Woody International Publish Limited
An Multidisciplinary Academic Journal Publisher
Journal of Theory and Practice in Sciences, Volume 2, 2025
https://www.woodyinternational.com/

Risk Identification and Evaluation Analysis of Engineering Audit Based on Data Mining

Liang Zhang

State Grid Liaoning Electric Power Co., Ltd. Comprehensive Service Center, Shenyang 110000, Liaoning, China

Abstract: ***With the rapid development of information technology, data mining technology has been widely applied in various fields. In the field of engineering auditing, data mining techniques can help auditors identify and evaluate risks more effectively, thereby improving audit efficiency and quality. This article first introduces the basic concepts and methods of data mining technology, and then explores in detail the application of data mining in engineering audit risk identification and evaluation, including key steps such as data preprocessing, risk identification model construction, and risk assessment. Through practical case analysis, the feasibility and effectiveness of data mining technology in engineering auditing have been verified. This study not only provides new methods and ideas for engineering auditing, but also provides reference for the application of data mining technology in other fields.***

Keywords: Data mining; Engineering audit; Risk identification; Risk evaluation.

1. INTRODUCTION

Engineering audit is an important means to ensure the compliance, efficiency, and safety of engineering projects. However, with the expansion and increasing complexity of engineering projects, traditional audit methods are no longer able to meet the demand for comprehensive and accurate identification of risks. Data mining technology, as an emerging information processing technique, can extract useful information from massive amounts of data and provide new solutions for engineering auditing. This article aims to explore the application of data mining techniques in the identification and evaluation of engineering audit risks, in order to provide useful references for the practice of engineering audit. Wu [1] developed cloud infrastructure for large-scale parallel computing in genetic disease analysis, while Tian et al. [10] improved brain tumor segmentation through a GSConv-enhanced UNet architecture with ECA attention mechanisms. Clinical applications have expanded with Shen et al. [11]'s LSTM-based AI system for anesthetic dose management in cancer surgery. The computer vision field has seen significant breakthroughs, particularly in 3D understanding. Peng et al. [3] proposed a dual-augmentor framework for domain generalization in 3D human pose estimation, complemented by their later work on 3D vision-language Gaussian splatting for enhanced scene representation [4]. Urban applications include Zhou et al. [6]'s ResNet-50 based garbage recognition system for sustainable development and Liu et al. [9]'s MiM-UNet for efficient building image segmentation. Financial and cybersecurity applications have leveraged modern AI techniques, with Deng et al. [5] employing transformer models for real-time fraud detection in cloud-optimized systems and Xu et al. [12] analyzing adversarial machine learning threats in cybersecurity. AI's role in financial management has grown through Shen et al. [13]'s data-driven robo-advisors and Chew et al. [14]'s e-commerce financial risk assessment models. Infrastructure and optimization research includes Chen [2]'s quantized framework for gig economy data integration and Xie et al. [7]'s GPU-accelerated Top-K selection for neural networks. Energy sector innovations feature Zhao et al. [8]'s CNN-Bi-GRU model for renewable electricity demand forecasting, demonstrating AI's cross-domain applicability.

2. OVERVIEW OF DATA MINING TECHNIQUES

Data mining refers to the process of processing and analyzing large amounts of data through specific algorithms to discover hidden and valuable patterns and knowledge. Data mining techniques include various methods such as association analysis, classification, clustering, prediction, etc. These methods can be selected and used according to different application scenarios and needs.

2.1 Basic Concepts of Data Mining Technology

Data mining technology is a process of extracting useful information and knowledge from large amounts of data. The term 'large amount of data' here refers to datasets that are massive in quantity, diverse in type, and complex in structure. These datasets may come from different data sources, including databases, files, networks, etc. The core

of data mining technology lies in discovering patterns and patterns in data, which can be association rules, classification models, clustering structures, etc. They can reveal the inherent connections and potential value between data. Specifically, association rules refer to the correlation between data items, that is, whether the appearance of one data item has some kind of association with the appearance of other data items; The classification model divides the samples in the dataset into different categories to achieve prediction and classification of new samples; The clustering structure divides the samples in the dataset into multiple groups or clusters, so that the similarity of samples within the same group is high, while the similarity of samples between different groups is low. The discovery of these patterns and patterns is of great significance for understanding data, predicting future trends, optimizing decisions, and other aspects.

2.2 Main methods of data mining

Data mining technology plays a crucial role in data analysis, helping staff to gain a deeper understanding of project data and uncover potential risks and issues through various methods. Taking engineering auditing as an example, the main methods commonly used in data mining techniques are shown in Table 1.

3. CURRENT STATUS OF RISK IDENTIFICATION AND EVALUATION IN ENGINEERING AUDIT

Engineering audit is an important means of reviewing and supervising engineering projects, aimed at ensuring their compliance, efficiency, and safety. However, in the process of engineering audit, auditors face various risks, which may come from various aspects of the engineering project, such as contract management, fund management, quality management, etc. Therefore, accurate identification and evaluation of engineering audit risks are of great significance [3].

Table 1: Analysis of Data Mining Methods

Number	Method	Describe	Related examples
1	Association analysis	Discovering the correlation between data items, that is, whether the appearance of one data item is related to other data items	Analyze the correlation between material procurement and cost overruns in engineering projects, identify potential cost control issues
2	Classification	Divide the samples in the dataset into different categories and use a classification model to predict the class of new samples	Based on historical audit data, classify engineering projects into high-risk, medium risk, and low-risk categories, and predict the risk level of new projects
3	Cluster	Divide the samples in the dataset into multiple groups or clusters, so that the similarity of samples within the same group is high	Cluster analysis of engineering projects to identify project groups with similar characteristics for targeted auditing and evaluation
4	Prediction	Predict the future based on historical data and use predictive models to forecast future trends and changes	Using time series analysis to predict the future cost trend of engineering projects, providing a basis for budget formulation and cost control

3.1 Difficulties in Identifying Engineering Audit Risks

In the field of engineering auditing, auditors face various challenges. Firstly, engineering projects involve a large and complex amount of data, with a wide variety of types, including but not limited to contract documents, construction drawings, financial statements, etc. These data are not only massive in quantity, but also diverse in format, making audit work tedious and arduous, and placing extremely high demands on the analysis and processing abilities of auditors. Secondly, engineering audit risks are concealed, and risk points are often hidden in various aspects of the engineering project, such as the ambiguity of contract terms, abnormal financial flows, etc. These risk points are not easily discovered and identified by conventional audit methods, which increases the difficulty of audit work. In addition, engineering auditing requires a high level of professional knowledge from auditors. It not only requires auditors to master knowledge in multiple professional fields such as engineering cost, engineering management, and financial management, but also requires them to have the ability to comprehensively apply this knowledge to deal with various complex issues in engineering auditing. Therefore, engineering audit is not only a highly technical job, but also a task that requires high standards of professional competence and

comprehensive ability for auditors. In this context, how to use modern technology to improve audit efficiency and accuracy has become a topic worthy of in-depth research.

3.2 Methods and limitations of engineering audit risk assessment

The purpose of engineering audit risk assessment is to quantitatively analyze the potential risks in engineering projects, in order to determine the degree of risk and the priority of its treatment, in order to ensure the smooth progress of engineering projects and effective control of risks. At present, the commonly used engineering audit risk assessment methods in the industry include expert scoring method, analytic hierarchy process, etc. Although these methods can assess risks to a certain extent, they also have some obvious limitations. Firstly, these methods often have strong subjectivity, especially the expert scoring method, which highly relies on the personal experience and judgment of experts, and is therefore susceptible to the influence of subjective biases of experts, which may lead to bias and uncertainty in the evaluation results. Secondly, these evaluation methods often require a lot of time and effort in the implementation process. For example, the Analytic Hierarchy Process requires in-depth analysis and complex comparisons of various details of engineering projects, which not only increases the workload of auditing but may also affect the project schedule due to time delays. Finally, due to the unique characteristics and risk points of different engineering projects, these traditional risk assessment methods often lack sufficient adaptability and are difficult to flexibly respond to the specific needs of various engineering projects.

4. THE ADVANTAGES AND PROCESS OF APPLYING DATA MINING IN THE IDENTIFICATION AND EVALUATION OF ENGINEERING AUDIT RISKS

4.1 Advantages of Data Mining in Identifying and Evaluating Engineering Audit Risks

Data mining technology has broad application prospects in identifying and evaluating engineering audit risks. With the continuous expansion and increasing complexity of engineering projects, the amount and types of data faced in the audit process are becoming increasingly diverse. Through data mining techniques, auditors can extract useful information and knowledge from massive amounts of data, identify potential risk points, and accurately evaluate risks, thereby improving audit efficiency and accuracy.

4.2 Application process of data mining in engineering audit risk identification and evaluation

4.2.1 Data Preprocessing

Data preprocessing is a crucial step in the process of data mining, which directly affects the effectiveness of subsequent analysis and the reliability of the results. High quality data is the foundation for building effective risk identification models. In engineering auditing, data preprocessing mainly includes data cleaning, data conversion, and data integration. Data cleaning aims to eliminate noise, errors, and missing values in data, for example, by verifying erroneous data in financial statements with relevant departments to improve data accuracy. Data conversion is the process of transforming raw data into a format and type suitable for data mining, such as converting text data into numerical data through word segmentation and stop word removal, or normalizing time series data to eliminate dimensional differences. Data integration involves merging data from different sources to form a complete dataset, such as integrating contract documents, construction drawings, and financial statements together for correlation analysis and comprehensive evaluation.

4.2.2 Construction of Risk Identification Model

After completing data preprocessing, auditors need to build a risk identification model to identify potential risk points. The process of model construction includes feature selection, model training, and model validation. Feature selection uses correlation analysis, principal component analysis (PCA), or tree based feature selection methods to screen out risk related feature variables from raw data, such as contract amount, construction period, and engineering quality, thereby reducing the dimensionality and complexity of the data and improving the efficiency and accuracy of the model. During the model training phase, known risk case data is utilized to train the model using machine learning algorithms such as decision trees, support vector machines (SVM), random forests, or neural networks. This enables the model to learn the characteristics and patterns of risk points, as well as the ability to predict and classify new data. Subsequently, the model was validated and evaluated using methods such as cross validation and confusion matrix to ensure its accuracy and reliability. If the model performance is not ideal, it can

be optimized by adjusting the model parameters or selecting other algorithms to improve its ability to identify risks.

4.2.3 Risk assessment

On the basis of risk identification, auditors need to evaluate and quantify the identified risks in order to develop corresponding response strategies. Risk assessment mainly includes risk measurement, risk ranking, and risk response strategy formulation. Risk measurement quantifies identified risks through methods such as probability statistics, fuzzy comprehensive evaluation, or Analytic Hierarchy Process (AHP) to determine their magnitude and severity, such as calculating the probability of risk occurrence and its impact on the project. Risk ranking prioritizes risks based on risk measurement results to determine which risks need to be prioritized and focused on, in order to allocate audit resources reasonably and improve audit efficiency and quality. Finally, based on the risk ranking results, corresponding risk response strategies are formulated, including risk avoidance, risk reduction, risk transfer, and risk acceptance measures. For high-risk projects, methods such as increasing audit frequency or introducing third-party audits can be adopted to reduce the likelihood and impact of risk occurrence.

5. CASE STUDIES

In order to verify the feasibility and effectiveness of data mining techniques in identifying and evaluating engineering audit risks, this paper selects a large-scale infrastructure construction project as a case study for in-depth analysis. This project involves multiple contractors, supervisory units, and rich financial and construction data. Auditors have systematically processed and analyzed the project data through the application of data mining techniques, aiming to improve the efficiency and accuracy of audit work.

5.1 Case Background

The engineering project studied in this case is a new subway project in a certain city, with a total investment of approximately 5 billion yuan and a construction period of five years. It involves multiple construction units, supervision units, and suppliers. During the project implementation process, the audit department needs to conduct a comprehensive review of the project's compliance, fund utilization efficiency, engineering quality, and safety management. However, due to the large scale of the project and the diverse types of data, including contract documents, construction drawings, financial statements, progress reports, etc., traditional audit methods are difficult to efficiently and comprehensively identify potential risks. Therefore, the audit department has decided to introduce data mining technology in order to enhance the ability to identify and evaluate risks through intelligent means.

5.2 Data Mining Process

5.2.1 Data Collection and Preprocessing

The auditors first collected various data related to the project, mainly including: 50 contract documents, covering the contract texts signed between various contractors and project parties; 200 construction drawings, including design drawings and construction change records; 500 financial statements, involving cash flow, cost accounting, etc; And 60 project progress reports, submitted once a month.

In the data preprocessing stage, the collected raw data was checked and about 10% of the data was found to have input errors and missing values, which were corrected. For example, in the financial statements, 5% of the amount input errors were corrected and 2% of the missing data were supplemented through verification with the finance department. Subsequently, different types of data are converted into formats suitable for data mining and analysis. The contract text is segmented and keyword extracted using natural language processing (NLP) technology, transforming unstructured text data into structured numerical data; Financial data has been standardized to eliminate dimensional differences between different data sources. Finally, integrate data from different sources to form a comprehensive dataset, such as linking contract amounts with actual expenditure data, construction progress with fund flow data, for cross source correlation analysis. After a series of data preprocessing work, a comprehensive dataset containing 10000 records and 50 feature variables was finally formed, significantly improving the quality and usability of the data.

5.2.2 Construction of Risk Identification Model

After completing data preprocessing, auditors begin building a risk identification model, with the following specific steps:

(1) Feature selection:

Through correlation analysis and principal component analysis (PCA), feature variables highly correlated with risk were selected from the raw data, including contract amount, construction period, number of engineering changes, abnormal cash flow, etc. After feature selection, the dimensionality of the model decreased from 50 to 15, reducing computational complexity while improving the accuracy of the model.

(2) Model training:

Train the filtered data using the random forest algorithm. Random forests have the advantages of handling high-dimensional data and resisting overfitting, making them suitable for the complex risk identification needs of this project. Auditors use 80% of the data for training and 20% for testing model performance.
(3) Model validation:

The model was rigorously evaluated through cross validation and confusion matrix. The accuracy of the model on the test set reached 88%, the recall rate was 85%, and the F1 value was 0.86, indicating that the model has high reliability and effectiveness in identifying potential risks.

5.2.3 Risk assessment and formulation of response strategies

On the basis of the risk identification model, auditors further quantified and prioritized the identified risks. In terms of risk measurement, the fuzzy comprehensive evaluation method is adopted to quantitatively analyze each identified risk point, evaluate its probability of occurrence and potential impact. For example, the ambiguity of contract terms is assessed as high risk, while abnormal cash flow is assessed as medium risk. Subsequently, based on the results of risk measurement, the Analytic Hierarchy Process (AHP) was used to prioritize the risks. The sorting results show that abnormal fund flow and ambiguity of contract terms are the highest priority risks that need to be addressed first. In response to these high priority risks, auditors have developed specific response strategies: for abnormal fund flows, they plan to strengthen fund monitoring, establish a multi-level approval mechanism, and regularly review the fund flow situation; For the ambiguity of contract terms, the contract terms will be re examined and clarified to ensure the clarity and specificity of the contract text, in order to reduce potential legal disputes.

5.3 Analysis of Data Mining Results

By applying data mining techniques, auditors successfully identified and evaluated multiple potential risk points in the newly built subway project in the city. The overall risk level of the project has decreased by about 45% compared to before. The specific data is as follows:

Table 2: Comparison of Risk Levels Before and After Risk Identification

Risk category	Identify the pre risk level	Identify the risk level after identification	Reduce the amplitude
Contract management risk	Tall	Centre	50%
Risk of fund management	Tall	Centre	40%
Engineering quality risk	Centre	Low	50%
Progress management risk	Centre	Low	50%

And by formulating and implementing corresponding risk response strategies, the audit efficiency of the project has been improved by about 25%, and the audit accuracy has been improved by about 20%, significantly enhancing the scientific and standardized nature of project management.

6. CONCLUSION

Data mining technology has broad application prospects in identifying and evaluating engineering audit risks.

Through data mining techniques, auditors can extract useful information and knowledge from massive amounts of data, identify potential risk points, and accurately evaluate risks. This article explores the application of data mining technology in the identification and evaluation of engineering audit risks, and verifies its feasibility and effectiveness through practical case analysis. However, the application of data mining technology in engineering auditing still faces some challenges and issues, such as ensuring data quality and interpretability of models. Therefore, in future research, it is necessary to further strengthen the application research and exploration of data mining technology in engineering auditing, in order to promote the continuous innovation and development of engineering auditing work.

REFERENCES

[1] Wu, W. (2024). Research on cloud infrastructure for large-scale parallel computing in genetic disease.
[2] Chen, J. (2025). Data Quality Quantized Framework: Ensuring Large-Scale Data Integration in Gig Economy Platforms.
[3] Peng, Q., Zheng, C., & Chen, C. (2024). A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2240-2249).
[4] Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., ... & Wu, Z. (2024). 3d vision-language gaussian splatting. arXiv preprint arXiv:2410.07577.
[5] Deng, T., Bi, S., & Xiao, J. (2025). Transformer-Based Financial Fraud Detection with Cloud-Optimized Real-Time Streaming. arXiv preprint arXiv:2501.19267.
[6] Zhou, Y., Wang, Z., Zheng, S., Zhou, L., Dai, L., Luo, H., ... & Sui, M. (2024). Optimization of automated garbage recognition model based on resnet-50 and weakly supervised cnn for sustainable urban development. Alexandria Engineering Journal, 108, 415-427.
[7] Xie, X., Luo, Y., Peng, H., & Ding, C. RTop-K: Ultra-Fast Row-Wise Top-K Selection for Neural Network Acceleration on GPUs. In The Thirteenth International Conference on Learning Representations.
[8] Zhao, S., Xu, Z., Zhu, Z., Liang, X., Zhang, Z., & Jiang, R. (2025). Short and Long-Term Renewable Electricity Demand Forecasting Based on CNN-Bi-GRU Model. IECE Transactions on Emerging Topics in Artificial Intelligence, 2(1), 1-15.
[9] Liu, D., Wang, Z., & Liang, A. (2025). MiM-UNet: An efficient building image segmentation network integrating state space models. Alexandria Engineering Journal, 120, 648-656.
[10] Tian, Q., Wang, Z., & Cui, X. (2024). Improved Unet brain tumor image segmentation based on GSConv module and ECA attention mechanism. arXiv preprint arXiv:2409.13626.
[11] Shen, Z., Wang, Y., Hu, K., Wang, Z., & Lin, S. (2025). Exploration of Clinical Application of AI System Incorporating LSTM Algorithm for Management of Anesthetic Dose in Cancer Surgery. Journal of Theory and Practice in Clinical Sciences, 2, 17-28.
[12] Xu, J., Wang, Y., Chen, H., & Shen, Z. (2025). Adversarial Machine Learning in Cybersecurity: Attacks and Defenses. International Journal of Management Science Research, 8(2), 26-33.
[13] Shen, Z., Wang, Z., Chew, J., Hu, K., & Wang, Y. (2025). Artificial Intelligence Empowering Robo-Advisors: A Data-Driven Wealth Management Model Analysis. International Journal of Management Science Research, 8(3), 1-12.
[14] Chew, J., Shen, Z., Hu, K., Wang, Y., & Wang, Z. (2025). Artificial Intelligence Optimizes the Accounting Data Integration and Financial Risk Assessment Model of the E-commerce Platform. International Journal of Management Science Research, 8(2), 7-17.