Journal of Economic Theory and Business Management
ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online) | Vol. 1, No. 3, 2024

Anomaly Detection Method for High-Frequency Financial Market Volatility Data Based on LLM

RATHORE, Vikramaditya Singh [1*] SHARMA, Kritika [2] VERMA, Siddharth [3]

[1] Jawaharlal Nehru University, India
[2] Indian Institute of Management, India
[3] Indian Institute of Technology, India

**** RATHORE, Vikramaditya singh is the corresponding author, E-mail: VIiill629@gmail.com***

Abstract: This article introduces the importance and challenges of detecting anomalies in high-frequency volatility data in financial markets. Traditional methods such as SV and GARCH models have been unable to cope with the rapidly changing and increasing complexity of the market environment, so new strategies must be developed to identify abnormal fluctuations quickly. This paper proposes a method based on local linear mapping (LLM), which aims to improve anomaly detection accuracy, monitor market fluctuations in real-time, and identify potential risk events, enhancing investment decisions and promoting financial market stability and sustainable development.

Keywords: Financial Market Volatility, High-frequency Data, Anomaly Detection, Large Language Models.

DOI: https://doi.org/10.5281/zenodo.11636947

1 Introduction

In financial markets, anomalous volatility behavior can arise from changes in market sentiment, significant events, or other factors, making volatility a crucial risk metric [1-3]. Detecting anomalous volatility enables timely identification of erratic market behavior, aiding financial institutions and investors in risk management to mitigate potential losses. Moreover, it enhances understanding and interpretation of market dynamics, facilitating adjustments in investment strategies for improved decision-making accuracy and success rates. Traditionally, scholars have relied on low-frequency data to study various financial market issues, employing SV models (Stochastic Volatility models) and GARCH models (Generalized Autoregressive Conditional Heteroskedasticity models) to analyze market volatility effectively. However, with rapid financial market evolution and globalization, volatility changes have become more frequent and complex, surpassing the capabilities of low-frequency data to track trends and detect anomalies in real time.

Therefore, developing methods for detecting anomalies in high-frequency financial market volatility data is crucial. Feng Huawei incorporated variance terms into traditional Random Forest models and utilized Deep Recursive Neural Networks (Deep RNNs) for data training [4-6], achieving a minimum 2% improvement in F1 scores compared to other algorithms. From a fund flow perspective, Ji Xun addressed data scarcity issues in anomaly detection models within financial systems by employing methods such as Support Vector Data Description (SVDD) and Principal Component Analysis (PCA) [7], conducting empirical analyses to detect and analyze anomalies in financial systems. However, these methods exhibit high computational complexity and poor real-time performance. The Local Linear Mapping (LLM) model effectively handles various financial market volatility data characteristics, including fat tails, jumps, and non-stationarity, making it a robust tool for analyzing complex data. This positions the LLM model as advantageous for applications in high-frequency data. Therefore, this paper proposes an anomaly detection method based on LLM [8] for high-frequency financial market volatility data. The technique aims to enhance anomaly detection accuracy, enable real-time monitoring of market volatility, identify potential risk events, and improve investment decision-making, contributing to financial market stability and sustainable development.

2 Related Work

2.1 High-frequency data preprocessing of financial market volatility

When selecting an anomaly detection method, it is essential to consider the characteristics of the samples carefully. High-frequency financial volatility data often lacks negative samples, making it unsuitable for traditional binary classification methods. Instead, Support Vector Data Description (SVDD) [9-11], which is suitable for low-dimensional data, addresses the issue of severely imbalanced data by mapping samples into a higher-dimensional space and enclosing positive samples with a

minimum hyper-sphere while introducing slack variables and penalty parameters for optimization to avoid overfitting [12]. For high-dimensional data, Principal Component Analysis (PCA) facilitates dimensionality reduction and captures sample variation by calculating deviations from the central position of the samples. SVDD and PCA effectively address anomaly detection problems across different data dimensions.

This study utilizes financial transaction data from a company on the Kefengdai platform in 2022, focusing on credit information indicators across 30 attributes. PCA is employed to reduce the data dimensions into a low-dimensional feature representation. Subsequently, an unsupervised learning approach constructs a single-class classification model, utilizing one-class [13] SVM as an SVDD model to describe the dataset. Since average data constitutes a large proportion in anomaly detection while abnormal data is relatively small, addressing class imbalance is crucial. Moreover, cautious handling is paramount due to the potential impact of misidentifying abnormal data in anomaly detection.

$$C = (P_r - \mu) / \sigma_r \tag{1}$$

To ensure comparability of all high-frequency financial market volatility data targets in metric values, standardizing daily average sequence data for each high-frequency [14] data target is performed before computing indicators.

2.2 Filtering and Processing of High-Frequency Financial Market Volatility Data

When performing graph analysis on high-frequency financial market volatility data, the constructed graph network contains many abnormal and normal data, making it challenging to mine the hidden anomalous data targets [5]. A subgraph filtering method is employed to process the high-frequency financial market volatility data to address this issue. The specific method is as follows [15-17]:

Edges' attribute values are extracted across dimensions and filtered using ReLU layers. Specifically, for the i-th dimension, let the edge attribute feature value between two nodes be \(f_i \), and set a threshold \(f_{i0} \). When the feature value \(f_i \) reaches the threshold \(f_{i0} \), the edge between the nodes is retained, and edges irrelevant to the dimension are removed, thereby obtaining filtered subgraphs for each dimension. This helps understand the influence of each dimension on the relationships between nodes in the dataset, facilitating more detailed observation and interpretation of the data structure and associations in data analysis and visualization.

2.3 Feature Extraction of High-Frequency Financial Market Volatility Data Using Multi-Dimensional Graphs

Based on the data filtering process, financial market volatility high-frequency data features are extracted through multi-dimensional graph mapping. Using the connected component algorithm [18], data with edge connections are divided into the same connected component, and connected island data is mined through node and edge traversal. Filtering across dimensions is performed to divide connected subgraphs. The hierarchical features are as follows:

1. Intra-group Individual Features [19]: Includes statistical indicators such as mean, standard deviation, maximum, and minimum values used to describe the variation of each financial market volatility data.

2. Cluster Topological Features [20]: Includes the number of edges within the group, average edge weights, and degree distribution used to describe the connections and relationships between nodes in the subgraph.

3. Cluster Importance Features [22]: Using the PageRank algorithm, the influence and importance of each data point in the network are assessed based on its connectivity with other nodes, calculating its importance value. Data points with high impact and significance in the financial market are identified by analyzing cluster importance features, facilitating a better understanding of market changes and trends for decision-making.

Compared to the initial M single features of high-frequency financial market volatility data, this feature extraction approach effectively enhances the feature dimension. [23-25] Increasing the quantity and diversity of features enables a more comprehensive exploration of the latent information and patterns in high-frequency financial market volatility data, enhancing the analysis and prediction capabilities for these data, thus providing more robust support for decision-making and risk management.

3 Methodology

3.1 Large language model anomaly detection

This section demonstrates how a multi-agent AI framework can be applied to financial market data - specifically, a daily S&P 500 index series from 1980 to 2023. This example explains how LLM-powered multi-agent models process and analyze real-world financial data, illustrating every stage of the process from anomaly detection to final decision-making. [26] Using the well-known S&P 500 series as a test case, I aim to highlight the framework's proficiency in navigating the complexity of financial datasets. The examples provided in this section are real-world results of a fully automated, custom-developed framework.

The initial phase of the demonstration involved anomaly detection, which was performed by applying the z-score method to the daily percentage change of the S&P 500 series. Choose a deliberately high z-score threshold (10) to highlight significant outliers, ensuring attention to the most critical deviations. As a result, three outliers were identified on October 19, 1987, October 13, 2008, and March 16, 2020 (Figure 1). In addition, to challenge the recognition capability of the framework, three missing values were deliberately inserted into the data set. This approach aims to assess not only the framework's ability to identify significant anomalies but also its ability to distinguish between natural anomalies and deliberately introduced inaccuracies. [27-29] The carefully introduced legal anomalies and potential errors create a delicate testing environment that allows for a thorough assessment of the proficiency of multi-agent AI models in managing the inherent complexity of real-world financial data. Once these data points are detected, they are converted into a format suitable for machine processing, as shown in Table 1, laying the foundation for subsequent analysis by the AI agent.

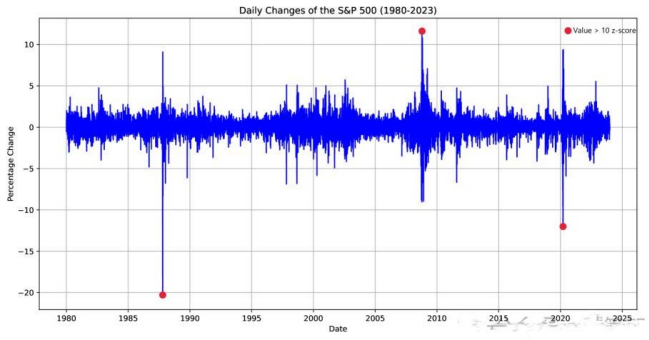

Figure 1. Anomalies identified in the S&P500 series

While preparing input data, integrating tabular data and its corresponding metadata is critical to the effectiveness of AI systems. Acquiring metadata - including details such as the name, origin, frequency, description, and data coverage - is necessary for AI to understand and contextualize the detected anomalies fully. This integration allows the AI system to interpret tabular data more accurately and maximize the knowledge gained during LLM pre-training.

3.2 Develop agents for data issues

After receiving the exception data and associated metadata, the agent responsible for formulating the data problem plays a crucial role in the initial phase of exception verification. The agent's output includes questions designed to explore the validity and context of the identified exception. The questions developed by the agent serve various purposes: they aim to confirm the nature of the detected anomalies, understand their significance in the historical and market context, and prepare relevant large-scale language models for further validation. The table below shows how agents guide and respond to unusual events within the S&P 500 index. [30] The output generated by the agent reflects the human-like response, indicating a satisfactory integration of tabular data and LLMs.

This refinement is reflected in the following key aspects:

- Situational awareness: Although there is no explicit event information in the data or metadata provided, agents infer and incorporate relevant historical context, such as Black Monday and the impact of COVID-19 on financial markets. This ability to correlate numerical anomalies with significant real-world events demonstrates an agent's situational understanding and the use of pre-trained data to enrich the analysis.
- Adaptability: The agent's problem seeks not only to verify the nature and accuracy of the anomaly data but also to intelligently speculate on possible explanations for these anomalies, such as suggesting whether the data points represent a percentage drop, a point drop, or some other measurement. This adaptability ensures a comprehensive verification process that considers the full range of possible causes of the problem.
- Efficiency: The agent effectively manages the context window by grouping similar questions and aggregating queries related to missing values into a single question. This approach optimizes interaction with subsequent LLM-based analysis phases, ensuring the problem remains within acceptable processing and analysis [31]. This efficiency is critical to maintaining the system's performance, scalability, and responsiveness.
- Pre-trained knowledge exploitation: The ability of agents to add additional information and provide possible interpretations based on pre-trained knowledge underscores the powerful integration of LLMs into the framework. This integration enables the system to leverage large amounts of historical data and insights to enhance the depth and accuracy of the anomaly verification process.

4 Conclusion

Demonstrating the potential of AI in financial market price analysis through multi-agent workflows reflects the potential for emerging technologies to improve data monitoring and anomaly detection. Integrating LLMs with traditional analytical methods can significantly improve market surveillance and decision-making accuracy and efficiency. This approach is expected to simplify data review, speed up the detection of market anomalies, and provide decision-makers with timely information. [32-33] The key to this approach's success and widespread application lies in effective metadata management and data governance. As an essential bridge, metadata can promote the transformation of tabular data into a structure conducive to LLM processing, enrich the data context, and improve the efficiency and accuracy of LLM-driven processes. Such

advances in AI-driven analysis of price data in financial markets herald a reconfiguration of data analysis and decision-making. As AI technology evolves, the future envisions a framework capable of autonomously performing increasingly complex analytical tasks, reducing the need for human supervision. [34-36] This evolution toward AI-centric approaches in financial market price data analysis is expected to streamline anomaly detection and review procedures and find applications in various fields that require sophisticated data analysis capabilities. Amid these promising developments and the prospect of AI in financial market price analysis, it is crucial to emphasize the indispensable role of human oversight in the development stage of AI technology. [37] The demands for accuracy, accountability, and adherence to ethical standards in AI applications call for vigilant human oversight. As AI systems gain autonomy and become more integrated in decision-making, the potential for systemic bias, inaccuracies, and unexpected outcomes underscores the need for continued human engagement. This engagement is necessary to validate AI outputs and steer these technologies in a direction that meets ethical standards and societal values.

Acknowledgments

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author Contributions

Not applicable.

About the Authors

RATHORE, Vikramaditya Singh

Financial information, Jawaharlal Nehru University (JNU), New Delhi, India.

SHARMA, Kritika

Business Administration, Indian Institute of Management (IIM) Bangalore, India.

VERMA, Siddharth

Electronic information engineering, Indian Institute of Technology (IIT) Kanpur, India.

References

[1] He, Zheng, et al. "Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering."

[2] Zhou, Y., Zhan, T., Wu, Y., Song, B., & Shi, C. RNA Secondary Structure Prediction Using Transformer-Based Deep Learning Models.

[3] Qian, K., Fan, C., Li, Z., Zhou, H., & Ding, W. (2024). Implementation of Artificial Intelligence in Investment Decision-making in the Chinese A-share Market. Journal of Economic Theory and Business Management, 1(2), 36-42.

[4] Li, Zihan, et al. "Robot Navigation and Map Construction Based on SLAM Technology." (2024).

[5] Fan, C., Ding, W., Qian, K., Tan, H., & Li, Z. (2024). Cueing Flight Object Trajectory and Safety Prediction Based on SLAM Technology. Journal of Theory and Practice of Engineering Science, 4(05), 1-8.

[6] Fan, C., Li, Z., Ding, W., Zhou, H., & Qian, K. Integrating Artificial Intelligence with SLAM Technology for Robotic Navigation and Localization in Unknown Environments.

[7] Liu, B., Cai, G., Ling, Z., Qian, J., & Zhang, Q. Precise Positioning and Prediction System for Autonomous Driving Based on Generative Artificial Intelligence.

[8] Tian, J., Qi, Y., Li, H., Feng, Y., & Wang, X. (2024). Deep Learning Algorithms Based on Computer Vision Technology and Large-Scale Image Data.Journal of Computer Technology and Applied Mathematics,1(1), 109-115.

[9] Xu, Jiahao, et al. "AI-BASED RISK PREDICTION AND MONITORING IN FINANCIAL FUTURES AND SECURITIES MARKETS." The 13th International scientific and practical conference “Information and innovative technologies in the development of society”(April 02–05, 2024) Athens, Greece. International Science Group. 2024. 321 p.. 2024.

[10] Wang, Yong, et al. "Machine Learning-Based Facial Recognition for Financial Fraud Prevention." Journal of Computer Technology and Applied Mathematics 1.1 (2024): 77-84.

[11] Song, Jintong, et al. "LSTM-Based Deep Learning Model for Financial Market Stock Price Prediction." Journal of Economic Theory and Business Management 1.2 (2024): 43-50.

[12] Bai, Xinzhu, Wei Jiang, and Jiahao Xu. "Development Trends in AI-Based Financial Risk Monitoring Technologies." Journal of Economic Theory and Business Management 1.2 (2024): 58-63.

[13] Jiang, W., Yang, T., Li, A., Lin, Y., & Bai, X. (2024). The Application of Generative Artificial Intelligence in Virtual Financial Advisor and Capital Market Analysis. Academic Journal of Sociology and Management, 2(3), 40-46.

[14] Wang, X., Tian, J., Qi, Y., Li, H., & Feng, Y. (2024). Short-Term Passenger Flow Prediction for Urban Rail Transit Based on Machine Learning.Journal of Computer Technology and Applied Mathematics,1(1), 63-69.

[15] Feng, Y., Li, H., Wang, X., Tian, J., & Qi, Y. (2024). Application of Machine Learning Decision Tree Algorithm Based on Big Data in Intelligent Procurement.

[16] Cui, Z., Lin, L., Zong, Y., Chen, Y., & Wang, S. Precision Gene Editing Using Deep Learning: A Case Study of the CRISPR-Cas9 Editor.

[17] Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. Predictive Optimization of DDoS Attack Mitigation in Distributed Systems using Machine Learning.

[18] Wang, Y., Zhu, M., Yuan, J., Wang, G., & Zhou, H. (2024). The intelligent prediction and assessment of financial information risk in the cloud computing model. arXiv preprint arXiv:2404.09322.

[19] Li, H., Wang, X., Feng, Y., Qi, Y., & Tian, J. (2024). Driving Intelligent IoT Monitoring and Control through Cloud Computing and Machine Learning.arXiv preprint arXiv:2403.18100.

[20] Qi, Y., Wang, X., Li, H., & Tian, J. (2024). Leveraging Federated Learning and Edge Computing for Recommendation Systems within Cloud Computing Networks. arXiv preprint arXiv:2403.03165.

[21] Ding, W., Zhou, H., Tan, H., Li, Z., & Fan, C. (2024). Automated Compatibility Testing Method for Distributed Software Systems in Cloud Computing.

[22] Ding, W., Tan, H., Zhou, H., Li, Z., & Fan, C. Immediate Traffic Flow Monitoring and Management Based on Multimodal Data in Cloud Computing.

[23] Qi, Y., Feng, Y., Tian, J., Wang, X., & Li, H. (2024). Application of AI-based Data Analysis and Processing Technology in Process Industry.Journal of Computer Technology and Applied Mathematics,1(1), 54-62.

[24] Tian, J., Li, H., Qi, Y., Wang, X., & Feng, Y. Intelligent Medical Detection and Diagnosis Assisted by Deep Learning.

[25] Yu, D., Xie, Y., An, W., Li, Z., & Yao, Y. (2023, December). Joint Coordinate Regression and Association For Multi-Person Pose Estimation, A Pure Neural Network Approach. In Proceedings of the 5th ACM International Conference on Multimedia in Asia (pp. 1-8).

[26] Cheng, Qishuo, et al. "Monetary Policy and Wealth Growth: AI-Enhanced Analysis of Dual Equilibrium in Product and Money Markets within Central and Commercial Banking." Journal of Computer Technology and Applied Mathematics 1.1 (2024): 85-92.

[27] Li, Huixiang, et al. "AI Face Recognition and Processing Technology Based on GPU Computing." Journal of Theory and Practice of Engineering Science 4.05 (2024): 9-16.

[28] Qin, Lichen, et al. "Machine Learning-Driven Digital Identity Verification for Fraud Prevention in Digital Payment Technologies." (2024).

[29] Wang B, Lei H, Shui Z, et al. Current State of Autonomous Driving Applications Based on Distributed Perception and Decision-Making[J]. 2024.

[30] Chen, Zhou, et al. "Application of Cloud-Driven Intelligent Medical Imaging Analysis in Disease Detection." Journal of Theory and Practice of Engineering Science 4.05 (2024): 64-71.

[31] Bao, Wenqing, et al. "The Challenges and Opportunities of Financial Technology Innovation to Bank Financing Business and Risk Management." Financial Engineering and Risk Management 7.2 (2024): 82-88.

[32] Cui, Z., Lin, L., Chen, Y., Wang, S., & Zong, Y. (2024). Drug Screening and Target Prediction Based on Machine Learning.

[33] Lu, W., Ni, C., Wang, H., Wu, J., & Zhang, C. (2024). Machine Learning-Based Automatic Fault Diagnosis

Method for Operating Systems.

[34] Shi, Y., Li, L., Li, H., Li, A., & Lin, Y. (2024). Aspect-Level Sentiment Analysis of Customer Reviews Based on Neural Multi-task Learning. Journal of Theory and Practice of Engineering Science, 4(04), 1-8.

[35] Ni, C., Zhang, C., Lu, W., Wang, H., & Wu, J. (2024). Enabling Intelligent Decision Making and Optimization in Enterprises through Data Pipelines.

[36] Bao, Q., Wei, K., Xu, J., & Jiang, W. (2024). Application of Deep Learning in Financial Credit Card Fraud Detection. Journal of Economic Theory and Business Management, 1(2), 51-57.

[37] Bai, X., & Braganza, M. (2024). Application of Artificial Intelligence Technology in Financial Risk Control. Academic Journal of Sociology and Management, 2(3), 47-53.