the application of random forest in quantitative stock selection[J]. Operations Research and Management. 2016 (03): 49-51.

[5] He Qing, Li Ning, Luo Wenjuan, Shi Zhongzhi. Overview of machine learning algorithms under big data [J]. Pattern Recognition and Artificial Intelligence. 2014 (04): 56-58.

[6] Wang Li. Research on forecasting method of stock price fluctuation based on artificial intelligence algorithm [D]. Jilin University. 2016.

[7] Zhang Xiao, Wei Zengxin. Application of random forest in stock trend prediction [J]. China Management Informationization. 2018 (03): 204-206.

[8] Sun Jiao. Multi-factor quantitative investment strategy and empirical test [D]. Nanjing University. 2016.

[9] Zhu Chenxi. Empirical analysis of China's a-share market multi-factor quantitative stock selection model[D]. Capital University of Economics and Business. 2017.

[10] Li Zhibing, Yang Guangyi, Feng Yongchang. Empirical test of the Fama-French five-factor model in China's stock market [J]. Financial Research, 2017 (6): 191-206.