Anomaly Detection of Financial Data using Machine Learning

Khirod Chandra Panda

Asurion Insurance, USA
Email: *khirodpanda4bank[at]gmail.com*
0009 - 0008 - 4992 - 3873

Abstract: ***Anomaly detection is critical in the financial sector, especially as financial environments evolve with increasing digitization, posing challenges for real - time anomaly detection. Recently, deep learning (DL) algorithms have emerged as promising solutions for this problem. This study presents a DL - based anomaly detection model utilizing various algorithms, including LSTM, GRU, and 1dCNN, applied to Tesla's stock market and Ethereum cryptocurrency data sets. Hyperparameter optimization is performed using grid search. Results show that the GRU algorithm achieves the highest prediction score in both datasets, while the 1dCNN algorithm performs the lowest. Additionally, anomaly values are graphically demonstrated using GRU for both datasets. Accurate bookkeeping is essential for legitimate business operations, yet the complexity of financial auditing requires new solutions. Supervised and unsupervised machine learning techniques are increasingly applied to detect fraud and anomalies in accounting data. This paper addresses the challenge of detecting financial misstatements in general ledger (GL) data, proposing seven supervised ML techniques, including deep learning, and two unsupervised ML techniques. Models are trained and evaluated on real - life GL datasets, demonstrating high potential in detecting predefined anomaly types and efficiently sampling data. Practical implications of these solutions in accounting and auditing contexts are discussed. The rapid development of computer networks brings both convenience and security challenges due to various abnormal flows. Traditional detection systems, like intrusion detection systems (IDS), have limitations, necessitating real - time updates to function effectively. With the advent of machine learning and data mining, new methods for abnormal network flow detection have emerged. This paper introduces the random forest algorithm for detecting abnormal samples, proposing the concept of an abnormal point scale to measure sample abnormality based on similarity. Simulation experiments demonstrate the superiority of random forest - based detection in terms of model accuracy and computing efficiency compared to other methods.***

Keywords: accounting; auditing; anomaly detection; general ledger, machine learning, Anomaly detection, LSTM, GRU, 1dCNN

1. Introduction

The application of machine learning (ML) techniques in financial auditing context is in high demand [1]. The intricate nature of manually managing audit - related tasks is significantly challenging, emphasizing the need for improved automation and intelligent solutions [2]. Financial markets enable traders to profit from trading financial instruments without needing to physically possess the underlying assets. While traders can hold assets for extended periods, as seen in options markets, the prices of assets in these markets typically fluctuate due to short selling, which attracts speculators focused solely on price movements. An ideal market is both liquid and efficient, ensuring that there are no abrupt, volatile changes in the prices and volumes of traded instruments, thus maintaining a relatively stable market [3]. In practice, undisclosed information related to a financial instrument and various trade - based manipulations can influence both the price and volume of the instrument in a financial market [4]. Market abuses are subjective and are generally decided by regulations and the guidelines governing the market. Nevertheless, although there is no standard definition of market abuse, there are two main categories of abuses that have been widely studied in the literature [5].

Today's financial market trading activities are predominantly conducted on electronic platforms, facilitated by the accessibility of information and the convenience of automation. This shift has heightened demands for transparency, compelling marketplaces to establish stringent regulatory measures for participants. Regulatory oversight now heavily relies on trading data to monitor market behavior and detect malfeasance such as price manipulation and insider trading. However, the vast quantity of data presents significant challenges in storage, processing, and analysis.

To manage the deluge of information, regulators have turned to rule - based systems that flag potential market anomalies based on predefined criteria. Such systems generate alerts for suspicious events, which are then scrutinized by human experts. Despite their utility, these systems are fraught with issues, notably the high volume of alerts they produce, which complicates the identification of genuine instances of market manipulation. As markets evolve and grow in complexity—evidenced by the diversity of trading activities in sectors like the physical power market—these rule - based systems struggle to adapt, potentially overlooking novel manipulation tactics. Consequently, while rule - based [6] surveillance is a cornerstone of current regulatory strategies, its limitations underscore the need for more dynamic and adaptable approaches to market oversight.

2. Key Challenges

At a fundamental level, the objective of financial market surveillance is to delineate a domain of standard behavior and flag any deviations within the data as potential manipulations. The process of algorithmically distinguishing between normal and abnormal market behaviors automatically is complex and fraught with practical challenges.

Firstly, anomalies are inherently infrequent, making the available labeled data for training machine learning models to identify such events scarce and expensive to produce.

DOI: https://dx.doi.org/10.21275/SR24403054826

Secondly, to differentiate between normal and anomalous events, it is necessary to define a boundary that accurately represents all conceivable normal activities. However, the definition of this boundary is often imprecise, leading to potential misclassification of data points that lie near it. In the context of anomaly detection algorithms, models are initially trained to assign scores to each data point, with those scoring highest flagged as anomalies. These are then scrutinized by human analysts to discern genuine anomalies, although many identified by the algorithms could be false positives due to their deviation from the modeled norm. Thirdly, the inherent variability in data generation, collection, and processing introduces noise, complicating the detection of true anomalies and increasing the likelihood of false positives. Fourthly, abnormal behavior in the market may not be isolated to a single action but could encompass a series of actions by an entity, where the sequence and timing are crucial. Therefore, it is essential for the detection system to account for the sequential nature and timing of market actions to accurately identify anomalies [7].

The first three challenges are common to anomaly detection across various domains. This paper delves into the evolution of anomaly detection techniques aimed at addressing these issues within the realm of market surveillance and other areas. The fourth challenge is unique to the development of systems based on machine learning that interpret patterns in time - series data for anomaly detection and prediction. We explore the fundamental aspects of this challenge, review research in the field, and examine technical hurdles in designing machine learning - based surveillance systems. Specifically, we conduct a comparative analysis of different machine learning methods for pattern recognition in time - series data, predictions, and anomaly detection, with a focus on their application to forecasting and spotting abnormal price movements in electricity trading markets. We share insights from this analysis on pattern learning, effective prediction of future prices, and the identification of anomalous price fluctuations in time - series data.

In machine Learning, The random forest algorithm is an integration and improvement based on the decision tree algorithm and is the result of the integrated learning of the decision tree algorithm. This paper uses two machine learning algorithms, decision tree and random forest, to construct a financial statement analysis - based system for judging financial irregularities of listed companies, which can predict and analyze financial irregularities, thereby helping to discover more potential unknown financial violations and risk of violations and promote the steady development of the financial system.

2.1 Decision Tree

In assessing whether a company has reported violations, the significance of different types of data varies. Regulatory bodies, such as the China Securities Regulatory Commission, prioritize certain data elements when evaluating potential breaches. This evaluative approach shares similarities with the logic underlying decision trees. Consequently, we have opted to employ the decision tree machine learning model for scrutinizing corporate financial data. The decision tree algorithm stands out in the realm of machine learning for its capability to address classification challenges within supervised learning contexts.

The decision tree algorithm functions by extracting patterns from training data, which are then applicable to new, unseen data. Beyond decision trees, the machine learning spectrum encompasses algorithms such as Naive Bayes, support vector machine - based classifiers, neural networks, K - means clustering, and fuzzy classification techniques.

In a decision tree model, every branch delineates a relationship between an entity's attribute and its respective value or category. Non - leaf nodes represent decision points linked to attributes, while branch paths align with attribute values meeting those decisions. Leaf nodes signify sets of values consistent with the conditions traced from the root to the leaf. The model's construction begins at the root, selecting attributes to segment the sample set into subsets, each representing a branch node, which are further divided until homogeneity, or a specific criterion is achieved. The decision tree model embodies a tree - like algorithmic structure, mirroring the structured thought processes humans engage in, particularly when analyzing a company's financials, such as profitability in each period.

Constructing a decision tree typically involves two phases: (1) generating the tree from training samples, and (2) pruning the tree to ensure accuracy and relevance by eliminating overfitting. However, our proposed method integrates multiple decision trees, each functioning as a "weak" classifier to avoid overfitting, thereby bypassing the pruning stage. The input for constructing the decision tree algorithm is described as follows.

$$I = \{(A_{00}\dots, A_{0j}\dots, A_{0n}, T_0)\dots(A_{i0}\dots, A_{ij}\dots, A_{in}, T_i)\dots\}, \quad (1)$$

where A_{ij} represents the value of the j - th attribute of the i - th sample in the set and T_i is the type mark of the i - th sample. The result of decision tree construction is a binary tree or multibranch tree. The binary tree is generally used for data collection whose attributes are all Boolean logic judgments.

Different decision tree classification algorithms use different judgment conditions to select split attributes. The two most important judgment conditions are information gain and information gain rate. Split attribute selection based on information gain. Suppose the training sample set is S, and the attribute set is

$$P = \{p_1, \dots p_i, \dots p_m\}. \quad (2)$$

Then, the proportion of samples belonging to the j - th category in the sample dataset is

$$P(C_j) = \frac{|S_{ij}|}{|S|}. \quad (3)$$

Currently, the information entropy of the sample dataset S is

$$\text{Entropy}(S, p_i) = \sum_{j=1}^{n} -P(C_j)\log_2 P(C_j). \quad (4)$$

Suppose that in the sample dataset, the value range corresponding to the attribute p_i vi is, and S_i (v) represents the subset of samples whose attribute pi takes the value v. Then, the information gain of the sample set S to the attribute p_i is

$$\text{Gain}\,(S, p_i) = \text{Entropy}\,(S, p_i) - \sum_{v \in v_i} \frac{|S_i\,(v)|}{|S|} \text{Entropy}\,(S, p_i)\,. \tag{5}$$

After the information gain of the sample set S is calculated, the split information of S on the attribute pi is calculated as

$$\text{SplitGain}\,(S, p_i) = -\sum_{v \in v_i} \frac{|S_i\,(v)|}{|S|} \log_2 \frac{|S_i\,(v)|}{|S|}. \tag{6}$$

Then, the information gain rate of the sample dataset S relative to the attribute p_i is

$$\text{GainRatio}\left(S, p_j\right) = \frac{\text{Gain}\left(S, p_j\right)}{\text{SplitInfo}\left(S, p_j\right)}. \tag{7}$$

2.2 Random Forest

A concise approach to evaluating company performance and random forest model construction is as follows:

Company Evaluation: Companies that have demonstrated profitability and steady operations over several years are preliminarily classified into a database of excellent companies. Conversely, companies with poor management histories and past losses warrant increased scrutiny and a deeper analysis to enhance our understanding of their operations.

Random Forest Model Construction: Central to developing a random forest model is the creation of multiple, independent, and diverse decision trees. Using raw data alone is insufficient for achieving a diversified decision tree model, as it does not fully leverage the benefits of ensemble learning. Therefore, data sampling based on specific rules is essential. Each decision tree within the random forest is trained on a subset of the total sample, minimizing data repetition and ensuring the uniqueness of the training data for each tree, thereby enhancing the diversity of the decision trees.

Random Forest Characteristics: The random forest is an ensemble classifier composed of decision trees. It features a hierarchical structure with root nodes (representing the entire training dataset), internal nodes (splitting problems based on attributes), and leaf nodes (data collections with classification labels). The decision tree algorithm operates on a top - down, greedy approach, selecting the best attribute for splitting data at each node to refine classification. The choice of the splitting attribute is critical, with selection criteria including information gain, information gain ratio, and Gini index. Various decision tree algorithms, such as ID3, C4.5, and CART, apply different methods for attribute selection.

3. Simulation

When dealing with tasks that require extensive numerical calculations, it's crucial for the algorithm to be straightforward and quick. Yet, distance - based detection approaches often suffer from prolonged computation times and substantial memory demands. The UC (University of California Irvine) machine learning database stands as a renowned repository for testing machine learning algorithms, frequently employed in algorithm modeling and validation. This study conducts simulation experiments using six standard datasets, wherein abnormal samples constitute 5% of the total dataset. It compares the performance of an abnormal sample detection method leveraging random forest technology, as introduced in this study, against two distance - based methods (RHM and robust Mahalanobis distance). Furthermore, it evaluates the effectiveness of these methods by examining the model's predictions after excluding the abnormal samples.

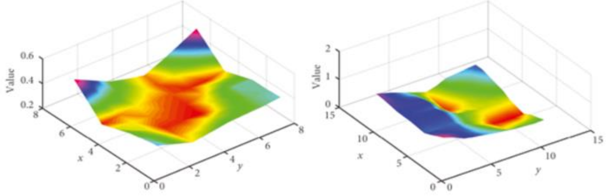

Figure 1: Evaluated Value

Additionally**,** the robustness of the three approaches is assessed using a model based on support vector machine **(**SVM**)** technology**,** with the results presented in Figure 1**.**

Initially, 5% of the "abnormal samples" are removed from each dataset using three distinct methods, after which these modified datasets serve as the foundation for constructing a random forest model. Depending on the collective sample size across the six datasets, a random forest comprising 500 to 1000 trees is established. For each node within the forest, the number of potential splitting attributes (q) is determined by the square root of the total attribute count in the dataset, and 5 - fold cross - validation is applied.

To further assess the robustness of the three detection methods, an additional experiment is performed: each dataset undergoes five - fold cross - validation to generate separate training and testing sets ([training set i], [test set i] for i = 1, 2,. . ., 5). The training sets are used to construct SVM models, which are then evaluated using the respective test sets. The mean accuracy from all five tests provides the 5 - fold cross - validation accuracy. Subsequently, the three methods are applied to remove abnormal samples from each [training set i], after which the refined [training set i] is used to build new SVM models. These models are tested against the [test set i] containing previously unremoved abnormal samples, and the accuracy of the 5 - fold cross - validation is calculated. SVM models are developed using the libSVM toolkit, employing a Gaussian kernel function and optimizing the penalty coefficient C through grid search.

The removal of abnormal samples by the three methods results in varied improvements in model accuracy, confirming the effectiveness of each abnormal sample detection approach. However, the comparison reveals that the random forest (RF) - based method for detecting abnormal samples outperforms the other two in enhancing model accuracy and demonstrates superior robustness.

The modeling and testing strategy involves removing abnormal samples from the training set while retaining them in the test set, thereby placing a premium on the model's capacity for generalization. The RHM method and the robust Mahalanobis distance method require the computation of the covariance matrix's inverse. However, the need for a pseudo - inverse arises when the covariance matrix is singular, diminishing the algorithm's robustness. As previously indicated, the robustness of the Random Forest (RF) method surpasses that of both the RHM and the robust Mahalanobis distance methods. RF ensures the model's generalization capability across all six datasets, a feat not matched by the other two methods, which exhibit reduced robustness and accuracy, notably on the heart dataset. Moreover, for large datasets, the memory usage and time consumption become significant impediments to the detection of abnormal samples, issues the RF algorithm does not encounter. This superiority facilitates the broader adoption of random forest - based methods for detecting abnormal samples.

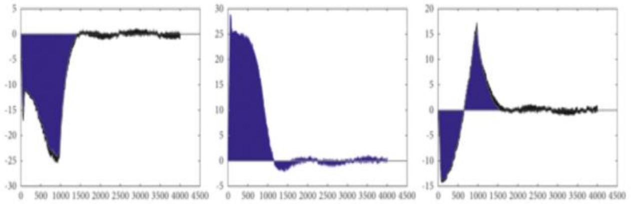

Figure 2: Showcases a comparison of the predictive accuracies

4. Conclusion

This study introduces the random forest algorithm for the identification of abnormal samples, integrating it with sample similarity to devise a novel metric called the abnormal point scale. This metric aims to quantify the abnormality level of samples, facilitating the filtration of outliers based on their scale values.

Beyond its application in outlier detection, the concept of sample similarity offers additional utility in forming dataset prototypes, mapping dataset dimensions, and filling in missing data within both training and test sets. The use of sample similarity via random forest uncovers a broader and more profound potential for exploring dataset characteristics. Nevertheless, the selection of scale thresholds for identifying abnormal samples currently relies on empirical outcomes, lacking a solid quantitative basis. This gap highlights an area ripe for future investigation.

References

[1] Baesens, B.; Van Vlasselaer, V.; Verbeke, W. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection; Wiley: New York, NY, USA, 2015. [Google Scholar]

[2] Zemankova, A. Artificial Intelligence in Audit and Accounting: Development, Current Trends, Opportunities and Threats - Literature Review. In Proceedings of the 2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), Athens, Greece, 8–10 December 2019; pp.148–154. [Google Scholar]

[3] F. Black, "Toward a fully automated stock exchange Part I", *Financial Analysts J.*, vol.27, no.4, pp.28 - 35, Jul.1971.

[4] F. Allen and D. Gale, "Stock - price manipulation", *Rev. Financial Stud.*, vol.5, no.3, pp.503 - 529, Jul.1992

[5] Y. Cao, Y. Li, S. Coleman, A. Belatreche and T. M. McGinnity, "Detecting price manipulation in the financial market", *Proc. IEEE Conf. Comput. Intell. Financial Eng. Econ. (CIFEr)*, pp.77 - 84, Mar.2014.

[6] D. Diaz, B. Theodoulidis and P. Sampaio, "Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices", *Expert Syst. Appl.*, vol.38, no.10, pp.12757 - 12771, Sep.2011.

[7] M. Atallah, R. Gwadera and W. Szpankowski, "Detection of significant sets of episodes in event sequences", *Proc.4th IEEE Int. Conf. Data Mining (ICDM)*, pp.3 - 10, Nov.2004.