Stream technologies of data analysis in organizational and technological design in construction

Pavel Kagan

Moscow State University of Civil Engineering, Yaroslavskoe shosse, 26, Moscow, Russia
E-mail: kagan@mgsu.ru

Abstract. The issues of intellectual processing of large data arrays in construction are considered. The concept of an automated system for analyzing big data is proposed. The technology of data mining assumes the possibility of estimating the time and cost of performing work on the basis of the used organizational and technological models and real indicators of time and labor observed by these models, and, accordingly, the formation of specific technical and economic indicators. Such indicators are of great importance in the current conditions of activity of construction organizations, they are used by the management of construction organizations in the preparation of tender documents, the development of construction organization projects, and in settlements with subcontractors. The proposed sequence of actions when using data mining technology to estimate the time and cost of work and resource consumption will allow at the stage of a feasibility study or decision to conclude contractual obligations to provide a better understanding of future costs and timing of work. This technique assumes the possibility of its use by both data analysis specialists, data scientist and process engineers, BIM managers, and the like and has a number of advantages due to the use of discriminant functions for data analysis.

1. Introduction

Currently, methods for processing and analyzing large amounts of data have become known. The general name of this direction, associated with a change in the methods of processing arrays of information, with the development of distributed storage and data processing systems that differ from classical databases, is Big Data [1, 2].

When analyzing data, an important point is their correct interpretation. The problem of using traditional data warehouses is the need to clean up the data and convert it to a specific format. Traditional approaches to the interpretation of data require pre-processing, for example, the introduction of dummy variables, etc. The methods used in the analysis of "big data" are not demanding on the type of data and allow you to work with unstructured data. At the moment, of interest is the identification and fixation of non-obvious relationships present in the analyzed data. The advantage of using big data is that you do not need to advance hypotheses for validation. Hypotheses appear in the process of data analysis. Analysis of the natural manifestations of dependencies is a basic approach when working with big data.

In the algorithms used to work with such data, data can be represented, for example, in the form of language words and use text analysis algorithms [3]. By converting quantitative and ordinal variables
to nominal, you can create a fully automated system for using data from various fields of human activity and data that are variables of various types.

With the development of modern technologies for processing big data, it becomes possible to use similar methods in such fields as construction (for example, [3, 4, 5, 10-16]). Analysis of a significant amount of input data allows to increase the level of reliability of reports and forecasts, to issue warnings when exceeding permissible indicators. The proliferation of models built on the basis of evidence arrays eliminates the limitations caused by inaccurate hypotheses and assumptions in the models.

2. Methods

When designing the technology for the production of construction works, an important point is the calculation and further use of technical and economic indicators, such as [6]:

- the duration of the work;
- labor costs and machine time;
- costing labor and machine time;
- work schedule;
- estimated cost calculations.

The calculation of technical and economic indicators should take into account the main, preparatory, concomitant and auxiliary processes: unloading, layout and storage of building structures and materials in the working area, organization of workplaces with the installation and fixing of scrubbing facilities, preparation and supply of solutions and other types of work [6, 7].

The formation of specific technical and economic indicators is an important stage in the calculations when designing the technology for the production of construction works [6, 7]. Material and technical resources, labor costs and machine time are tied to enlarged meters, for example: on the area - 10, 100 or 1000 sq.m.; per volume - 10, 100 or 1000 cubic meters; at a distance of 100 or 1000 meters; on weight - 100 or 1000 tons; for the quantity - 10 or 100 pieces.

Such indicators can be formed, in particular, on the basis of statistical characteristics obtained from the construction of facilities. These indicators can be supplemented by calculations of workers' wages, costs of machinery, equipment and materials [7].

Such indicators are of great importance in the current conditions of activity of construction organizations, they are used by the management of construction organizations in the preparation of tender documents, the development of construction organization projects, and in settlements with subcontractors.

To increase the competitiveness of the construction organization, it is recommended to use progressive, more stringent rules and regulations when designing technological processes. To calculate the resource requirements, production, departmental and local standards are used. In specific construction organizations, time standards for workers and machines can be developed according to timekeeping data at the organization’s construction sites or based on statistical generalizations of the results of production activities. The costs of labor and machine time are determined by the product of the volumes of work (in the process or operation) and the corresponding norms of time [7].

The implementation of technical regulation, on the basis of which detailed time standards for construction work are obtained, entails significant costs and difficulties. Not always construction organizations have the ability to carry out such work or order the development of the necessary standards. In a number of cases, a generalized rationing of work on the basis of the applied organizational and technological models and real indicators of time and labor observed by these models may turn out to be an acceptable way out. This approach can use any of the above techniques.

To this end, they determine the scope of work to be standardized, compose their classifier, characteristic organizational and technological models for their implementation, explicitly outline the factors affecting the duration of the work. Next, they fix the practical terms of work and other indicators at the facilities in relation to the created models and the classifier.
After the accumulation of a certain amount of data, they are analyzed, distributed by factors, and, ultimately, they receive the average values of the norms within the framework of the adopted classifier.

This approach, built on the principle of "from simple to complex", allows you to quickly get the first rough values of the norms, and then subsequently refine and modify them as data and experience of standardization accumulate.

An important task of the organization of standardization is the planning and implementation of periodic reviews of norms and standards in order to ensure their progressiveness, to prevent weakening of the norm and to reduce its level of tension. Foreign experience [8] indicates that most enterprises revise standards at relatively short intervals - from 6 months to 1 year, constantly tightening them even with small improvements in the production and labor processes, and minor organizational and technical measures.

Thus, an important task is the correct processing [8] and analysis of the results of measurements of the parameters of technological processes.

The discriminant analysis allows you to study the differences between two or more groups of objects for several variables (factors) at the same time [9], to identify the most important factors and categories. Modern classification methods are based on obtaining discriminant functions of a new type, which make it possible to assign this object to one of the groups. The discriminant functions also allow us to determine the importance of factors for the classification problem. To solve the classification problem, it is necessary to conduct preliminary studies on existing data to obtain discriminant functions.

3. Results

The developed method of data mining allows you to perform calculations, store information about factors, and provide classification based on available factors.

In practice, the calculation of discriminant functions is a difficult computational task, requiring the use of concurrent programming methods and the use of significant computing power. After calculating the discriminant functions, further work with the new data occurs in real time, fully in automatic mode.

In statistics, many well-known outcomes are usually called the values of the dependent variable. The value of the dependent variable can be predicted based on factors (predictors). The values of the dependent variable and factors are categorical, i.e. variables that take one of the given values (categories).

Allowed values consist of Russian and Latin letters, numbers. Spaces are not allowed. It is desirable to structure the input data in the form of tables (Table 1.).

| Dependent variable | Factor1 | Factor2 | Factor3 | ... | FactorN |
|--------------------|---------|---------|---------|-----|---------|
| Value1             | Category1 | Category5 | Category7 | ... | CategoryM |
| ...                | ...     | ...     | ...     | ... | ...     |

A table (training set) is necessary for constructing discriminant functions. Each category of the dependent variable will have its own discriminant function. Using discriminant functions, data is assigned to one or another group. The discriminant functions are calculated based on the available data with a known outcome (based on the known categories of the dependent variable), that is, the system is trained.

Discriminant functions are a set of rows. Suppose we have rows, which consists of the Factor__Categories \( F_j \__C_k(i) \) and its weight \( W_{jk}(i) \), separated by a comma: \( F_j\__C_k(i), W_{jk}(i) \).

For example, a discriminant function is a set of lines of the form:
Factor 25__Category1, 3.960233516483516
Factor3__Category5, 3.891928571428571
Factor 13__Category2, 3.846153846153846
The weight $W_{jk}(i)$ for $F_{jk}C_i(i)$ is measured in relative units, not absolute. The lines of the discriminant function are sorted in descending order of weights. An analysis of the importance of factors is based on the weights $W_{jk}(i)$ of the corresponding $F_{jk}C_i(i)$ discriminant functions. In the first places are the rows with the largest weights, in the last - with the smallest. $F_{jk}C_i(i)$ with the maximum weight after sorting is in the first line. The greater weight of $F_{jk}C_i(i)$ corresponds to the greater significance of $F_{jk}C_i(i)$. An analysis of the rows of the discriminant function allows us to identify the most important factors $F_{jk}C_i(i)$. From the discriminant function, it is possible to identify the “influence” of $F_{jk}C_i(i)$, rather than “positivity” or “negativity”.

The value of the discriminant function is calculated according to the following rule: if a specific $F_{jk}C_i(i)$ of the i-oó discriminant function exists for the sample element, then the corresponding weight $W_{jk}(i)$ is multiplied by 1, otherwise, by 0, the obtained values are added.

As new data accumulate, system training can be repeated. With the growth of the training database, the accuracy of discriminant functions increases.

The basis for finding the weights of discriminant functions is their sequential calculation for various elements of the training sample.

The algorithm consists of the following steps:

Step 1. Calculation of the initial values of the weights of the discriminant function for each group. For each $i$-th discriminant function, the initial values of its weights $W_{jk}(i)$ are calculated, for example, the frequency of occurrence of this $F_{jk}C_i(i)$ in the training set for this group divided by the total number of all $F_{jk}C_i(i)$ (taken as the weight for a specific $F_{jk}C_i(i)$ in the training sample).

Step 2. Calculation of base values for each group. For the discriminant function $D_i$ of each category of the dependent variable, the base value $B_i$ is calculated by multiplying all the weights of the discriminant function by 1.

Step 3. Fixed $i$, $j$, $k$ and weight. Investigation of the value of the weight $W_{jk}(i)$ for $F_{jk}C_i(i)$ - of the $i$-th discriminant function. The calculation of the deviation of the values of the discriminant function $D_j$ for the elements of its group $G_i$ is the value of $S_i$ and other groups $G_j$, where $j\neq i$ ($S_p$, $j\neq i$). $B_i$ is taken as the average.

The idea of calculating the weights $W_{jk}(i)$ of the discriminant function is that when substituting the weight at $F_{jk}C_i(i)$ into the discriminant function, the deviation $S_i$ for the group of elements $G_i$ should decrease as the deviations $S_j$, $j\neq i$ for other groups increase.

Step 4. Fixed $i$, $j$, $k$ and iterating over all weight values. Loop over the set of studied values of the weight $W_{jk}(i)$ of the factor category $F_{jk}C_i(i)$. Repeat Step 3 for the entire weight change interval $W_{jk}(i)$ with $F_{jk}C_i(i)$.

Step 5. Fixed $i$, iterating over all $j$ and $k$. The loop over all for $F_{jk}C_i(i)$ of the discriminant function $i$. Steps 3-4 are repeated for all $F_{jk}C_i(i)$ of the $i$-th discriminant function, and the factor categories $F_{jk}C_i(i)$ with the maximum weight ($F_{jk}C_1(i)$, $F_{jk}C_2(i)$,...) are selected first.

Step 6. The loop for all discriminant functions. Steps 3-5 are repeated for all discriminant functions of different $G_i$ groups (categories of dependent variable).

4. Discussion

This technique implies the possibility of use by both data analysis specialists and other specialists (technologists, BIM managers, etc.).

The main advantages of the proposed methodology:

1. After receiving the discriminant functions, users can use them in the future independently in their systems. Using, for example, decision trees does not allow to obtain such functions.

2. The algorithm is resistant to the quality of input data and does not require thorough data cleaning.

3. After training the system and calculating discriminant functions, the classification problem can be solved in real time on the basis of incoming data and without the involvement of experts.
4. In the process of constructing discriminant functions, “hidden dependencies” in the data are automatically detected and taken into account in discriminant functions.
5. For factors, categorical (nominal) variables are used, for example, the “semaphore” type (red, yellow, green) or interval1, interval2, or word1, word2, etc. It is easy to convert quantitative (in the form of intervals) and ordinal variables into categorical variables.
6. The analyzed data can be considered not only as tabular, but also as text messages. The ability to analyze text messages allows you to analyze, for example, texts of regulatory, technical, organizational, technological or executive documentation.
7. With the growth of computing capabilities, the quality of the obtained discriminant functions also increases.

5. Conclusions
The main conclusion:
The issues of intellectual processing of large data arrays in construction are considered. The concept of an automated system for analyzing big data is proposed.
The proposed data mining technology suggests the possibility of estimating the time and cost of performing work on the basis of the used organizational and technological models and real indicators of time and labor observed by these models, and, accordingly, the formation of specific technical and economic indicators.

References
[1] Foreman J. 2017 Many numbers. Big Data Analysis with Excel (Moscow, Alpina Publisher)
[2] Prokopec A. 2018 Learning Concurrent Programming in Scala (Moscow, DMK Press)
[3] Kagan P.B. 2018 Analytical studies of large data sets in construction Promyshlennyye i grazhdanskiye stroitel'stvo vol 3 pp 80-84
[4] Vikentieva O.L., Deryabin A.I., Shestakova L.V., Kychkin A.V. 2017 Synthesis of an information management system for subsystems of technical support of intelligent buildings Vestnik MGSU vol 12 Issue 10(109) pp 1191-1201
[5] Ivanov N., Gnevanov M. 2018 Big data: perspectives of using in urban planning and management MATEC Web of Conferences vol 170, 01107
[6] Osipovich L.M. 2016 BASIS OF QUALITY CONTROL OF CONSTRUCTION PRODUCTS - TECHNOCAL CARDS FOR PRODUCTION OF WORKS Proceedings of Novosibirsk State University of Architecture and Civil Engineering (Sibstrin) vol 19 2(62) pp 62-73
[7] MDS 12-29.2006. "Methodological recommendations for the development and design of the technological map" [8] Foreign experience of labor standards. Recommendations for use in domestic practice https://www.jobgrade.ru/2007/01/27/
[9] Afonin P.N., Afonin D.N. 2015 Statistical analysis using modern software (SPb: Information Center Intermedia) p 100
[10] Larin V.S., Zhelenzov M.M., Kagan P.B. 2019 PROCESSING BIG DATA FOR AUTOMATED CREATION OF SPECIFICATIONS IN MODERN BIM PROJECTS Sistematikhna stroitel'stva. Kiberfizicheskiye stroitel'nyye sistemy (Moscow: Publishing House MISI-MGSU) pp 372-377
[11] Volkov A.A., Kagan P.B., Afanasyeva A.A. 2019 CONSTRUCTION UNDER THE CONDITIONS OF DIGITAL ECONOMY DEVELOPMENT Sistematikhna stroitel'stva. Kiberfizicheskiye stroitel'nyye sistemy (Moscow: Publishing House MISI-MGSU) pp 63-67
[12] Kagan P. 2019 Big data sets in construction. E3S Web Conf. Volume 110, International Science Conference SPbWOSCE-2018 “Business Technologies for Sustainable Urban Development” https://doi.org/10.1051/e3sconf/201911002007
[13] Valpeters M., Kireev I., Ivanov N., 2018. Application of machine learning methods in big data analytics at management of contracts in the construction industry. MATEC Web of Conferences, 170, 01106

[14] Konikov A., Kulikova E., Stifeeva O., 2018. Research of the possibilities of application of the Data Warehouse in the construction area. MATEC Web of Conferences, 251, 03062

[15] Konikov A., 2019 The method of iterative use of the ABC-XYZ analysis in the construction industry. E3S Web Conf. Volume 110, International Science Conference SPbWOSCE-2018 “Business Technologies for Sustainable Urban Development”. DOI https://doi.org/10.1051/e3sconf/201911001073

[16] Ginzburg A., Shilova L., Adamtsevich A., Shilov L. 2016 Implementation of BIM-technologies in Russian construction industry according to the international experience / Journal of Applied Engineering Science, 14 (4), p. 457-460.