Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/44/20230134

# Application analysis of financial data mining in investment decision

**Rongying Zeng**

Fuzhou University, Fuzhou, Fujian Province, 350108, China

rongying.zeng.2021@mumail.ie

**Abstract.** Amidst the escalating complexities that define the contemporary financial market and the rapid proliferation of information, traditional methods of formulating investment decisions confront increasingly formidable challenges. In response to these intricate dynamics, the realm of financial data mining has emerged as a prominent avenue of scholarly investigation within the investment domain. This paper's fundamental objective is to conduct a comprehensive retrospective analysis of the diverse applications of financial data mining in the context of investment decision-making.This scholarly pursuit entails a meticulous synthesis of existing academic inquiries, concurrently proposing potential avenues for future advancements in this field. By undertaking this academic endeavor, the paper strives to make substantive contributions to the refinement of methodologies essential for adeptly navigating the multifaceted landscape of modern investments. As the financial landscape continues to evolve, this study aspires to offer insights that not only enhance the efficacy of investment strategies but also foster a deeper understanding of the intricate interplay between data mining techniques and decision-making processes. Through the synthesis of empirical findings and theoretical perspectives, this paper seeks to underscore the pertinence of leveraging data-driven approaches in investment practices, thereby promoting a more informed and sophisticated investment landscape.

**Keywords:** Data Mining; Machine Learning; Deep Learning; Time Series Model.

## 1. Introduction

In the present era characterized by dynamic transformations in financial markets and a proliferation of information, investment decision-making confronts intricate and challenging circumstances. Conventional investment methods, predominantly reliant on subjective factors like fundamental analysis, technical analysis, and expert judgment, struggle to effectively address the heightened market uncertainty and rapid fluctuations. In response to these demands, financial data mining has emerged as a nascent discipline that leverages computer science and statistical techniques to extract latent patterns and regulations from financial data. Consequently, it has garnered significant attention and research interest as a means to enhance the precision of capturing market trends, identifying investment opportunities, and mitigating risks. This paper aims to comprehensively review the diverse applications of financial data mining in the realm of investment decision-making, while providing a comprehensive synthesis and evaluation of pertinent research endeavors.

## 2. Overview of data mining technology

Data mining is the process of extracting hidden but useful information and knowledge from a large number of fuzzy, random and incomplete actual data [1]. In essence, it is a method of data analysis, mining potential useful information contained in a variety of data, such information is hidden, and it needs to establish models or rules for in-depth analysis and collation to improve the connection between information, and condense these data into knowledge. Data mining has two main tasks, one for description and the other for prediction. Description is to summarize the characteristics of the data in the database to be processed. Prediction is a kind of inference and evaluation, which is to process and predict the data and information currently mined. Data mining functions can be divided according to task content or task nature, mainly including concept description and various analysis forms, including association rule analysis, regression analysis, cluster analysis, sequence pattern analysis and classification analysis, etc. [2].

## 3. Research on the application of financial data mining in investment decision

Financial data mining is used in various fields of investment decision-making, including market trend analysis, high-frequency trading, algorithmic trading, risk assessment and many other aspects. Different aspects of these studies reveal the potential benefits and challenges of using data mining techniques in the investment arena. By studying various applications of financial data mining, it aims to enhance the decision-making process, improve investment performance, and better manage risk in a dynamic market environment.

### *3.1. Market forecasting and stock selection*

The inception of financial data mining can be traced back to the early years of this century. Pioneering research endeavors, exemplified by the efforts of the research team at the Hong Kong University of Science and Technology, yielded successful predictions regarding stock market trends. Their accomplishment was underpinned by the systematic application of data mining technology, enabling precise projections of the Hang Seng Index's trajectory [3]. Nevertheless, the ongoing evolution of the machine learning domain has ushered in a slew of innovative methodologies, tailored to address the intricate intricacies intrinsic to financial markets.

Classical time series analysis techniques and conventional machine learning methodologies, while enduringly pertinent, have ceded ground to cutting-edge deep learning approaches, culminating in significant advances within the realm of finance. Notably, Jin Yujia's seminal work in 2017 introduced a fuzzy time series model for stock price prognostication [4]. Nevertheless, contemporary paradigms such as the Transformer and BERT architectures have gradually ascended to prominence in financial time series forecasting, uniquely positioned to adeptly capture intricate non-linear relationships and perturbations characterizing market fluctuations.

Moreover, cognizant of the multidimensional nature and the intrinsic complexity of financial data, novel methodologies such as factor analysis and the OPTICS-Plus algorithm have been posited, fostering heightened analytical efficacy and the swifter convergence of computations. While these techniques bear significance in the domain of stock price prediction, their applications transcend into the prognostication of financial exigencies. Concurrently, the ascendancy of deep learning methodologies like Long Short-Term Memory networks (LSTM) [5] has endowed the financial community with potent instruments for discerning underlying time series patterns. However, it is prudent to entertain more advanced architectures like Gated Recurrent Units (GRU) and the Transformer model, as these structures are better poised to assimilate the intricate non-linear dynamics and voluminous datasets emblematic of the stock market.

In the field of market prediction and stock selection in this study, with the goal of demonstrating the effectiveness of emerging machine learning methods in practical applications, we selected the Alpha Vantage dataset as the research object to illustrate its characteristics, application cases, and the use of new machine learning methods in detail.

### *3.1.1.* Data selection - Alpha Vantage dataset

Data Source: Alpha Vantage is an online platform that provides diverse financial data, encompassing high-frequency data such as stock prices and indices.

Characteristics: The Alpha Vantage dataset embodies a salient feature set, inclusive of high-frequency stock prices and trading data, with granularity down to per-minute or per-second intervals. Notably, this dataset undergoes real-time updates, continuously reflecting the latest market conditions. This facet endows researchers with the capacity for real-time analysis of market dynamics and patterns, enabling a heightened precision in tracking the instantaneous state of the market.

### *3.1.2. Application*

In this study, the Alpha Vantage dataset serves as an illustrative exemplar, elucidating the application of novel machine learning methodologies employed for the prediction of stock price trends. The procedural intricacies of acquiring data from Alpha Vantage are exhaustively expounded, further undergoing preprocessing and feature engineering to enhance data quality. Subsequently, we harness cutting-edge deep learning models, such as the Transformer, to execute market forecasting endeavors. Inclusive in this exploration is a comprehensive evaluation of the model's performance across varying temporal scales, coupled with a comprehensive exploration of its aptitude in addressing the challenges presented by high-frequency market volatility. Through the application of the Alpha Vantage dataset, we not only accentuate the challenges stemming from market volatility and information incompleteness but also delineate how novel methodologies adeptly confront these challenges.

The analysis facilitated by the Alpha Vantage dataset vividly exposes the salient challenges posed by market dynamics and data incompleteness, concurrently portraying the efficacy of innovative methodologies in countering these challenges. This empirical undertaking substantiates the practicality of the adopted machine learning techniques in furnishing robust and precise insights within the domain of market prediction and analysis.

## *3.2. High-frequency trading and algorithmic trading*

In the realm of financial innovation, high-frequency trading (HFT) and algorithmic trading have emerged as vanguard subjects, persistently endeavoring to forge nimble and intelligent trading models capable of navigating the complexities of the market landscape. Eminent scholarship in this domain, exemplified by the work of Chakole et al. in 2016 [6], has yielded noteworthy accomplishments through the fusion of reinforcement learning principles with trend-following strategies. Nonetheless, as the landscape evolves, the ascendancy of deep reinforcement learning has steered the trajectory of research. Methodologies such as the Deep Q Network (DQN) and Proximal Policy Optimization (PPO) have surfaced as instrumental conduits to more pronounced advancements within the sphere of high-frequency trading.

Analogously, He Qidong's seminal contribution in 2017 [7] introduced a pioneering paradigm—a deep reinforcement learning stock trading model synergistically amalgamated with trend analysis—to unearth optimal trading strategies. However, within the contemporary milieu, the unfolding panorama heralds a surge of more sophisticated reinforcement learning algorithms, among which Trust Region Policy Optimization (TRPO) and Soft Actor-Critic (SAC) have assumed pivotal roles in HFT, rendering a discernible footprint through their attainment of more remarkable outcomes.

## *3.3. Risk management and portfolio risk analysis*

Within the expanse of financial discourse, risk management and portfolio risk analysis have perennially stood as pivotal domains, underpinning the imperative of evaluating and mitigating risk through the prism of risk models and optimization methodologies, while concurrently maximizing the yields of portfolios. Classical paradigms, such as the value at risk (VaR) model and the Markowitz mean-variance model, have historically assumed a pivotal role within this milieu. However, the advent of machine learning has ushered in a transformative era, imbuing the arena with an influx of data mining techniques.

These encompass, among others, the utilization of deep learning for volatility prediction and the application of adversarial networks for the generation of risk scenarios.

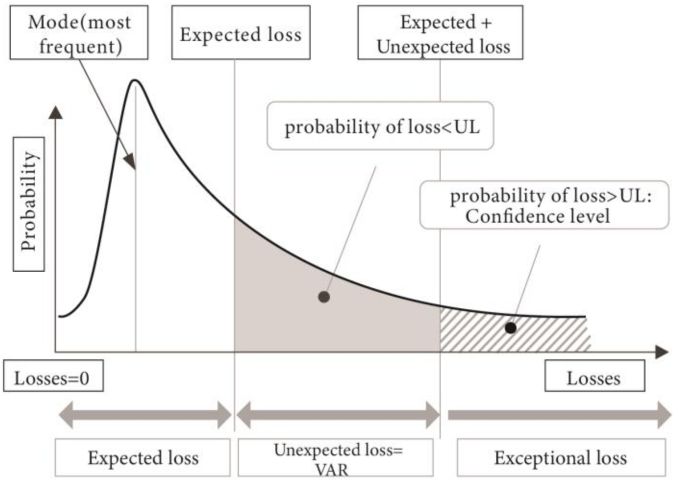

**Figure 1.** Value at risk model (VaR).

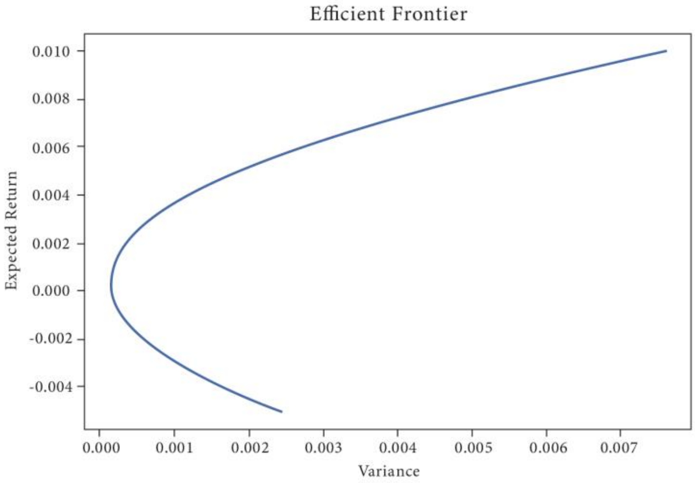

**Figure 2.** Markowitz mean variance model.

In the ambit of empirical investigation [8], the GARCH model, LSTM model, and GCN model have been harnessed in the pursuit of constructing optimal portfolios, each yielding a discernible impact across distinct market stages. Nevertheless, the ever-evolving landscape of research is animated by an enduring quest to unravel the nuanced potentialities of these models, probing how they might be optimally harnessed within the sphere of risk management and portfolio optimization to encapsulate the dynamic facets of financial markets with heightened fidelity. Of particular salience is the escalating prominence of the Graph Convolutional Network (GCN) model, which has swiftly attained the mantle of a favored option in both scholarly and pragmatic applications. This model, distinguished by its stability, measured volatility, and adaptability to divergent market contexts, crystallizes as an embodiment of contemporary prowess.

Recently, Wang Da and Zhou Yingxue [9] constructed a risk identification model based on the macroeconomic data set of 25 characteristic variables from 17 countries including the United States to conduct a comprehensive analysis of China's systemic risk probability. The empirical results show that the GDL model is significantly better than the traditional logistic regression model in capturing systemic risk, and can better depict the trend of risk probability in China.

### *3.3.1. Macroeconomic data set of 25 characteristic variables for 17 countries*

Data description: Referring to Fan Xiaoyun [10] and Laeven &Valencia [11], countries with serious systemic risks (risk loss exceeding 10% of GDP) or multiple risks occurring during 1990-2002 were selected as training samples, and a total of 16 countries were screened out. They include South Korea, Malaysia, Thailand, Turkey, Mexico, Paraguay, Venezuela, Finland, Japan, Ecuador, Jamaica, Bulgaria, Hungary, Brazil, Argentina and Indonesia. Finally, the training model is used to analyze China's systemic risk comprehensively. It should be noted that in order to better determine whether the final model can accurately identify risks outside of the training time, the sample time of the above training set and test set is 1990-2012, while the sample time of China is 1990-2019. For the features of the machine learning model, 25 indicators were determined on the basis of data availability, covering multiple dimensions of macroeconomic finance.

Model building: Firstly, comparing the performance of traditional logistic regression model and gradient lifting tree model, and exploring the optimal risk lead time; secondly, adjusting the generalization ability of the model hyperparameter optimization model; finally, applying the trained model for China risk identification.

Conclusion: The mainstream machine learning model of gradient lift tree is applied to the problem of systemic risk identification. The sample data composed of 25 features from 16 countries such as Argentina during 1990-2012 is trained and assisted by the same feature data from the United States during 1990-2012 for testing. A systematic risk identification model is constructed. Based on this model, the systematic risk probability of China during 1991-2019 is comprehensively analyzed. The results show that the machine learning model is superior to the logistic regression model in terms of prediction accuracy, generalization performance and capturing ability of systemic risk, and the established model can accurately capture the changing trend of the probability of systemic risk in China.

## 4. Limitations of financial data mining methods

The application of financial data mining in investment decision-making has great potential, but it also faces various challenges and limitations. Through a comprehensive review of the existing literature, this paper describes the main challenges and limitations encountered in financial data mining, which can be classified as follows.

### *4.1. Data quality and reliability*

The quality and reliability of financial data serve as fundamental prerequisites for conducting data mining. Financial data is often sourced from various channels and originates from multiple dimensions and attributes, rendering it susceptible to issues such as noise, missing values, and outliers. Moreover, the reliability of financial data is influenced by the credibility of data sources and the data acquisition process. In their study on sentiment analysis methods, Wang Ting and Yang Wenzhong [12] highlight the importance of addressing data quality issues in sentiment analysis. Given the subjectivity and diversity of sentiment texts, challenges arise in mitigating problems related to text noise and subjective sentiment labeling. Hence, ensuring the accuracy, completeness, and consistency of data emerges as a critical challenge in financial data mining.

### *4.2. Implicit information and market irrationality*

Financial markets are subject to numerous factors that involve hidden information and market irrationality. Factors such as investor sentiment, market manipulation, and media influence can significantly impact market prices and trends. However, accurately quantifying and capturing these factors proves challenging. Moreover, irrational behavior in financial markets can lead to abnormal volatility and inefficient pricing. In 2021, Xiong Yuning [13] conducted research on multi-channel investor sentiment analysis methods specifically targeting the stock market. The study emphasized the significant influence of implicit information and market sentiment irrationality in text data, necessitating in-depth exploration of sentiment and subjective information within textual data. This exploration includes investigating the correlation between social media sentiment and news media sentiment, as well as the impact of sentiment intensity.

### *4.3. Dimensional disasters and data sparsity*

Financial data often exhibits a high-dimensional feature space, which gives rise to challenges associated with the curse of dimensionality and data sparsity. In high-dimensional spaces, the distances between data points become large, making pattern recognition and prediction difficult. In a study conducted by Tao Jiaming [14], a three-stage feature selection algorithm called Relief-PIMP-Pearson/Mic was employed to mine relevant factors from quarterly financial statements and stock trading data, resulting in improved accuracy in predicting high stock dividends. However, due to the sparsity of financial data, the applicability of the algorithm on the dataset still requires further investigation.

## 5. Conclusion

At present, more and more enterprises and individuals apply financial data mining technology in the field of investment. The technology is used for trend prediction and risk assessment through phases of problem definition, data collection, data preprocessing, algorithm modeling, model training and evaluation. This process not only involves the application of machine learning and deep learning, but also combines methods such as natural language processing and time series analysis to achieve the cross-fusion of multiple technologies.

This paper reviews the application of financial data mining in investment decision making, and discusses the challenges and limitations in this field. Literature review shows that financial data mining is widely used in market prediction and stock selection, high-frequency trading and algorithmic trading, as well as risk management and portfolio risk analysis. However, in practical applications, financial data mining faces challenges such as data quality and reliability, dimensional disaster and data sparsity, as well as implied information and market irrationality.

In view of the above limitations, future financial data mining research can be broken through from the following aspects:

(1) Introduction of domain expert knowledge: Incorporating the knowledge of domain experts into the process of financial data mining can provide a deep understanding and interpretation of financial data, thereby improving data quality and effectively conducting feature selection. Combining domain knowledge with data mining technology can improve the accuracy and interpretability of financial data mining methods.

(2) Integrate multi-source data: Integrate data from multiple sources such as financial data, market indicators, social media data and news public opinion data to obtain a more comprehensive and diversified feature representation. At the same time, research on how to effectively deal with high dimensional and sparse data, such as using graph-based methods and deep learning methods.

(3) Integrated sentiment analysis and unstructured data mining: Combined with natural language processing and machine learning technology, emotion analysis and emotion factor extraction are carried out on financial news, social media data and company reports to better understand the impact of market sentiment and market irrational factors on investment decisions.

(4) Personalized investment decision support system: Develop personalized investment decision support system for individual investors, combine individual investors' preferences, risk tolerance and goals, and provide them with personalized investment advice and optimized portfolio allocation. Through personalized investment decision support, we can better meet the needs of investors, and improve the effect and credibility of investment decisions.

Financial data mining has great potential in investment decision-making. With continued research and practice, financial mining technology will become an indispensable tool in investment decisions, leading to better outcomes and higher returns for investors and financial institutions.

## References

[1] Zhu, T. and Luo, S. (2015) Application of data mining technology in university management decision. Computer age, (3): 39 - 40.

[2] Gong, G. (2016) Discussion on the application mode of data mining in the management decision of university students. Information and computer, (3): 135 - 136.

[3] Yan, F. (2018) Empirical analysis of Chinese value stock investment. Time finance, 709(27): 162 - 166.

[4] Jin, Y. (2017) Stock market analysis and prediction based on data mining technology. Jilin University of Finance and Economics.

[5] Zhang, W. (2022) Stock time series prediction based on long short-term memory neural network. Information and Computers (Theory), (9): 68 - 72.

[6] Chakole, J. and Kurhekar, M. (2020) Trend Following Deep Q-Learning Strategy for Stock Trading. Expert Systems, 37.

[7] He, Q. (2022) Deep reinforcement learning stock trading strategies combined with trends. Computer science and applications, (3): 673 - 681.

[8] Tang, Y. (2022) Research on optimal portfolio based on graph Convolutional neural network. Lanzhou University of Finance and Economics.

[9] Wang, D. and Zhou, Y. (2023) Can machine learning methods identify the probability of systemic financial risk in China? Financial market research, 134: 48 - 58.

[10] Fan, X. (2006) Behind prosperity -- the essence, measurement and management of financial systemic risk. China Finance Press, Beijing.

[11] Laeven L. and Valencia F. (2013) Systemic banking crises database. IMF Economic Review, 61(2): 225 - 270.

[12] Wang, T. and Yang, W. (2021) A review of text sentiment analysis methods. Computer engineering and applications, (12): 11 - 24.

[13] Xiong, Y. (2021) Research on multi-channel investor sentiment analysis for stock market. Southwest University of Finance and Economics.

[14] Tao, J. (2022) Research on high-transfer stock prediction based on three-stage feature selection and Enhanced GBDT-Logit: A case study of China A-share market. Zhejiang Gongshang University.