Reasons and Control Strategies of Construction Project over Budget Based on Big Data Analysis

Honglie Jiang*
Liaoning Jianzhu Vocational College, Liaoyang111000, China

*Corresponding author e-mail: 5013543@qq.com

Abstract. During the construction process of construction projects, the cost budgeting work is an important guarantee for the entire project construction. Due to the influence of the social and economic market and the error problems caused by the related cost staff during the work process, these external factors will affect the construction project. The overall cost budgeting work has had a serious impact. The internal cost of construction projects has received the full attention of the company's internal management staff. It is mainly aimed at common problems and loopholes in cost work, which is manifested as the phenomenon of over budget in engineering. In this case, serious problems such as out-of-control costs in building construction will also arise, and eventually the overall economic benefits of construction enterprises will decline. Based on the above background, the purpose of this paper is to study the reasons for the over-budgeting of construction costs and their control strategies based on big data analysis. This article mainly addresses the problem that traditional anomaly data detection methods cannot detect anomalous data that is related to the description of the list. This paper proposes to use a list classifier and use the clustering center obtained by K-means clustering as a label. This kind of abnormal data is detected, and the experimental results of the proposed method are compared with traditional methods. Among them, the accuracy of the method of using the list classification method to detect abnormal data of the integrated unit price is 0.9107, while the accuracy of the traditional distance-based method is 0.8304. The experimental results show that the method proposed in this paper using the inventory classification method to detect abnormal data of the integrated unit price is more accurate than the traditional method, and can solve the problem of over-budgeting of construction costs to a certain extent.

Keywords: Construction Cost, Big Data, Comprehensive List Price, Abnormal Data Detection

1. Introduction
The construction industry has been a pillar industry in China for a long period of history. Due to the existence of a large amount of engineering cost estimation, auditing and settlement in real life, a field of engineering cost has been derived [1]. But at present, China's engineering cost big data industry is
still in the initial stage of data collection and standardization. At present, the main method for inventory classification is still a classification method based on rule base matching. This method is not universal and difficult to adapt to changing the actual situation is susceptible to the subjective factors of the rule base maker [2]. As for abnormal data monitoring, the industry generally only detects abnormal data that differs greatly from historical data, and lacks detection methods for cases where the data does not match the description [3].

As the construction cost budget control can not only promote the smooth development of the construction of the project, but also achieve the established economic goals, many research teams have begun to study the construction cost and achieved good results [4-5]. For example, by studying the core technology of BIM, according to the implementation ideas of fine engineering cost management, fully digging the application value of BIM technology and various links in engineering cost management, and proposed the establishment of a fine engineering management model based on BIM [6]. From the perspective of risk management, the risk management method was used to identify and analyze the risk factors that affect the project cost at each stage of the construction project, and established a risk evaluation model. Using the risk identification method combining the WBS-RBS method and the expert survey method, a risk identification list was obtained [7].

With the advent of the era of big data, significant changes have taken place from parallel machine architecture, computing resource expansion capabilities, and industrial application models. Through case studies, Janssen identified factors that influence decision-making based on big data (BD). BD is collected from different sources with various data quality and processed by various organizational entities, thus creating a big data chain that uses big Data is an evolutionary process, in which the gradual understanding of the potential of big data and the dailyization of the process play a vital role [8]. Cai provides a functional framework for identifying areas of acquisition, management, processing, and mining of IoT big data, and defines and describes several related technical modules based on their main features and functions. Then it analyzes the current research in the application of the Internet of Things, and in addition, identifies the challenges and opportunities related to the research of Internet of Things big data [9]. Smith and Nichols discuss "big data" human neuroimaging studies with a large number of topics and large amounts of data, which provide huge opportunities for new discoveries about the brain [10]. Fetene research analyzes the energy consumption rate (ECR) and mileage of battery electric vehicles (BEV), and uses the big data from real-life driving to understand the factors affecting its energy consumption [11]. Tang introduced a layered distributed fog computing architecture to support the integration of a large number of infrastructure components and services in future smart cities [12].

The main research content of this paper is the cause and countermeasures of over-budgeting construction cost. A method for detecting abnormal data that is difficult to detect by traditional methods such as the integrated unit price of the list and the description of the list is not proposed. This method uses the integrated unit price of the list as the Classification label and list description are used as classification objects. The method of list classification is used to analyze the relationship between the integrated unit price and the description in the normal data, so as to detect abnormal data that does not match the comprehensive list price and description. Among them, the accuracy of the method of using the inventory classification method to detect abnormal data of the integrated unit price is 0.9107 and the recall rate is 0.6198, while the accuracy of the traditional distance-based method is 0.8304 and the recall rate is 0.8683. It can be seen from the experimental results that the outlier data detection method proposed in this paper is better than the traditional method in detecting integrated unit price anomaly data, and can detect the abnormal data that the traditional method misses.

2. Proposed Method

2.1 Big Data Problem of Construction Cost
Engineering cost refers to the construction price of a series of projects such as construction engineering, landscaping engineering, road traffic engineering, etc. Due to the existence of a large
amount of engineering cost estimation, review, settlement and other tasks in real life, a field of engineering cost has been derived. After long-term development, a large amount of historical data has been accumulated in the field of engineering cost, mainly some structured text files. The mining and summary of the knowledge contained in these structured text files has a guiding role in the development of the future engineering cost field. Because the accumulated data is massive, combining the engineering cost field with big data technology will bring major changes to a series of tasks such as data storage, knowledge analysis and mining. However, China's engineering cost big data industry is still in the initial stage of data collection and standardization. This paper selects two specific problems, which are inventory classification and inventory comprehensive unit price abnormal data processing, as entry points to solve the problems encountered in the initial stage of China's engineering cost big data.

2.2 Defects of Traditional Abnormal Data Detection Methods
At present, the traditional abnormal data detection method for the integrated unit price of the list mainly detects such abnormal data that the difference between the integrated unit price data of the list and the historical data is too large. The occurrence of such abnormalities is mainly caused by artificial recording errors. Due to the long-term accumulation of data in the engineering cost industry, a large amount of historical data is saved, and the method of detecting such abnormal data with a large gap with historical data is more mature through the statistical rules. However, in practice, there is still a large number of other types of anomalies. Although the overall unit price of the list is slightly different from the historical data as a whole, the exceptions that are seriously inconsistent with the description of the list cannot be detected well by traditional detection methods for such anomalies.

2.3 Influence of Parameter K in K-means Clustering on The Accuracy of Inventory Classification
As the abnormal data detection problem of the list does not explicitly classify the normal list data, but in practice, there is a connection between the processes and material models in the description of similar lists. The final classification effect cannot be guaranteed, and if there are too few categories in the cluster, the subsequent effect of distinguishing normal data from abnormal data based on the cluster label and list description cannot be achieved.

3. Experiments
In this experiment, a list classifier is used, and the cluster center obtained by K-means clustering is used as a label to detect such abnormal data that the comprehensive unit price of the list does not match the description of the list. Because only the normal inventory data is used when training the inventory classifier, the inventory classifier will only learn the association between the normal inventory comprehensive unit price and the inventory description. When the normal list test data is subsequently input, the classifier can classify the list into the category represented by the correct comprehensive unit price clustering center based on the learned knowledge according to the list description. At this time, the comprehensive unit price and list of the list are considered The description match is normal data. If the input is abnormal list data, since the classifier has not learned the correspondence between the comprehensive unit price and the description of the list, the classifier is more inclined to classify the description of the list into the label it should actually correspond to. Or the correlation between the output list description and its own integrated unit price tag is weak. At this time, it is considered that the description of the list does not match the integrated unit price, which is abnormal data.

This experiment mainly studies the influence of the number of clusters K in the K-means on the effectiveness of the method for detecting abnormal data in the inventory. The evaluation indicators mainly include the accuracy rate and the recall rate of identifying abnormal data. The influence of the value of K in the K-means cluster on the two indexes of recall and accuracy was tested. The K value of the test is 5,7,9.
Table1. Schematic table of abnormal data detection indicators

|                | Actually abnormal data | Actually normal data |
|----------------|------------------------|----------------------|
| Recognized as abnormal data | TP | FP |
| Recognized as normal data    | FN | TN |

Combining Table 1, the formulas for accuracy and recall are as follows:

$$Recall\ rate = \frac{TP}{TP + FN}$$ \hspace{1cm} (1)

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$ \hspace{1cm} (2)

3.1 Experimental Steps

(1) Use traditional methods to determine whether the data is abnormal.
(2) If the traditional method considers the data to be normal, extract the inventory description attributes of the input inventory data as a classification basis.
(3) Obtain the two cluster centers A and B that are closest to the comprehensive unit price of the list, preprocess them and enter them into the trained list classifier.
(4) Calculate the list describing the probabilities Pa and Pb belonging to A and B, and take the smaller of 1-Pa and 1-Pb as the data anomaly index.
(5) Comparative analysis of experimental anomaly results.

4. Discussion

4.1 Influence of K-means Parameters in Abnormal Data Detection

This experiment mainly tests the influence of the value of K in the K-means cluster on the two indexes of recall and accuracy. The K value of the test is 5, 7, 9.

Table2. K value experiment table for abnormal data detection

| K value in K-means | 5     | 7     | 9     |
|-------------------|-------|-------|-------|
| Recall            | 0.8583 | 0.8876 | 0.9089 |
| Accuracy          | 0.9107 | 0.9015 | 0.8972 |

This experiment mainly tests the influence of the value of K in the K-means cluster on the two indexes of recall and accuracy. The K value of the test is a relatively stable K value of 5, 7, 9. It can be seen from the experimental results that the overall abnormal data detection method has a high recall rate for abnormal data. When the value of K is larger, the more types of training data are clustered, the higher the recall rate for abnormal data will be. Improvement, that is, the finer the granularity of the category of test data, the more abnormal data can be detected from the test data, but the corresponding accuracy rate is decreasing, mainly because when the number of categories is increased, some normal data will be affected. As anomalous data, the overall accuracy is reduced. In practical applications, when the recall rate is within an acceptable range, and at the same time, in order to reduce the manual review of abnormal data in the later period, after considering two indicators comprehensively, this article chooses K in the actual abnormal data detection work. K-means clustering method with a value of 5.
4.2 Comparison of Effect between Traditional Method and This Method

This article mainly proposes a method for the problem that traditional anomaly data detection methods cannot detect anomalous data that has correlation with the list description. Here we compare the method in this paper with the traditional anomaly data detection method based on distance.

Table 3. Comparison table between traditional method and this method

| Evaluation index         | Accuracy ratio | Recall ratio |
|-------------------------|----------------|--------------|
| Distance-based method   | 0.8304         | 0.6198       |
| List-based classification method | 0.9107         | 0.8683       |

The traditional method based on distance is mainly to calculate the difference between the test data and the training data. If the number of training data whose difference from the test data is less than a certain value is greater than the threshold, the test data is considered normal data, otherwise it is abnormal data.

From the experimental results, it can be seen that the method proposed in this paper for the use of inventory classification method to detect abnormal data of the integrated unit price is more accurate than the traditional method, and it has a better recognition effect for abnormal data. The main reason is that the traditional method only considers the distribution law of the integrated unit price in normal data, but does not consider the connection between the integrated unit price and the list description, which makes the traditional method difficult to detect abnormal data such as the difference between the list description and the integrated unit price. Not effectively.

5. Conclusions

The method proposed in this paper is more accurate than traditional methods for detecting abnormal data. The main reason is that the traditional method only considers the distribution law of the integrated unit price in normal data, but does not consider the connection between the integrated unit price and the description of the list. Detecting such anomalous data that does not match the comprehensive unit price is difficult to detect and the detection effect is not good. Therefore, it is of great research value to use big data analysis to solve the problem of over-budgeting of construction costs. This article shows the advantages of the application of big data analysis to over-budgeting of construction costs by using the advantages of big data analysis and traditional methods.
References
[1] Aladdin Alwisy., Ahmed Bouferguene., Mohamed Al-Hussein. Factor-based target cost modelling for construction projects[J]. Canadian Journal of Civil Engineering, 2018, 45.
[2] Bo Li, Songtao Guo, Yan Wu. Construction and Resource Allocation of Cost-Efficient Clustered Virtual Network in Software Defined Networks[J]. Journal of Grid Computing, 2017, 15(4):1-17.
[3] André O. de Araujo, Mauro Rezende Filho, Floriano C. M. Pires. A ship construction cost data management system[J]. Marine Systems & Ocean Technology, 2017, 12(3):93-103.
[4] Tolga Çelik, Saeed Kamali, Yusuf Arayici. Social cost in construction projects[J]. Environmental Impact Assessment Review, 2017, 64:77-78.
[5] B. Shao, X. Li. Research on hidden quality cost during construction of architectural engineering project based on bayesian network[J]. Journal of Xian University of Architecture & Technology, 2017, 49(3):317-321.
[6] J.-L. Wu, S.-R. Zhang, A.K. Singh. Optimum layout of freeway construction area based on crash cost control[J]. Changan Daxue Xuebao, 2017, 37(5):97-103.
[7] Yong Zhou, Zhigang Zou, Wenguang Tu. Construction of unique two-dimensional MoS2-TiO2 hybrid nanojunctions: MoS2 as a promising cost-effective cocatalyst toward improved photocatalytic reduction of CO2 to methanol[J]. Nanoscale, 2017, 9(26):9065.
[8] Janssen M, Haiko V D V, Wahyudi A. Factors influencing big data decision-making quality[J]. Journal of Business Research, 2017, 70.
[9] Cai H, Xu B, Jiang L, et al. IoT-Based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges[J]. IEEE Internet of Things Journal, 2017, 4(1):75-87.
[10] Smith S M, Nichols T E. Statistical Challenges in "Big Data" Human Neuroimaging[J]. Neuron, 2018, 97(2):263.
[11] Fetene G M, Kaplan S, Mabit S L, et al. Harnessing big data for estimating the energy consumption and driving range of electric vehicles[J]. Transportation Research Part D Transport and Environment, 2017, 54:1-11.
[12] Tang B, Chen Z, Hefferman G, et al. Incorporating Intelligence in Fog Computing for Big Data Analysis in Smart Cities[J]. IEEE Transactions on Industrial Informatics, 2017, 13(5):2140-2150.