European Journal of AI, Computing & Informatics

Vol. 1 No.1 2025

Article **Open Access**

The Application of Machine Learning Combined with Graph Databases in Financial Technology

Xuanrui Zhang [1,*]

[1] College of Engineering, University of California, Berkeley, CA, 94720, USA
* Correspondence: Xuanrui Zhang, College of Engineering, University of California, Berkeley, CA, 94720, USA

Received: 02 April 2025
Revised: 05 April 2025
Accepted: 18 April 2025
Published: 22 April 2025

Abstract: Against the backdrop of rapid advances in financial technology, the surge in data volume and the diversity of data structures have posed great challenges to traditional data analysis methods. The combination of machine learning and graph databases provides a new means of data processing and analysis for the financial industry. Graph databases have unique advantages in handling complex relational data, while machine learning can mine deep patterns in large amounts of data. This article mainly studies the application of these two technologies in the field of financial technology, focusing on enhancing data analysis, financial forecasting, and risk control capabilities. By combining these two technologies, financial institutions can more efficiently complete tasks such as fraud detection, credit evaluation, anti-money laundering, and marketing. Detailed discussions were held on the technical architecture and methods required to implement these technologies, including key technologies such as data integration, graph model construction, and distributed computing, to help the financial industry move towards intelligent operations.

Keywords: machine learning; graph database; financial technology; fraud detection; risk management

1. Introduction

With the rapid development of financial technology, the scale and complexity of data involved in the financial industry have significantly increased, making traditional data processing methods difficult to meet practical needs [1]. As a core component of the field of artificial intelligence, machine learning has demonstrated strong practicality in financial forecasting and optimization. Meanwhile, graph databases are particularly important in handling complex network analysis due to their outstanding performance in storing and retrieving associated data. The integration of machine learning and graph database technology has greatly improved the efficiency of mining potential patterns and relationships in data, and can also bring new technological solutions for financial decision-making and risk control, promoting the development of the financial industry towards intelligence.

2. The Importance of Combining Machine Learning with Graph Databases

2.1. Enhance Data Analysis and Decision-Making Abilities

The integration of machine learning and graph database technology highlights outstanding performance in the field of data analysis, especially in analyzing complex relational data [2]. By using the node and connection architecture of graph databases to store information, the interconnection of data becomes clearer and more intuitive. Machine learning technology, on the other hand, can delve into these intricate data through its algorithm models, unearth hidden patterns and patterns, and assist in data analysis. This combination of technologies enables enterprises to quickly uncover deep connections between data and also cope with unstructured or changing datasets that are difficult to handle with traditional analytical methods. In the financial field, using graphical databases to record the correlation between customers and transactions, and integrating intelligent analysis technology, can effectively reveal hidden fund flow trajectories and discover potential risk factors, which helps companies make more accurate strategic decisions [3].

2.2. Enhance Financial Forecasting and Risk Control Capabilities

In the financial field, predicting market dynamics and effectively managing risks have become the core aspects of work. Integrating machine learning techniques into graph databases can extract important data from complex financial relationship networks. Graph databases, with their efficient storage capabilities, manage customer transaction information, credit chains, and other data, presenting a multidimensional structure of financial relationship networks. Meanwhile, machine learning can detect unusual patterns or market trends in trading networks through algorithm training. For example, in the field of credit risk assessment, graph databases can assist in building a network diagram of customer credit connections, while machine learning predicts possible default risks or abnormal fund flows by mining customer interaction behavior and past activity records. This combination of technology endows financial institutions with more comprehensive and accurate risk identification capabilities in the face of market changes [4].

3. Technical Architecture and Methods Combining Machine Learning with Graph Databases

3.1. Data Integration and Processing

When integrating machine learning and graph database technology, data integration and processing constitute a key part of the technical framework, with the aim of integrating heterogeneous data from multiple sources into a unified structure, thereby optimizing the efficiency of data utilization. In the financial field, data often involves multiple sources such as user profiles, transaction history, and market information, which generally suffer from high complexity and inconsistent standards. After the data integration process, the data can be cleaned, deduplicated, and formatted, and the processed data can be imported into a graph database for graphic analysis and machine learning algorithm training [5].

For example, a bank needs to consolidate customer transaction details, equipment usage records, and geographic location information from different business departments in its anti-fraud system. The formats of these information sources are inconsistent, and there is a common occurrence of data redundancy and incompleteness. After steps such as data cleaning and feature extraction, the system unifies this information and creates a graph database. Transaction details and device information are presented in the form of nodes, while the association between customer accounts is connected through edges. In the process of model training, to quantify the strength of the correlation between different data sources, the following formula is used:

$$W = \sum_{i=1}^{n} (Ai \cdot Bi) \tag{1}$$

Among them, W represents the calculated weight, and Ai and Bi represent the values of the features (e.g., customer transaction behavior or credit score) in the data

source. This formula helps quantify the strength of relationships between data and provides input data for subsequent machine learning models. Through this method, financial institutions can conduct integrated research on data information from multiple channels, thereby optimizing the speed and efficiency of data processing.

3.2. Integration of Graph Models and Machine Learning

Graph databases store complex associated information through the architecture of nodes and edges, while machine learning techniques explore the inherent patterns of nodes and edges within these graphs. In the field of financial technology, the integration of graph models and machine learning mainly uses graph convolutional neural networks (GCN) and other methods to analyze user transactions and behavior patterns, enabling risk estimation and customer analysis.

For example, a bank stores customers' account information, transaction history, and social network data in a graph database, connecting customers through transaction behavior and guarantee relationships. With the help of these associated data, a graphical model is constructed, and machine learning algorithms (graph convolutional networks) can analyze the node attributes and connection methods within these graphical structures, thereby predicting customers who may be at risk.

$$gv = \sigma(W \cdot \sum_{u \in N(v)} g_u + b) \quad (2)$$

Among them, gv represents the representation of node v, $N(v)$ is the neighboring node of node v, W is the weight matrix, σ is the activation function, and b is the bias term. With this integration strategy, banks can screen out potentially high-risk customer groups, identify unusual fund flow activities through analyzing transactions between customers, and enhance the accuracy and effectiveness of risk prediction.

3.3. Real Time Data Flow and Dynamic Analysis

In the combination of machine learning and graph databases, real-time data streams and dynamic analysis occupy a core position, and its technical framework relies on complex data processing stages, aiming to provide fast and real-time decision assistance. During specific execution, the data collection module is responsible for collecting real-time data streams from transaction logs, social platform dynamics, and other channels. The raw data is cleaned and standardized by the data preprocessing module, ensuring the accuracy and completeness of the data. The data stream is transmitted through the distribution module to the graph database for storage, which stores nodes and associated information in the graph model. The real-time analysis module adopts graph traversal algorithm and machine learning technology to comprehensively mine the stored data. At the same time, the prediction and feedback module will provide analysis results to the business system, which can be applied to various business scenarios such as fraud identification and risk assessment. Figure 1 shows the main process and key steps of real-time data flow and dynamic analysis.

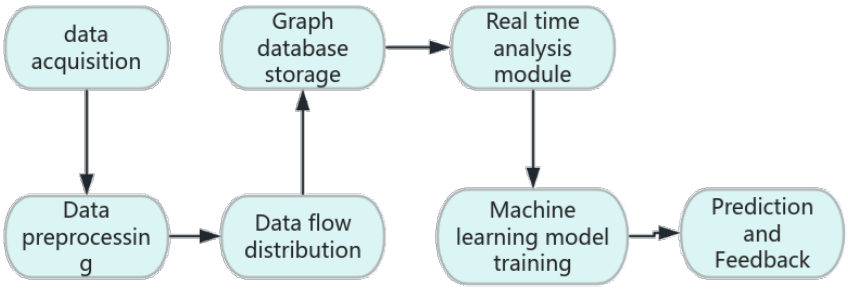

Figure 1. Real Time Data Flow and Dynamic Analysis Flowchart.

3.4. Distributed Computing and Performance Optimization of Graph Database

To cope with large-scale datasets and complex relational networks, graph databases adopt a distributed computing structure to improve storage and processing capabilities. This database slices graphic information and distributes it to different server nodes, using distributed computing frameworks (Apache Spark or Hadoop) to achieve cross node computing tasks and accelerate data processing speed.

For example, a securities company has introduced advanced distributed processing architecture in its anti-fraud detection system to cope with massive financial transaction data. The system achieves efficient retrieval and analysis of transaction links by dispersing customer, transaction, and account data to different server nodes, and significantly improves processing speed through parallel processing technology. The system adopts graph traversal, index optimization, and data caching strategies to enhance performance when performing data retrieval tasks. In this distributed computing architecture, the update of node characteristics is carried out through the following formula:

$$v_i = \frac{\sum_{j \in N(i)} e_{ij} \cdot v_j}{\sum_{j \in N(i)} e_{ij}} \tag{3}$$

Among them, v_i represents the updated feature value of node i, $N(i)$ is the set of neighbors of node i, and e_{ij} represents the weight of the edge between node i and neighboring node j, where v_j is the feature value of the neighboring node. This formula describes the updating process of node features through weighted averaging of neighbor features, and the weights of edges reflect the importance of relationships between nodes. By utilizing a distributed architecture for data storage and processing, this method demonstrates extremely high efficiency in handling large amounts of graph data.

4. The Specific Application of Combining Machine Learning with Graph Databases in Financial Technology

4.1. Financial Fraud Detection

The detection of fraudulent behavior in the financial field can be achieved through the integration of graph databases and machine learning technology, enabling efficient identification of potential threats in complex transaction networks. This technology significantly enhances the detection accuracy of fraudulent behavior by constructing a correlation graph covering multiple dimensions such as users, trading accounts, operating devices, and geographic locations, and combining machine learning algorithms to conduct in-depth analysis of various nodes and connection paths in the network.

Taking credit card transactions as an example, financial institutions can use graph databases to construct a connection graph between customers, transaction terminals, and accounts, transforming each transaction into an edge in the graph that connects the initiator and receiver of the transaction. On this basis, machine learning techniques such as graph convolutional neural networks or anomaly detection algorithms are applied to model and analyze graph data, enabling real-time detection of abnormal transaction paths. In the flow of funds involving multiple accounts, if the flow of funds along a certain path shows an abnormal concentration trend, especially between several specific nodes, it often indicates the existence of illegal activities such as "fake account" operations or "money laundering". The fraud probability score of the transaction path can be calculated by multiplying the weights of each side in the path:

$$P_i = \prod_{(j,k) \in R(i)} w_{jk} \tag{4}$$

Among them, P_i represents the fraud probability score of paths i, $R(i)$ is the set of all edges on path i, and w_{jk} is the edge weight between nodes j and k on path i. By calculating the total weight of the transaction path, we can evaluate the likelihood of anomalies along the path. The edge weight can be defined based on characteristics such as transaction frequency, amount, and regional anomalies to comprehensively evaluate the degree of suspicion of each transaction in the path. In practice, this technology can

quickly identify high-risk abnormal paths in the transaction network, particularly for detecting complex behaviors such as cross-account money laundering or circular fraud.

4.2. Credit Rating and Risk Management

Integrating machine learning technology with graphical databases in the field of credit scoring and risk management can provide financial institutions with more accurate credit scoring and decision assistance by analyzing the complex relationships and behavioral characteristics between customers. The graph database constructs a relationship graph covering multiple dimensions such as customers, loans, transactions, and guarantees. Machine learning effectively identifies potential credit risk models through in-depth mining of associations in the graph.

For example, a bank has developed a risk control platform based on graph databases and machine learning to optimize its credit evaluation system. When a customer submits an application for a credit loan, the platform will activate and retrieve the applicant's social and business association information through the graph database, covering joint guarantee individuals, business relationships, and previous repayment behavior. Through in-depth analysis using artificial intelligence algorithms, the platform identified abnormal fund flow patterns in the applicant's network, such as frequent large transfers by certain parties involved in the transactions or negative credit histories of guarantors. Based on these analyzed data, the platform has formed preliminary recommendations for credit ratings and prompted loan review agencies to conduct more detailed review procedures for relevant applicants.

Based on the classification of credit scores and risk management, Table 1 shows the credit score ranges, risk levels, and corresponding strategic recommendations for different customer groups.

Table 1. Customer Classification Table for Credit Scoring and Risk Management.

Customer group	Credit score range	Risk level	recommendation strategy	Loan interest rate recommendation	Approval time limit
High net worth clients	seven hundred and fifty-eight hundred and fifty	low	Approval of high limit loans	Low interest rate	Quick approval
Medium risk clients	six hundred-seven hundred and forty-nine	in	Increase credit review	Medium interest rate	Regular approval
high-risk customers	three hundred-five hundred and ninety-nine	high	Limit loan amount	High interest rates	Strict approval
Student users	five hundred and fifty-sixty	in	Provide low credit limit products	Medium interest rate	Regular approval
new user	five hundred-six hundred	high	Conduct a detailed background check	High interest rates	Strict approval

Through this table, financial institutions can accurately adjust credit policies based on customers' credit scores and risk levels, improve risk control capabilities, and optimize resource allocation.

4.3. Anti-Money Laundering and Compliance Monitoring

The combination of machine learning and graph databases can effectively identify hidden suspicious transaction paths in anti-money laundering and compliance monitoring. The graph database is responsible for building a correlation graph involving accounts, transaction information, and fund flows. Machine learning conducts in-depth analysis of trading behaviors in complex graphs, quantitatively evaluates the risk level of each path, and accurately identifies high-risk trading behaviors.

For example, a financial institution discovered in its daily transaction monitoring that these accounts attempted to distribute funds across multiple fields through continuous inter account trading operations. According to the analysis results of the graph database, funds quickly gathered into a key account in a relatively short period of time. Using machine learning algorithms to analyze trading paths and screen key features, identifying the characteristics of the trading path that are highly consistent with typical abnormal fund flow patterns. Therefore, the system determines that the path has high risk and activates the automatic risk warning mechanism. After further investigation, it was confirmed that these fund trading activities were indeed abnormal, and the bank immediately took measures to effectively avoid potential risks and hidden dangers. When evaluating the risk level of the trading path, the system uses the following risk calculation formula:

$$R_P = \sum_{(i,j)\in P} \frac{w_{ij}}{\sqrt{t_{ij}}} \tag{5}$$

Among them, R_P represents the risk score of paths P, (i,j) represents a pair of trading nodes in path P, w_{ij} is the trading weight between node i and node j, such as trading amount or risk coefficient, and t_{ij} represents the trading time interval. This formula evaluates the risk of each path by considering the inverse relationship between transaction weights and time intervals, meaning that transactions with higher weights completed in shorter timeframes are deemed higher risk. The path of completing numerous bulk transactions within a shorter time frame will be deemed high-risk, thereby enhancing the accuracy of detecting unconventional transactions.

4.4. Marketing and Customer Analysis

In marketing and customer analysis, the combination of machine learning and graph databases enables financial institutions to explore customer interaction networks, target potential consumers accurately, and develop personalized marketing strategies.

For example, a bank has established a network diagram of customer interaction and consumption through a graph database, and integrated transaction data to screen out customer groups interested in similar investment products. Through in-depth research on customer interrelationships with intelligent algorithms, it was found that some customers share common social or consumption chains. Based on these results, the bank successfully identified potential consumers closely associated with existing customers and implemented targeted marketing activities, such as launching personalized preferential policies or recommendation plans. This targeted marketing strategy increases customer conversion rates and reduces marketing costs. This strategy relies on the analysis of customer relationships and consumption patterns, helping financial institutions quickly locate their target customer groups and significantly enhancing the accuracy and efficiency of marketing.

5. Conclusion

The combination of machine learning and graph databases provides a new auxiliary tool in the field of financial technology, and its applications in financial fraud detection, credit scoring, anti-money laundering monitoring, and marketing demonstrate great development potential. With the efficiency of graph databases in processing complex relational data, as well as the ability of machine learning in pattern recognition and future

trend prediction, the financial industry is able to develop towards intelligence and precision. The combination of machine learning and graph databases provides a new auxiliary tool in the field of financial technology, and its applications in financial fraud detection, credit scoring, anti-money laundering monitoring, and marketing demonstrate great development potential. In the future, with the continuous advancement of big data and artificial intelligence technologies, the integration of machine learning and graph databases will inject more creative vitality into the fintech field and provide solid technical support for industry transformation and upgrading.

References

1. G. Karuna, G. B. Santhi, M. Al-Farouni, G. V. Reddy, A. Shukla, and C. P. Patnaik, "Enhancing financial data traceability using graph neural networks for provenance methods," in *Proc. 2024 Int. Conf. IoT, Commun. Autom. Technol. (ICICAT)*, Nov. 2024, pp. 1334–1339, doi: 10.1109/ICICAT62666.2024.10923449.
2. Q. Sun, X. Wei, and X. Yang, "GraphSAGE with deep reinforcement learning for financial portfolio optimization," *Expert Syst. Appl.*, vol. 238, p. 122027, 2024, doi: 10.1016/j.eswa.2023.122027.
3. P. Azad, C. Akcora, and A. Khan, "Machine learning for blockchain data analysis: Progress and opportunities," *Distrib. Ledger Technol.: Res. Pract.*, 2024, doi: 10.1145/3728474.
4. R. Mitra, A. Dongre, P. Dangare, A. Goswami, and M. K. Tiwari, "Knowledge graph driven credit risk assessment for micro, small and medium-sized enterprises," *Int. J. Prod. Res.*, vol. 62, no. 12, pp. 4273–4289, 2024, doi: 10.1080/00207543.2023.2257807.
5. L. Zhao, Z. Yao, C. Liu, Y. Du, H. Hou, and Z. Hu, "Construction and empirical analysis of data governance system based on knowledge graph," in *Proc. 2nd Int. Conf. Big Data, Comput. Intell., Appl. (BDCIA)*, vol. 13550, Mar. 2025, pp. 179–186, doi: 10.1117/12.3059088.