Construction Project Cost Management Model Based on Big Data

Jing Xu¹,*

¹Yunnan University Of Business Management, Kunming, Yunnan, 650106, China
*Corresponding author E-mail: xujing@ynjgy.edu.cn

Abstract. The advent of the era of big data and the development of machine learning have brought new opportunities to the field of engineering cost in China. This paper mainly studies the construction project cost management model based on big data. On the analysis of the listing of large amount of data that exist in the proper nouns, the characteristics of dispersion characteristics of distribution, based on these features to implement and compare several listing classification method, the effect of and analyzes the reason of the result, refer to the classification results of traditional method at the same time, puts forward a project cost listing standardization based on polynomial bayesian classification method. In this paper, we propose a method to detect abnormal data by using the comprehensive unit price as the classification label, and study its association with the description of list according to the list classification method, and verify the effectiveness of this method through experiments.

Keywords: Big Data, Construction Cost, Standardized Classification of List, Abnormal Data Detection

1. Introduction
The survival and development of construction enterprises depend on completing more engineering construction tasks and creating considerable economic and social benefits through organizing and implementing construction projects. Cost management is an important part of modern enterprise management, as well as an effective measure to measure the resource consumption and supply of construction projects. It is also an important means of reasonable cost control and investment control. Scientific cost management is the embodiment of the core competitiveness of enterprises and the basic way to realize economic benefits. However, due to the limitations of traditional management concepts and the lag of business management methods, there are still many problems in construction project cost management, which cause high project cost and heavy project burden, which is not conducive to long-term development [1]. With the accelerated accumulation of informatization in the engineering field, new technologies represented by mobile Internet, Internet of Things and cloud computer computing are widely applied in the engineering field. With the rapid expansion of information body, the construction engineering cost industry has accumulated huge data resources, and the engineering
cost has entered the era of big data. In the context of big data, it is particularly urgent to make full use of these data resources, classify and integrate them, and effectively discover their value, so as to promote the development of the construction engineering industry.

Foreign engineering cost industry data collection and accumulation of the project cost information more seriously, by setting up special cost information consulting organization unit, timing and summary of various kinds of engineering cost information, used for analysis of all kinds of engineering is widely used in the field of project cost information data, and through the platform to provide a reference for industry personnel cost in project cost. Mishra proposed a model based on case-based reasoning, and used artificial neural network and multiple regression analysis to predict the project cost based on more than 100 sample projects, so as to solve the problem of making cost prediction in the case of limited information in the early stage of the project. Yet et al. studied an artificial neural network-based prediction model for hydropower stations to evaluate the feasibility of hydropower station projects. The above premise for project cost is a large number of standardized data, which requires related work of project cost information management in the field of engineering cost.

Main work of this article is in view of the project cost in the initial stage of construction of big data project listing standardized classification and comprehensive unit price list two major problems of abnormal data detection, respectively, this paper proposes a listing classification method based on polynomial bayesian classification method and a kind of comprehensive unit price as category labels, applying the idea of listing classification comprehensive unit price method of abnormal data detection.

2. Construction Project Cost Management Model under Big Data

2.1 Big Data and Construction Cost

(1) Necessity
In the context of the vigorous development of information management of project cost, the information update of China's current construction project cost management is slightly lagging behind, which leads to the inaccuracy and imprecision of cost management. Therefore, it is imperative to accelerate the construction cost information management.

1) the information processing in today's era of big data, through the accumulation and analysis the project cost management information, establishing the project related to more comprehensive information collection, transmission, processing, maintenance and use of network platform system, can have the ground effect using data to predict the future cost of change has occurred and the trend of development, in order to achieve the project cost achieve the goal of reasonable and effective control of.

2) Requirements of management: An effective project management platform system can efficiently sort, analyze and query various rapidly changing cost information, which will help realize the professionalism, standardization and standardization of construction engineering cost management;

3) Need for information communication: Establish an information platform to sort out and upload original technical and economic data of construction enterprises' management technical quota, relevant national policies, equipment and material price information of various places, and large and medium-sized construction projects, and provide basis for enterprises' decision-making by analyzing and forecasting data.

4) The need to save office costs: establish an information platform to realize electronic office and paperless office, effectively reduce management costs and realize green office.

(2) Information Management Process
The design unit shall prepare and complete investment estimation, preliminary budget estimate, construction drawing budget or quantity management documents and design change documents as required; The construction unit is the main body of the implementation of the project construction, and
it should prepare the fund demand plan, submit the design change, implement the design change and apply for the progress payment according to the requirements; The supervising unit shall, as required, review the fund demand plan, review the design change, review the project visa, review the progress payment, etc.; The evaluation unit shall conduct the evaluation in accordance with the completion settlement materials, drawings and settlement books provided by the construction unit, and finally issue the settlement evaluation report \[^{5-6}\].

Compared with the traditional construction cost management, the construction cost information management focuses on the realization of cost lean management by information technology. In the initial stage, make deviation analysis according to typical cases of the platform, and timely make rectification measures and error analysis;

In the stage of construction drawing design, the latest equipment, materials and materials should be listed on the platform to ensure the accuracy of construction drawing budget.

In the project settlement stage, through the comparison with recent typical cases, find out the reasons for the deviation, and form the cost analysis;

After the completion of the project, collect all the data that can be researched and settled, upload them to the platform database, and then check and process the background data to provide reference basis for the next project management decision.

2.2 List Classification and Abnormal Data Detection

(1) Inventory Classification

At present, the standardized classification of list in the field of engineering cost mainly relies on the traditional classification method based on rule matching. This method has some problems, such as manual rule summary, inability to find the implicit knowledge in the data and low efficiency, unstable rule base and poor generalization, etc. \[^7\]. Proper nouns list based on the analysis of the data quantity, key distribution characteristics of scattered, still focus on traditional methods focus on listing the main attributes, inspired by text classification methods at the same time, put forward a more intelligent, more can learn data list of implicit knowledge and is applicable to their own characteristics based on polynomial listing standardized classification of the bayesian method.

When the polynomial model is applied to the list classification, each list is regarded as a set of a series of features, and according to the conditional independence hypothesis, the probability of Listing D about a category tag Ci is expressed as the product of the probability of all features about the feature tag. The difference with Bernoulli's model is that Bernoulli's model only considers the occurrence of the feature words in a list without considering their occurrence frequency, while the polynomial model also takes the occurrence frequency of the feature words in a list as the reference standard, so Listing D is represented as the sequence of len feature words \[^8\]. Then the formula of D belonging to Ci category is as follows:

\[
P(C_i|d) = P(C_i) \prod_{j=1}^{len} P(\text{word}_j|C_i)
\]

Among them, \(P(C_i)\) is more convenient to calculate. The formula can be obtained by dividing the number of lists of all \(C_i\) categories in the list training set by the total number of training lists:

\[
P(C_i) = \frac{\text{count}(C_i)}{\text{count}(C)}
\]

When calculating \(P(\text{word}_j|C_i)\), combine all \(m\) list vectors belonging to \(C_i\) and get the sum of tf-idf values of all feature words in \(C_i\) list, and then divide the sum of tf-idf values of each feature word by the total tf-idf values, which can be expressed by the following formula:

\[
P(\text{word}_j|C_i) = \frac{\sum_{k=1}^{m} \text{tf-idf}_{j,k}(\text{if } C=C_i)}{\sum_{k=1}^{m} \sum_{j=1}^{n} \text{tf-idf}_{j,k}(\text{if } C=C_i)}
\]

\[
P(\text{word}_j|C_i) = \frac{\text{count}(\text{word}_j|C=C_i)}{\text{count}(\text{word}_j)|C=C_i)}
\]

(2) List Exception Data Detection
This paper proposes a comprehensive unit price list as the tag list based on classification of the comprehensive unit price list of abnormal data detection method, first of all a listing description of normal according to the comprehensive unit price clustering, clustering center as a label to each type of listing, according to normal after listing to do training data list of training a classifier to analyze list data associated with listing labels, for the test data on the list of data analysis and comprehensive unit price belongs to the connection degree of category labels, if less than a threshold, which associated listing the comprehensive unit price of abnormal data.

Simply selecting the comprehensive unit price of each listing as a label makes the category of the classification of each listing to produce too much too little data in each category, and as a result of chosen the same process and material model of normal comprehensive unit price list is in a reasonable range of fluctuation, this method can produce serious fitting issue. Through the distribution of the normal comprehensive unit price data obtained in this paper, it can be seen that the normal data has an obvious clustering trend, which inspires this paper to put forward the idea of taking the clustering center as the label of each type of data after the list comprehensive unit price clustering.

By K - means clustering method to cluster, according to the comprehensive unit price list for K - means algorithm eventually tend to choose the clustering center of the gap between each other, at the same time due to all the data points to the clustering center distance is relatively close, the characteristics and listing the comprehensive unit price obvious boundary between the whole and similar to listing description a list of the comprehensive unit price within a reasonable range of fluctuation in correspondence with the characteristics of the.

3. List Classification and Abnormal Data Detection Experiment

3.1 Parameter Tuning Experiment of Polynomial Bayesian List Classification Algorithm
List of naive bayesian classification algorithm is used in the introduction of Laplacian smoothing parameter method to prevent because some key P (wordj | Ci) to 0 and cause failure problem list of naive bayesian classification algorithm, and the experimental results show the Laplacian smoothing algorithm in the parameter list of values for the naive bayesian classification algorithm has great influence on the classification accuracy, thus setting up the experiment to compare between 0 and 1 under four different group of granularity of alpha parameters of naive bayesian classification accuracy of the algorithm and choose listing the classification effect of the optimal parameter. According to the experimental results, the Laplacian parameter with the highest accuracy is selected as the parameter in practical application.

3.2 Listing Abnormal Data Detection Methods
By using the list classifier trained before and using the clustering center obtained by K-means clustering as the label, it can detect such abnormal data that the comprehensive unit price of the list is inconsistent with the description of the list. Because only the normal list data is used in training the list classifier, the list classifier only learns the normal association between the list sum and the list description.

When the normal list test data is input later, the classifier can classify the list into the category represented by the correct comprehensive unit price clustering center according to the learned knowledge. At this time, it is considered that the comprehensive unit price of this list matches the list description and is the normal data. If the input is abnormal listing data, the comprehensive unit price because of the classifier is studied, the corresponding relationship between listing description of the classifier are more inclined to describe this list of classified into it should correspond to the actual label, or output listing description and its comprehensive unit price tag associations between weak, now think don't match the listing description and comprehensive unit price, is the abnormal data.

4. List Classification and Test Results of Abnormal Data
4.1 Effect of List Classification Method is Compared

Table1. Comparative experiment of classification methods

| List classification method          | Accuracy | Hundred sorting time |
|-----------------------------------|----------|----------------------|
| Bernoulli Bayes’ method           | 0.8592   | 0.18s                |
| Polynomial Bayesian method        | 0.8976   | 0.17s                |
| Convolutional neural network      | 0.7814   | 0.19s                |

Figure1. Comparative experiment of classification methods

As shown in table 1 and figure 1, the element list of bayesian classification accuracy of the method is much higher than the accuracy in listing convolutional neural network classification method, and the comparison of two list of naive bayesian classification model, a polynomial of bayesian method classification accuracy is higher than Bernoulli classification model, but it can be seen that polynomial bayesian model and naive bayes classification efficiency is basic on average every one hundred data classification time difference, and the classification of the naive bayesian classification method of time is slightly higher than the convolutional neural network classification method, but the gap is not big, in the acceptable range.

4.2 Abnormal Data Detection

(1) Influence of K-means Parameters in Abnormal Data Detection

Table2. Abnormal data detection K value experiment

| K value | 5          | 7          | 9          |
|---------|------------|------------|------------|
| Recall rate | 0.8517    | 0.8903    | 0.9172    |
| Accuracy | 0.9124    | 0.9046    | 0.8976    |

As shown in table 1, the total abnormal data detection method for abnormal data of the recall rate is very high, when the K value, the greater the value the more categories of training data clustering, the recall rate can be improved for abnormal data, that is when the test data, the category of the particle size of the thinner, the more can detect abnormal data from the test data, but the corresponding accuracy is on the decline, the main reason is that when the increase in the number of category classification, part of the normal data would be considered abnormal data, so that the overall accuracy.
In the practical application, the recall rate should be within the acceptable range, and in order to reduce the manual review of abnormal data in the later period, after comprehensive consideration of the two indicators, the k-means clustering method with K value of 5 is finally selected in the actual abnormal data detection work in this paper.

(2) Compared with Traditional Methods

As shown in Figure 2, the list classification method proposed in this paper is more accurate than the traditional method in detecting the abnormal data of the comprehensive unit price, and has a better recognition effect for the abnormal data. The main reason is that the traditional method only considers the distribution law of the comprehensive unit price in the normal data, but fails to consider the relationship between the comprehensive unit price and the list description, which makes the traditional method more difficult to detect the abnormal data inconsistent with the list description and the comprehensive unit price, and the detection effect is not good.

Figure 2. Traditional method is compared with the method in this paper

5. Conclusions

Main work of this article is in view of the project cost in the initial stage of construction of big data project listing standardized classification and comprehensive unit price list two major problems of abnormal data detection, respectively, this paper proposes a listing classification method based on polynomial bayesian classification method and a kind of comprehensive unit price as category labels, applying the idea of listing classification comprehensive unit price method of abnormal data detection. According to these two methods, a system architecture with the function of normalizing the classification of lists and the function of detecting the abnormal data of the lists is designed. According to the two methods proposed in this paper, the system architecture with the function of list normalization classification and list synthesis unit price anomaly data detection is designed.

References

[1] Stanka Šebela. Monitoring of Active Tectonic Structures-Project COST 625[J]. Acta Carsologica, 2016, 34(2):471-488.

[2] Mishra D , Mahanty B . A study of software development project cost, schedule and quality by outsourcing to low cost destination[J]. Journal of Enterprise Information Management, 2016, 29(3):454-478.

[3] Yet, Barbaros, Constantinou, et al. A Bayesian network framework for project cost, benefit and
risk analysis with an agricultural development case study.

[4] Li, Man R Y . Relevant Data in the Rising Tide of Big Data: A Text-Mining Analysis in Construction Safety Index[J]. 2015, 10.1007/978-3-319-12430-8(Chapter 4):61-73.

[5] Lu W , Shen Y,Wang T,et al. Fast Failure Recovery in Vertex-Centric Distributed Graph Processing Systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(4):733-746.

[6] Jalilzadehazhari E, Johansson P, Johansson J, et al. Developing a decision-making framework for resolving conflicts when selecting windows and blinds[J]. Architectural Engineering & Design Management, 2019, 15(5):357-381.

[7] Kim B C . Integrating Risk Assessment and Actual Performance for Probabilistic Project Cost Forecasting: A Second Moment Bayesian Model[J]. IEEE Transactions on Engineering Management, 2015, 62(2):158-170.

[8] Kulczycki P, Mazgaj A. PARAMETER IDENTIFICATION FOR ASYMMETRICAL POLYNOMIAL LOSS FUNCTION[J]. Information Technology & Control, 2015, 38(1):51-60.

[9] Lines B C, Sullivan K T, Hurtado K C, et al. Planning in Construction: Longitudinal Study of Pre-Contract Planning Model Demonstrates Reduction in Project Cost and Schedule Growth[J]. International Journal of Construction Education & Research, 2015, 11(1):21-39.

[10] Xiang Z, Fenghua W, Jian F U, et al. Mechanical Condition Monitoring of On-load Tap Changers Based on Chaos Theory and K-means Clustering Method[J]. Proceedings of the CSEE, 2015, 35(6):1541-1548.