Anomaly detection in Financial Data

Yetong Li

School of Management, Wuhan University of Technology, Wuhan, 430070, China

309076@whut.edu.cn

Abstract. The business world is a colorful and diverse world in which the integration and communication of many fields will be involved. In order to explore this world, it is necessary to master the basic language of this world - financial statements; however, the prerequisite for the stable operation of the business world is that the financial data of each company are in legal, fair and objective. These fraudulent practices can have multiple effects, which may affect the rights and interests of stakeholders such as consumers, investors and even the entire business world, which requires the presence of auditors, but they need to invest a lot of manpower to deal with the complex and laborious work. find some detection algorithms that can help auditors to increase the efficiency of their work while increasing the detection of financial data sets.

Keywords: accounting, auditing, financial data, financial statements, anomaly detection

1 Introduction

As times progress and financial markets continue to move at high speed, many things are happening at the same time and different people or roles are responsible for monitoring these activities. For example, in the systems back office, the IT team carefully monitors the operation and performance of networks, servers, communication links, etc.; at the application level, an entirely different monitoring team monitors other category factors such as the speed of market data transmission, the time taken to complete a transaction, or the user experience; at the business level, the relevant business people analyze products according to the characteristics of customer transactions. The business operations of companies and other commercial entities that keep books of accounts in the order in which economic transactions occur and are completed are called journals, which are a fundamental part of the accounting system, and the detection of anomalies in financial data is a daily task for auditors who review financial statements. Within the general ledger data, transactions are recorded in journal entries. The original supporting documents for journal entries include monetary amounts and debit and credit symbols. A journal entry by itself does not carry any information about the transaction, and the permutation of different account types within a journal entry creates patterns of economic activity. Possible anomalies in this data include, but are not limited to: incorrect

F. Balli et al. (eds.), *Proceedings of the 2023 2nd International Conference on Economics, Smart Finance and Contemporary Trade (ESFCT 2023)*, Advances in Economics, Business and Management Research 261,
https://doi.org/10.2991/978-94-6463-268-2_46

combinations of account types, incorrect symbols for monetary amounts, and anomalous amounts relative to the standard account pattern. Such deviations from the standard in general ledger data may be erroneous or deliberately fraudulent. Experienced auditors assess the financial statements based on their expertise using the manual sampling method of general ledger data, which is a labor-intensive exercise. In such cases, the assessment of the accuracy of the data is based on the known scope of the relevant accounting standards and regulations. In addition, the sampling of data means that the auditor's work is limited and not every item of data in every financial statement is audited. The auditor's evaluation work will lead directly to the ability to identify material errors and risks in the company and thus provide a complete picture of the company's financial performance. In this case, due to the limited number of known account patterns, errors and fraudulent behavior, while constantly improving over time. In the absence of advanced automated tools, I wanted to find detection algorithms that would help auditors focus on the riskier areas and significantly improve their productivity. As more and more data detection solutions are adopted by audit firms, there is a growing expectation that they will reduce audit costs while improving audit quality [1]. The problem of anomaly detection has been studied extensively in the field of statistics since the 1980s, and as its areas of application have expanded and methods and techniques from other fields have merged, researchers have proposed many different detection methods. Hypothesis testing was one of the first statistically based methods used to identify anomalies in data samples based on the discrimination of low probability events, with the main drawback being the prior assumption that the data set conforms to a particular distribution model. In recent years, some progress has been made in data mining-based anomaly detection research, which is considered from a global perspective to detect isolated points by calculating the distance between data points or objects, which is not effective when the data set contains multiple distributions or when the data set is mixed by different density subsets. The density-based anomaly detection algorithm LOF (Local Outlier Factor), this algorithm overcomes the detection error caused by the mixing of different density subsets, and the detection accuracy is relatively high, but the complexity is too high when dealing with large data sets, and the satisfactory response speed cannot be achieved [2].

2 Anomaly detection

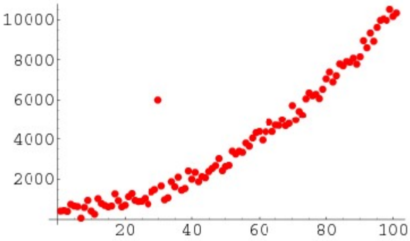


Fig. 1. The point shown by the (30,6000) coordinates in the figure is the outlier

2.1 Rationale

Given a training set X1, X2, X3.... .Xn. (First a set of databases has to be trained)

In the study, a function must be found to detect whether the input x belongs to the training set (whether it belongs to the same class as the data in the training set), and then a threshold is given. Below this threshold, the confidence level is considered "normal" if it is greater than this threshold, and "abnormal" if it is less than this threshold. (To compare: normal if included in the training set; outlier if new, outlier, exception, etc.) The point with coordinates (30,6000) in the figure 1 is the outlier point. The figure 2 shows the basic principles of anomaly detection.

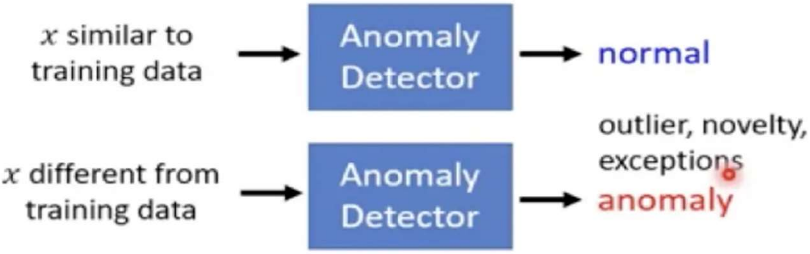

x similar to training data
Anomaly Detector
normal
outlier, novelty, exceptions
x different from training data
Anomaly Detector
anomaly

Fig. 2. Basic principles of anomaly detection

2.2 Financial data anomaly detection method steps

1. **Obtain financial data from the trained database for a pre-determined period of time and divide the financial data into sub-datasets of multiple classes on the basis of the pre-determined data category fields.**
2. **After a data elimination process based on the business subject fields in the sub-data-set, divide each class of sub-data-set into positive and negative datasets.**
3. **Generating feature width tables for positive and negative datasets respectively by feature construction, and performing group baseline partitioning on the positive data-set feature width table to obtain a positive data-set group baseline-transformed feature width table.**
4. **The group baseline-transformed feature width table of the positive data set is fed into the isolated forest algorithm model to obtain the first anomaly data.**
5. **Comparing the subject fields of the first anomaly data with the subject fields in the negative data set feature-width table to identify the risk data in the first anomaly data.**

The figure 3 shows an example of standard normal distribution in anomaly detection.

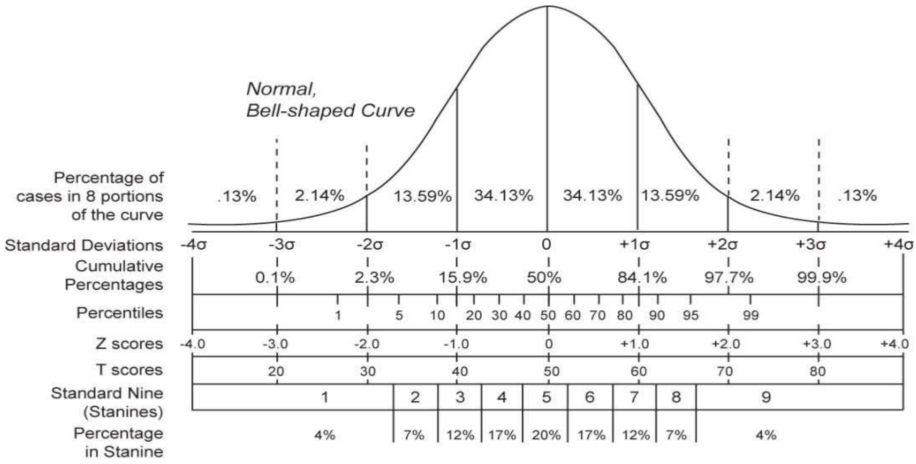
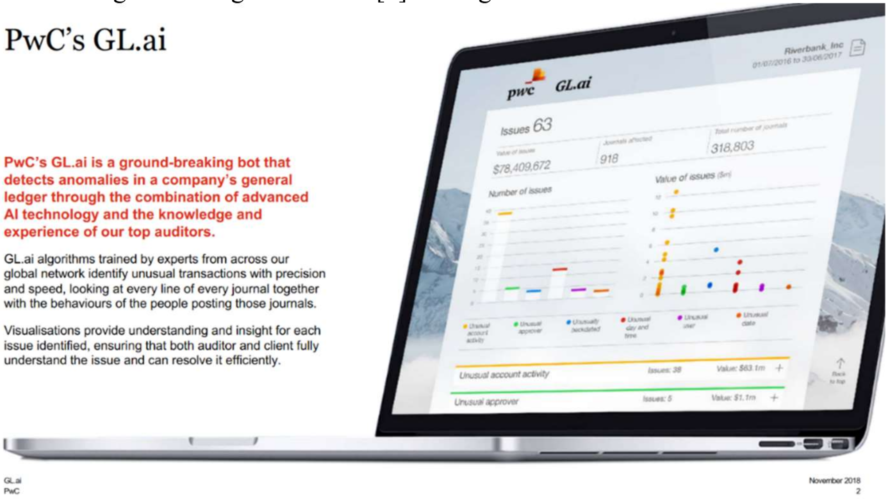


Fig. 3. Standard normal distribution in anomaly detection

2.3 Specific applications

There are many large accounting firms that have developed their own ML-based exception detectors, such as the Big 4 firms Ernst & Young and Pricewaterhouse Coopers. EY has developed the EY Helix General Ledger Exception Detector, an exception identification tool to improve auditor efficiency. [3]

PwC, one of the largest providers of accounting services, has developed a programmer called GL.ai. This uses machine learning and statistical techniques to identify anomalies in general ledger data sets. [4] The figure 4 illustrates PwC's GL.ai.


PwC's GL.ai

PwC's GL.ai is a ground-breaking bot that detects anomalies in a company's general ledger through the combination of advanced AI technology and the knowledge and experience of our top auditors.

GL.ai algorithms trained by experts from across our global network identify unusual transactions with precision and speed, looking at every line of every journal together with the behaviours of the people posting those journals.

Visualisations provide understanding and insight for each issue identified, ensuring that both auditor and client fully understand the issue and can resolve it efficiently.

GL.ai
PwC

November 2018
2

Fig. 4. PwC's GL.ai

2.4 Problems that may be encountered

Risk sampling.

Manual financial statement analysis requires a random sample of data. However, because the sample size is too small relative to the total, there is a significant likelihood that the misstatements detected will occur in a larger, non-sampled group.

Data volume and time efficiency.

Manual audit work becomes more costly as data volumes increase.

Complexity of patterns.

Pattern complexity within financial data is high.

Hidden fraud.

The development of fraudulent financial misstatements is carefully organized with the aim of being dissipated in the course of normal day-to-day financial activities.

Deviation report.

A number of data science solutions can be used to report accounting data discrepancies.

3 Examples of some general calculation methods

3.1 Isolated Forest Algorithm

Isolated Forest is an unsupervised anomaly detection learning algorithm that replaces the most traditional method of normal point profiling with the concept of isolated outliers. An outlier is an observation or statistical event that is so abnormal that it is suspected of having a cause different from that which occurs elsewhere. Outliers in large data sets can follow very complex patterns. In the vast majority of cases, they are difficult to detect "by eye". Therefore, machine learning techniques are well suited to the field of anomaly detection. The most popular anomaly detection techniques focus on creating a 'normal' profile: Anomalies in a data set are reported as instances that do not fit the normal profile.

The figure 5 shows an example of Isolated Forest Algorithm.

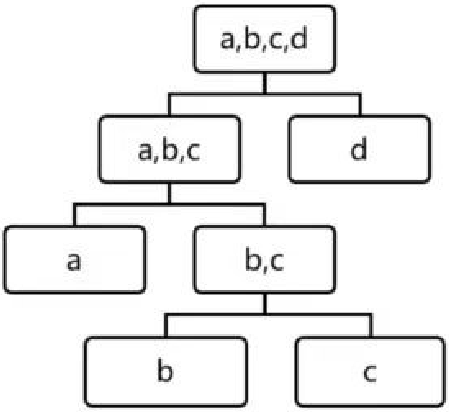


Fig. 5. Isolated Forest algorithm (The d in the graph is separated out first and the data is most likely to be anomalous)

3.2 Local Outlier Factor Method

The Local Outlier Factor (LOF) algorithm is an unsupervised outlier detection method that calculates the local density deviation of a given data point compared to its neighbors. Typically, the number of neighbors under consideration (the neighbors parameter) is greater than the minimum number of samples that should be covered by a cluster, so that other samples could be local outliers relative to that cluster, and less than the maximum number of sample closures that could theoretically be local outliers. Such knowledge is usually not available in practice. Neighbors = 20 seems to work well in general. [5].

The figure 6 shows an example of Local Outlier Factor Method.

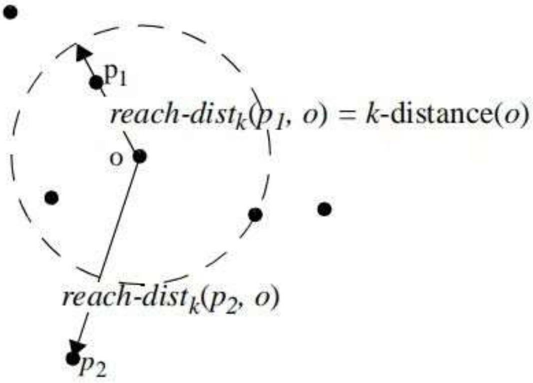


Fig. 6. Local Outlier Factor Method

3.3 Machine learning

In the presence of labels, tree models (gbdt, xgboost, etc.) can be used for classification. The disadvantage is that in anomaly detection scenarios the data labels are not uniform, but the advantage of using machine learning algorithms is that different features can be constructed.

4 Abnormal pattern identification of financial data

4.1 Abnormal financial data pattern identification methods

To determine whether a financial account's transaction behavior is normal, abnormal financial account data pattern identification methods can be used from the following two aspects.

Vertical comparison mode: compare the new data with its own historical behavior pattern to determine whether it is abnormal or not, e.g. if the account deposits or withdraws an amount that does not match its historical pattern at a particular time, it can be considered abnormal.

Side-by-side comparison model: comparison of one's own behavior patterns with the behavior patterns of comparable accounts. Comparing one's own history is prone to misinterpretation due to the existence of fluctuating cycles in consumption and business behavior, such as year-end bonus payments and low and high business seasons, and this can be effectively improved by cross-referencing. By collecting information from a sample of typical, normal accounts with similar business characteristics to those to be identified, establishing a paradigm for their transaction behavior and then comparing the behavioral characteristics of the accounts to be identified with them, anomalies can be detected or false positives can be eliminated. This significantly improves the quality of the screening. For example, it is perfectly normal transactional behavior to see an increase in funds deposited into an account that can be interpreted as a year-end bonus. Therefore, if the peer group all have a significant increase in deposits at the same time, false positives can be ruled out. [6]

4.2 Analysis of financial data transaction behavior

Financial accounts contain a lot of information about the user, according to the knowledge of the financial field, for the original transaction records, the account daily transactions of information generally include: account number debit amount, credit amount, balance transaction time counter-party account counter-party account name, summary, usage, transaction code and other attributes. However, some of this information is not relevant to the identification of anomalies. Therefore, attributes such as the name of the counter-party account are removed. By analyzing the financial account records and summarizing the transaction patterns of the accounts, we can identify those transactions that behave in a significantly different way and are used as an indication to analyze suspicious transactions.

5 Conclusion

There is no doubt that anomaly detection is an innovative tool that can help auditors quickly identify unusual financial data. It is like finding a needle in a haystack, where there may be 10 records of concern in a database of 100 million records. It is based on knowledge of the client, including its business, accounting policies and governance. It enables experienced auditors to make quick judgments. The use of machine learning and statistical techniques to identify anomalies in general ledger data sets has greatly improved the efficiency of auditors. As time goes on, anomaly detection is sure to gain significant traction in the accounting field, so let's wait and see.

References

1. Bakumenko Alexander & Elragal Ahmed. (2022). Detecting Anomalies in Financial Data Using Machine Learning Algorithms. Systems(5). doi:10.3390/SYSTEMS10050130.
2. Zhou Dazhang, Liu Yuefen & Ma Wenxiu. 2011 A new species of the genus Phyllostachys (Coleoptera, Staphylinidae) from China. (2008). Time series anomaly detection. Computer Engineering and Applications (35), 145-147.
3. EY. How an AI Application Can Help Auditors Detect Fraud. Available online: https://www.ey.com/en_gl/better-begins-with-you/how-an-ai-application-can-help-auditors-detect-fraud
4. PwC. GL.ai, PwC’s Anomaly Detection for the General Ledger. Available online: https://www.pwc.com/m1/en/events/socpa-2020/documents/gl-ai-brochure.pdf
5. Abhisu Jain, Mayank Arora, Anoushka Mehra & Aviva Munshi. (2021). Anomaly Detection Algorithms in Financial Data. International Journal of Engineering and Advanced Technology (IJEAT) (5).
6. Deng, Senlin & Chen, Weidong. (2015). Anomaly pattern recognition of financial data based on a class of support vector machines. Journal of the University of Information Engineering (02), 251-256.