Theory and Practice of Science and Technology
2025, VOL. 6, NO. 4, 144-148
DOI: 10.47297/taposatWSP2633-456925.20250604

A Literature Review on Spark-Based Optimization of Machine Learning Algorithms for Cyber Fraud Detection

Junyi Zou

Ludwig-Maximilians-Universität(LMU), Munich, Bayern, 80539, Germany

ABSTRACT
With the explosive growth of internet data and the diversification of cyber fraud techniques, traditional machine learning algorithms face challenges such as low efficiency, poor real-time performance, and insufficient dynamic adaptability when processing large-scale fraudulent information. Addressing the economic losses and social harm caused by cyber fraud requires new high-efficiency technical solutions. The Spark framework, leveraging its distributed computing capabilities, in-memory acceleration, and machine learning libraries, offers a novel technical pathway for optimizing fraud detection. This paper systematically reviews recent research progress in Spark-based machine learning algorithms for cyber fraud information processing, focusing on the real-time data processing capabilities of the Spark platform, optimization of distributed algorithms, and innovative models tailored to fraud characteristics. It reveals that existing studies have achieved significant algorithm acceleration and improved classification performance. However, limitations persist in multi-source data adaptability, dynamic fraud pattern recognition, and cross-scenario generalization. Future research should integrate deep learning with real-time stream processing to develop more robust fraud detection frameworks. This paper provides a technical pathway reference for optimizing cyber fraud information processing, offering both theoretical significance and practical value for enhancing the real-time responsiveness and precision of cybersecurity defenses.

KEYWORDS
Spark; distributed computing; machine learning algorithms; fraud detection

1 Introduction

With the rapid development of information technology, the internet has achieved widespread global adoption and become an indispensable part of daily life and work. In the networked environment, information spreads at unprecedented speeds, and data volumes have grown exponentially, posing significant challenges for data processing and analysis. To address these challenges, big data technologies have emerged and yielded remarkable research achievements. Among them, Spark, as a fast, general-purpose, and scalable big data processing engine, has gained extensive application in the field of big data processing due to its efficient in-memory computing capabilities and rich machine learning libraries. Many studies have focused on leveraging Spark's advantages to optimize various data processing tasks and improve efficiency and accuracy.

Against the backdrop of increasingly complex cyber environments, cybersecurity threats faced by individuals and organizations have grown severe, with phishing, scams, and fraudulent activities becoming rampant. Cyber fraud not only causes substantial financial losses to individuals and businesses but also severely undermines network security and stability. To effectively identify and prevent cyber fraud, machine learning algorithms have been widely applied in processing fraudulent information. By analyzing and mining large-scale datasets such as network transaction records and user behavior data, these algorithms can uncover latent fraud patterns and features, enabling accurate fraud detection. However, as data volumes continue to expand, traditional machine learning algorithms face challenges such as inefficiency and limited accuracy when processing large-scale fraudulent information.

This paper focuses on optimizing cyber fraud information processing using Spark-based machine learning algorithms. While existing research has explored the application of machine learning to fraud detection, few studies have fully utilized Spark's advantages to address efficiency issues in large-scale data processing. This paper aims to leverage Spark's distributed computing capabilities and machine learning libraries to optimize existing algorithms, thereby improving both the efficiency and accuracy of fraud information processing. Through this research, we seek to provide a more efficient and precise methodology for cybersecurity applications, contributing to the reduction of fraud-induced losses and the enhancement of network security and stability.

2 Current Research Status

2.1 Spark's Distributed Computing Capabilities and High Processing Efficiency

Spark-based big data analysis and platform construction leverage its distributed computing capabilities and high processing efficiency to collect, process, and analyze massive datasets. This approach facilitates the development of

intelligent big data analysis platforms aimed at extracting data value and addressing real-world challenges. Tao Huang and Liting Gao proposed a real-time data collection and processing method based on Spark. Operating in a distributed environment, Spark compensates for computational limitations by efficiently handling large-scale data. By integrating Flume and Kafka, which aggregate diverse data sources, Spark enables real-time monitoring of data streams from heterogeneous sources. The Spark Flow module further processes these streams in real time, with processed data transmitted to downstream components or databases [1]. Lei Zhao, Ji'an Xia, Yang Wu implemented three machine learning algorithms—feedforward artificial neural networks, support vector machines, and random forests—using Spark's MLlib. They evaluated the operational and classification performance of these algorithms on big data platforms, demonstrating Spark's adaptability to advanced analytical tasks [2]. Shuyan Jia designed a big data intelligent analysis platform by combining Spark with a data analysis orchestrator, referencing the MapReduce (MR) data network quality analysis framework. Experimental results showed that the platform achieved an acceleration ratio exceeding 9, operational efficiency of approximately 99%, and low average absolute error (0.5%–0.8%) in data analysis [3]. Yongjun Bai developed a network anomaly traffic capture system based on Spark's real-time stream data processing technology, which enables efficient data collection, real-time processing, and precise anomaly detection [4]. Murat AYDOĞAN and Ali KARCI developed a classification application using the Naive Bayes method from Apache Spark's MLlib. Their system achieved high accuracy in analyzing spam and non-spam datasets, highlighting Spark's superior data processing performance compared to other platforms. As illustrated in their results, Spark significantly outperformed conventional data analysis platforms in terms of average uptime for processing equivalent data volumes [5].

Fig. 1 Murat AYDOĞAN,Ali KARCI(September 2018)

İskender Ülgen Oğul, Caner Ozcan, and Özlem Hakdağlı et al. implemented a Naive Bayes method on Apache Spark for rapid text classification. By leveraging Spark's distributed in-memory data collection architecture, they analyzed annotated data from the open-source social news platform Reddit and presented the results in tables and charts [6]. A. N. M. JayaLakshmi and K. V. Krishna Kishore conducted sentiment analysis on the Apache Spark distributed framework. They employed Term Frequency-Inverse Document Frequency (TF-IDF) and the unsupervised Spark Latent Dirichlet Allocation (LDA) clustering algorithm for feature extraction and selection. The performance of their framework—evaluated using Deep Neural Networks (DNNs), Support Vector Machines (SVMs), and tree ensemble classifiers—was tested in both single-node and cluster environments, aiming to enhance computational efficiency, particularly in reducing runtime [7].

2.2 Algorithmic Model Innovations for Network Data Detection and Classification

In the field of network-related detection and classification research, optimized machine learning algorithms and novel methodologies have been proposed to achieve precise detection and classification of network intrusions, traffic anomalies, and traffic types. Jiacheng Wang investigated Spark-based network intrusion detection algorithms, exploring approaches to effectively identify network intrusions [8]. Shihao Lü designed and implemented a Spark-powered network traffic anomaly detection system from a system development perspective, revealing methodologies for constructing such systems under current technical constraints [9]. Jiaqi Yao focused on Spark-based network traffic classification techniques to enhance classification accuracy and efficiency [10]. Yang Hu, Xuegang Hu, Peipei Li, proposed a distributed rapid short text data stream classification method using Spark. This approach leverages external corpora to build Word2vec word vector models, addressing the high-dimensional sparsity of short texts. An extended word vector library was constructed to adapt to text variability, and an integrated Logistic Regression (LR) classifier model was employed for classification, while distributed processing improved efficiency in handling massive short text streams [11]. Xianchun Zhou, Heng Xiao, and Pingping Jiao introduced a malware maximal frequent subgraph mining method on Spark, offering new pathways for malware detection and analysis [12].

In addressing the limitations of traditional machine learning algorithms in processing high-dimensional, sparse

vulnerability text features and ignoring context-specific vulnerability information, Yazhou Li developed the TextCGRU vulnerability classification model. This novel framework combines convolutional neural networks (CNNs) for extracting local features of vulnerability text vectors with recurrent neural networks (RNNs) for capturing global contextual features. The fusion of these complementary models enables richer vulnerability feature representation [13]. Deniz Kılınç proposed a Spark-based big data analytics framework for real-time sentiment prediction in streaming data. To enhance the quality of sentiment analysis, the framework incorporates a fake account detection service. Training and testing were conducted using the Naive Bayes model from Apache Spark's MLlib, achieving sentiment classification accuracies of 86.77% (offline) and 80.93% (real-time) [14].

Rafał Kozik integrated NetFlow data with Extreme Learning Machine (ELM) classifiers trained in Apache Spark's distributed environment. Leveraging the MapReduce programming model to scale and distribute ELM training, the method demonstrated reliability in detecting malware activities based on NetFlow data [15]. Raed A. Hasan et al. introduced an adaptive clustering and classification algorithm for Twitter data stream processing on Spark. The two-stage approach employs modified fuzzy C-means clustering with adaptive particle swarm optimization for preprocessing and a modified Support Vector Machine (MSVM) classifier for Higgs data classification. Experimental results showed superior performance across multiple metrics compared to existing methods [16].

2.3 Research on Spam and Cyber Fraud Information Detection

Apache Spark-based technologies have been widely applied in text classification and big data analytics, enabling efficient processing of large-scale datasets and accurate classification and analysis of fraudulent information. Research on machine learning techniques for spam and cyber fraud detection aims to identify and filter deceptive content across platforms (e.g., emails, social networks, websites, review forums) to ensure cybersecurity and information authenticity. Andronicus A. Akinyelu conducted an in-depth analysis of spam detection (e.g., email, web, social media, and review spam) and proposed a survey of machine learning and nature-inspired spam detection techniques. This work provides guidance for designing effective spam-filtering systems, highlighting the suitability of deep learning and big data solutions for spam detection. Future directions include leveraging large datasets and advanced algorithms to develop high-performance detection frameworks [17].

Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada analyzed the impact of misinformation in online reviews, revealing that most studies focus on supervised learning methods while neglecting the potential of big data analytics for review spam detection. Their work offers a comparative synthesis of current research and outlines methodologies for future investigations [18]. Naw Safrin Sattar, Shaikh Arifuzzaman, Minhaz F. Zibran, and Md Mohiuddin Sakib introduced an innovative network graph-based predictive modeling approach for web spam detection. By computing multiple graph metrics and employing five machine learning classifiers, their method achieved high accuracy. The study also compared graph-based features with content-based textual features, demonstrating that graph metrics enable robust detection of web spam in any input network structure [19]. Shodh Sagar and Balachandr et al. proposed an advanced detection framework combining N-gram TF-IDF feature selection with a deep multilayer perceptron neural network. Further optimized via a distribution-based balancing algorithm, their method outperformed existing approaches on benchmark datasets, effectively capturing complex features and reducing false positives [20].

Pinnapureddy Manasa et al. addressed social media tweet spam detection by designing a swarm optimization-based solution. The framework trains machine learning models on tweet datasets, employs the Whale Optimization Algorithm (WOA) with Stochastic Gradient Descent (SGD) for feature selection, and utilizes an AdaBoost (AB) classifier to detect spam tweets, achieving high accuracy and rapid detection [21]. Anikait Kapoor focused on SMS spam detection using the UCI machine learning repository's SMS spam dataset. After preprocessing and feature extraction, various machine learning techniques were evaluated, identifying an optimal filtering algorithm that significantly reduced the error rate of prior state-of-the-art models [22]. Caixu Wu and Jing Zhang investigated Spark-based anomaly detection in big data terminals, analyzing factors influencing terminal anomalies and their interrelationships through data-driven methods [23]. Ning Wang proposed a distributed image segmentation model leveraging the Spark platform, offering novel insights for image-based fraud detection [24].

Muhammad Salman et al. introduced a new dataset of over 68,000 SMS messages to address challenges in SMS spam detection. A longitudinal analysis revealed evolving spam patterns, while semantic and syntactic feature evaluations exposed limitations of shallow machine learning techniques, urging researchers to refine detection frameworks [25]. Konstantinos I. Roumeliotis et al. tackled email spam and phishing detection by fine-tuning advanced language models (e.g., GPT-4, BERT, RoBERTa) for email filtering. This approach overcomes the limitations of traditional convolutional neural networks (CNNs), achieving precise identification of spam and phishing content while minimizing false positives, thus advancing spam-filtering technologies [26].

3 Conclusion

Researchers worldwide have achieved significant progress in Spark-based machine learning algorithms for cyber fraud information processing, offering valuable insights and foundational frameworks for the field. Domestic and international researchers have excelled in network-related detection and classification, big data analytics, and platform development. For instance, extensive studies have focused on network intrusion detection, traffic anomaly identification, and traffic classification, while others have advanced real-time data collection, algorithm performance evaluation, and intelligent big data platform construction. Innovations in Apache Spark applications, such as classification frameworks, large-scale analytics systems, and multi-method spam detection techniques, further demonstrate the versatility of Spark in combating cyber fraud.

However, despite these advancements, challenges persist. Domestic research often lacks specificity in fraud information processing, with a predominant focus on broader areas like network monitoring and big data analysis, rather than integrating machine learning algorithms deeply into fraud detection workflows. International studies, while widely applied in spam detection, frequently overlook the adaptability and stability of methods in complex network environments. Many rely on homogeneous datasets, limiting their ability to reflect real-world fraud dynamics. Additionally, both domestic and international efforts fall short in addressing the real-time responsiveness and dynamic adaptability required to counter rapidly evolving fraud tactics.

Future directions should prioritize cross-disciplinary integration, combining deep learning with Spark's real-time stream processing to enhance fraud pattern recognition. Researchers must focus on multi-source data utilization to develop frameworks that leverage diverse, large-scale datasets, improving generalizability across scenarios. Dynamic model optimization is critical to creating adaptive algorithms capable of adjusting to emerging fraud strategies. Collaborative efforts between academia and industry are essential to bridge gaps between theoretical innovations and practical implementations, ensuring solutions align with real-world cybersecurity demands. This review underscores the transformative potential of Spark-optimized machine learning in cybersecurity while advocating for targeted, scenario-driven solutions to address existing limitations.

References

[1] Tao Huang and Liting Gao. "Real-Time Data Collection and Processing Based on Spark." Journal of Hebei University of Architecture and Civil Engineering, vol. 40, no. 04, 2022, pp. 176–179+188.

[2] Lei Zhao, Ji'an Xia, Yang Wu, et al. "Performance Comparison of Classification Algorithms Based on the Spark Platform." Computer and Digital Engineering, vol. 52, no. 03, 2024, pp. 688–691+704.

[3] Shuyan Jia. "Construction of a Big Data Intelligent Analysis Platform Based on Spark Technology." Journal of Binzhou University, vol. 39, no. 06, 2023, pp. 86–91. DOI:10.13486/j.cnki.1673-2618.2023.06.012.

[4] Yongjun Bai. "Research on Network Anomaly Traffic Data Capture Methods Based on Big Data Technology." Cybersecurity and Informatization, no. 07, 2024, pp. 37–39.

[5] Murat Aydogan and Ali Karci. "Spam Mail Detection Using Naive Bayes Method with Apache Spark." 2018.

[6] İskender Ülgen Oğul, Caner Ozcan, and Özlem Hakdağlı. "Fast Text Classification with Naive Bayes Method on Apache Spark." 2017.

[7] A. N. M. JayaLakshmi and K. V. Krishna Kishore. "Performance Evaluation of DNN with Other Machine Learning Techniques in a Cluster Using Apache Spark and MLlib." 2022.

[8] Jiacheng Wang. "Research on Spark-Based Network Intrusion Detection Algorithms." Master's thesis, Nanchang University, 2023. DOI:10.27232/d.cnki.gnchu.2023.001559.

[9] Shihao Lü. "Design and Implementation of a Spark-Based Network Traffic Anomaly Detection System." Master's thesis, Beijing Jiaotong University, 2023. DOI:10.26944/d.cnki.gbfju.2023.001326.

[10] Jiaqi Yao. "Research on Spark-Based Network Traffic Classification Technology." Master's thesis, North University of China, 2023. DOI:10.27470/d.cnki.ghbgc.2023.000755.

[11] Yang Hu, Xuegang Hu, and Peipei Li, et al. "A Distributed Fast Short Text Data Stream Classification Method Based on Spark." Computer Engineering and Applications, vol. 56, no. 14, 2020, pp. 138–147.

[12] Xianchun Zhou, Heng Xiao, and Pingping Jiao, et al. "Malware Maximal Frequent Subgraph Mining Method Based on Spark Platform." Modern Computer, vol. 29, no. 14, 2023, pp. 57–61.

[13] Yazhou Li. "Research on Software Vulnerability Automatic Classification Method Based on NLP Neural Networks." PhD diss., Yanshan University, 2020. DOI:10.27440/d.cnki.gysdu.2020.001926.

[14] Deniz Kılınç. "A Spark-Based Big Data Analysis Framework for Real-Time Sentiment Prediction on Streaming Data." 2019.

[15] Rafał Kozik. "Distributing Extreme Learning Machines with Apache Spark for NetFlow-Based Malware Activity Detection." 2018.

[16] Raed A. Hasan, Royida A. Ibrahem Alhayali, Nashwan Dheyaa Zaki, and Ahmed Hussien Ali. "An Adaptive Clustering and Classification Algorithm for Twitter Data Streaming in Apache Spark." 2019.

[17] Andronicus A. Akinyelu. "Advances in Spam Detection for Email Spam, Web Spam, Social Network Spam, and Review Spam: ML-Based and Nature-Inspired-Based Techniques." 2021.

[18] Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. "Survey of Review Spam Detection Using Machine Learning Techniques." 2015.

[19] Naw Safrin Sattar, Shaikh Arifuzzaman, Minhaz F. Zibran, and Md Mohiuddin Sakib. "Detecting Web Spam in Webgraphs with Predictive Model Analysis." 2019.

[20] Shodh Sagar and Balachandar Paulraj. "Leveraging Machine Learning for Improved Spam Detection in Online Networks." Universal Research Reports, 2024.

[21] Pinnapureddy Manasa, Arun Malik, Khaled N. Alqahtani, et al. "Tweet Spam Detection Using Machine Learning and Swarm Optimization Techniques." IEEE Transactions on Computational Social Systems, vol. 11, 2024, pp. 4870–4877.

[22] Anikait Kapoor. "SMS Spam Detection Using Machine Learning Approach." International Journal of Research in Science and Technology, 2024.

[23] Caixu Wu and Jing Zhang. "Simulation of Potential Anomaly Identification in Big Data Terminals Based on Spark Computing." Software, vol. 45, no. 06, 2024, pp. 31–33.

[24] Ning Wang. "Research on Distributed Image Segmentation Methods Using the Spark Platform." PhD diss., University of Chinese Academy of Sciences (Aerospace Information Research Institute), 2022. DOI:10.44231/d.cnki.gktxc.2022.000072.

[25] Muhammad Salman, Muhammad Ikram, and M. Kâafar. "Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models." IEEE Access, vol. 12, 2024, pp. 24306–24324.

[26] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. "Next-Generation Spam Filtering: Comparative Fine-Tuning of LLMs, NLPs, and CNN Models for Email Spam Classification." Electronics, 2024.