Application of Blockchain Technology and Data Mining Technology in Public Utilities Management

Cirenlajie Ci(✉)

Fujian Agricultural and Forestry University, Fuzhou, Fujian, China
c13375000971@163.com

Abstract. As a government management evaluation system, it must be relatively independent and integrated. However, the existing research results lack the research on the technical platform of performance, which makes the design limitations and the repair of the index system, can not connect the index design with the performance evaluation, and can not make full use of the existing relevant data and information, resulting in the weak practicability of the designed index. This paper mainly introduces the econometric analysis method of extracting conceptual data from a large number of data by using the decision tree method of data mining technology in the process of public utility management performance evaluation, so as to improve the reliability and effectiveness of evaluation and save evaluation time.

Keywords: Decision tree · Data mining · Public utilities management · Performance evaluation

1 Introduction

Public utilities are social public affairs that take the common interests and living standards of all the public as the basic content, and include the activities and results of necessary economic affairs, mainly including education, science and technology, culture, health, sports, social security, environmental protection and so on. Public utilities management is a process of adjusting and controlling public affairs according to law under the guidance of the government, so as to promote the coordinated development of the overall interests of the society. The research on the performance of public utilities management began from the perspective of human resource management. Campbell, McCloy, Borman and other foreign scholars have made outstanding contributions to the research on performance.

With the development of information technology and the implementation of transaction automation and e-government, a large amount of data has been accumulated in the field of public utilities management. Traditional query and data processing technology can not solve the problem of real information explosion. The decision-making errors and efficiency caused by information redundancy perplex managers. This requires making

I. Ahmad et al. (Eds.): STSIoT 2021, LNDECT 122, pp. 604–613, 2023.
https://doi.org/10.1007/978-981-19-3632-6_71

full use of the information processing capacity of the computer in order to automatically and efficiently obtain useful knowledge and information from the vast amount of data. This is KDD (knowledge discovery in database) technology. KDD is an automatic process of discovering useful, new and available information from a data set. Generally speaking, it includes three stages: data preparation, data mining (DM) and knowledge expression, interpretation and verification. DM is the core process of KDD. It mainly includes classification, clustering, regression analysis, generalization, construction of dependent patterns, change and deviation analysis, pattern discovery and path discovery, using decision tree method, neural network method, statistical method and so on. Among them, decision tree method has the advantages of high speed, high precision and simple generation mode, which is widely popular in data mining. The evaluation of job performance is not a new concept. A considerable number of experts and scholars at home and abroad have studied and explored it. The research contents mainly include the following aspects: first, study the variables, indicators and definitions of performance; Second, research performance evaluation; The third is to study the structure of performance; The fourth is to study the input and output of performance. Different schools have different perspectives and different analysis tools, and the conclusions are often different or even contradictory. Therefore, it is very necessary to strengthen the quantitative refinement of the evaluation indicators and integrate accurate quantitative analysis methods in the analysis. It can be divided into the following steps: (1) query the data sets related to the task to generate a multidimensional data view; (2) Summarize and summarize the original data to a higher level of abstraction; (3) Further summarize and standardize; (4) Describe the knowledge found. This paper consists of the following parts. The first part introduces the relevant background and significance of this paper, the second part is the related work of this paper, and the third part is data analysis. The fourth part is example analysis. The fifth part is conclusion.

2 Related Work

The authors propose an algorithm of actions to follow when introducing occupational standards in a higher education institution; examine the problem of applying occupational standards in the context of training bachelors and masters in economics and housing and public utilities management; and show that the HPU industry suffers from a lack of methodical literature on using the requirements of the occupational standard in working training programmes [1]. Aim of Bresciani et al. investigate the governance of Italian public utilities whose top management is engaged in balancing the conflicting pressures of the business model and the social functions [2]. This research study analyses the project management of agricultural facilities that protect agricultural food and public health from associated toxic hazardous landfill emissions and risks [3]. Lukmanova et al. present the needs to improve the controlling systems in public utilities [4]. Golladay et al. consider PES to include a range of knowledge necessary to develop a resilient and sustainable water management strategy based on ecological, environmental, and engineering principles [5]. Liu used pattern discriminant analysis technology and BP neural network model to build a local public utility management performance prediction model and uses 11 regions in the east, middle, and west of the country as samples to predict the local public utility management performance [6]. Based on the above background, the purpose of Ref. [7] is the application of big data information system in the field of public utilities management. At present, there are many evaluation systems for public health, but their function is different, so this paper provide a new evaluation system for public health, its mainly function is emergence management [8]. However what is the fundamentals of public utilities management? The paper shared the answer for us [9]. Other influential work includes Refs. [10]. The core process of public health management is artificial intelligence, which is realized through machine learning, statistical calculation and logical derivation. However, public health management itself is not a simple combination of multiple technologies, but a complete whole. It needs to connect various technical means and information resources to complete a series of tasks such as data collection, preprocessing, information analysis and result description, and finally present the results.

2.1 Data Mining Concepts

Data mining is to find the “knowledge Nuggets” hidden in the “data mine”, so as to help enterprises reduce unnecessary investment and improve capital return at the same time. Practice has proved that the potential return on investment brought by data mining to enterprises is almost endless. Innovative companies around the world have begun to use data mining technology to determine which customers are their most valuable customers, so as to re formulate their product promotion strategy (promote products to the people who need them most) and get the best sales with the least cost. For another example, telecom companies can use data mining technology to find customers with high probability of loss (no longer using the company’s services), possible fraud (no payment after call), potential large customers and so on. In short, data mining is to extract knowledge or mine knowledge from a large amount of data. Data mining is a basic step in the process of database knowledge discovery. The knowledge discovery process includes the next step, such as Fig. 1.

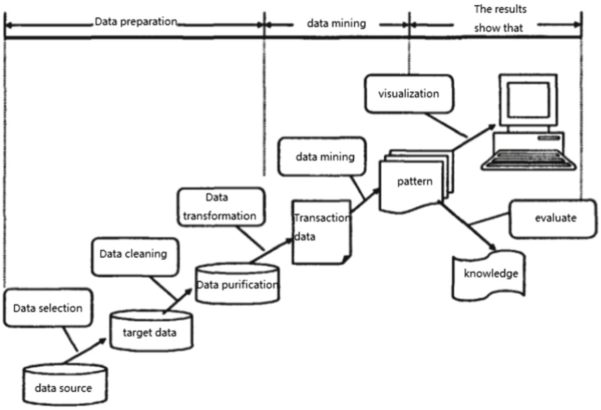

Fig. 1. Data mining and knowledge discovery

Data mining is related to the integration of multiple disciplines and technologies. This includes database technology, statistics, machine learning, high-performance computing, model recognition, neural network, data visualization, information extraction, image and signal processing, and spatial data analysis. Through data mining, interesting knowledge, rules or advanced information can be extracted from the database for observation or query from different angles. The discovered knowledge can be used in decision-making, process control, information management, query processing and so on. Data mining is one of the most important and promising subjects in the information industry.

The purpose factor of performance evaluation if we want to establish the performance evaluation index system of local government public utilities management, we must first determine its establishment basis, which is the basis of performance evaluation. The selection of performance evaluation indicators is determined by the purpose of performance evaluation. The fundamental indicators of performance evaluation are different in different regions. If the government wants to improve the performance of public utilities, the index system should pay more attention to the analysis of potential factors affecting the performance evaluation than the simple performance evaluation, and have an in-depth discussion on how to improve it. However, in real life, there are many kinds of performance evaluation indicators for local government public utilities management, and the number can not be ignored. In the evaluation, we can not be comprehensive and specific, so we can only evaluate the government management performance according to the evaluation purpose. Otherwise, not only the performance evaluation is meaningless, but also the operability is not strong, and the due value of evaluation is lost.

2.2 Data Mining System Structure

The essence of local government's function factor performance is the performance quality and degree of government's public utilities management function. Based on the government's management function, the local government's public utilities management performance evaluation index system is established φ The content of the indicators shall be carefully marked according to the management functions, so as to fully and accurately understand the main functions of local government management. From a macro perspective, the main functions of local governments include: education, scientific and technological development, cultural communication, health supervision, sports training, infrastructure construction, environmental protection, etc. θ The indicators to evaluate these functions need to establish corresponding performance evaluation systems, such as education performance evaluation system, science and technology development performance evaluation system, cultural communication performance evaluation system, health supervision performance evaluation system, sports training performance evaluation system, equipment construction performance evaluation system, social security performance evaluation system, environmental protection performance evaluation system, etc. Formulate corresponding performance evaluation indicators according to different functions. China is rich in land resources. Different regions have different historical factors and natural conditions. Therefore, when establishing indicators, we should fully consider that different functions affect the formulation of indicators. It is necessary to formulate not only a special index system, but also a general index system, which is analyzed and formulated by using quantity, so as to minimize the functional differences in different regions and the errors and negative effects caused by the differences. 3. The subjective and objective conditions of local governments will affect the performance evaluation system due to regional differences. This mainly includes three aspects: first, according to the requirements of the superior leaders and the general objectives put forward by the superior leaders, convey and complete the task allocation, refer to the relevant provisions, laws and regulations of the plan, etc.; Second, the quality of local personnel, economic and social development, the quality and degree of achievement of objectives, government management system, etc.; Third, the direct impact of the external environment, that is, the development quality and speed of relevant enterprises, regional characteristics, departments, etc. Take it as the basis of establishing the system, and establish the management performance evaluation index system of local government public utilities. Only by truly and comprehensively understanding, mastering, researching, analyzing and making use of these conditions can we reasonably formulate the management performance evaluation indicators of local government public utilities.

A typical data mining system has the following main components, and its system structure is shown in Fig. 2:

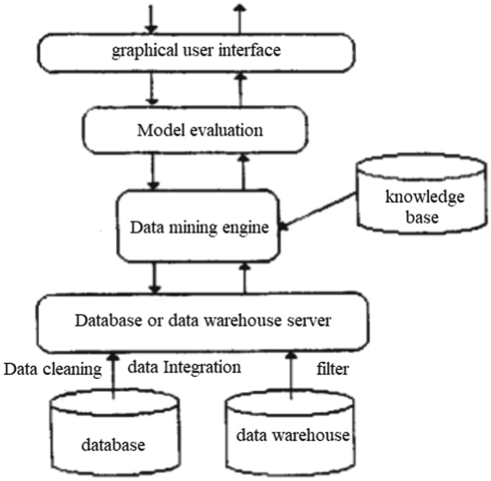

Fig. 2. Typical data mining system structure

2.3 Research on Decision Tree Algorithm

In many methods of data mining, classification algorithm is an important topic, which is the most widely used and studied by scholars. It can be used for prediction and decision-making. By learning and analyzing the experience data that have been classified in the past, the classification algorithm establishes the corresponding algorithm model to describe the differences of each category in the training data, and can classify the new data of unknown categories. By predicting the class labels of these new data, we can have a better understanding of the categories in the data and obtain the category knowledge of the new data, even though they are unknown.

Recursive top-down method is a common way in the learning process of decision tree. The general process is as follows: firstly, the attribute values of different variables to be compared are compared in the inner nodes of the tree, and then the branch is decided based on the different attribute values of different variables, According to this step, the classification of the variable attribute is obtained in the leaf node of the tree. Therefore, from the root node to the leaf node of the tree, each path in the middle is essentially a disjunctive rule. To sum up, a decision tree is essentially a set of disjunctive rules.

According to the learning process of decision tree, the algorithm of generating decision tree is divided into two steps: the first step is the generation of decision tree, the root node stores all the data to be classified during initialization, and then the data is segmented by recursion. The second step is to prune the decision tree. The essence of pruning is to eliminate noise or abnormal data similar to noise. When the data in an internal node belongs to a category or the attribute cannot be further divided, the data segmentation can be stopped.

3 Data Analysis

ID3 uses greedy method and recursively uses top-down divide and conquer method to generate decision tree. Whenever the branch node of the tree is selected, the information gain is measured, that is, the amount of information required to classify the current data is calculated before division. Then, after the segmentation according to the segmentation, the amount of information required to classify the current data is calculated again. The difference between the amount of information before and after splitting is the information gain of the splitting attribute. After calculating the information gain of all the current split attributes, the split attribute with the largest information gain is finally selected, which can make the amount of classification information of data decrease the fastest. Most of the decision tree induction algorithms follow this top-down classification method. The decision tree is generated from the training tuple set and their associated class labels. The decision tree consists of three types of nodes: root node, internal node (decision node) and leaf node. The root node and the inner node correspond to an attribute in the training tuple set, while the leaf node is the set of class label attributes in the classification. As the tree grows, the training set is recursively divided into smaller training subsets. The growth process stops until the leaf node, when the entropy is zero, and the instances in the corresponding instance set of each leaf node belong to the same class.

Assuming that the number of samples belonging to category s is, the amount of information needed to classify a given data object is Formula 1:

$$AOI(C_1, C_2, \cdots, C_m) = -\sum\nolimits_{i=1}^{m} p_i \log_2 p_i \tag{1}$$

The faster the entropy decreases, the better. In this way, we can get a decision tree with the smallest height. The smaller the value is, the purer the result is. For a given subset, the information quantity formula 2:

$$AOI\left(C_{1j}, C_{2j}, \cdots, C_{mj}\right) = -\sum\nolimits_{i=1}^{m} p_{ij} \log_2 p_{ij} \tag{2}$$

Split information degree spiltinfo is used to measure the breadth and evenness of an attribute's split data. The calculation formula is 3:

$$IGR(A) = \frac{IG(A)}{SpiltInfo(A)} \tag{3}$$

Information gain rate is a kind of compensation for the number of attribute values. The importance of attributes will decrease with the increase of split information degree. When the number of branches generated by data set splitting is small, the gain rate will increase accordingly, which effectively solves the problem of attribute multi value bias.

For numerical data, ID3 algorithm can't deal with continuous attributes, but can only deal with discrete attributes, so it discretizes the data set; for descriptive data, every node or concept of spanning tree is actually a range or interval, so we use grid method to quantify the statistical indicators.

4 Example Analysis

4.1 Generation Algorithm of Public Utility Management Performance Decision Tree

For numerical data, ID3 algorithm can not deal with continuous attributes, but only discrete attributes, so the data set is discretized; For descriptive data, each node or concept of the spanning tree is actually a range or interval. We use the grid method to quantify the statistical indicators. Here, taking the education management in public utilities as an example, we select three indicators for evaluation: the proportion of education in GDP (g), the number of full-time teachers per 100 students (s), and the proportion of college students in the population (P). The data are from cities in Hunan Province (Zhangjiajie, Changsha, Shaoyang, Zhuzhou, Xiangtan, Hengyang, Chenzhou, Xiangxi, Yiyang, Loudi and Yueyang), in order to avoid unnecessary impact caused by the evaluation, the specific unit name is hidden in the text, supplemented by a, B, C… Instead. Interested readers can refer to relevant literature or the local statistical yearbook of that year.

The virtual government education management performance indicators are constructed, and the statistical indicators of each region are weighted and averaged to form a virtual evaluation index. According to the given index weight, the proportion in the education management performance evaluation is 0.5391, 0.2971 and 0.1638 respectively through factor conversion; the classification comparison results are obtained by comparing the actual indicators with the virtual indicators.

The principle of data preprocessing and sorting is to compare the regional indicators with the virtual indicators and convert them into a unified format suitable for data mining. The corresponding value range of each type is that if 20% of the performance of the virtual government is excellent and 10% is good, generally less than - 10% is poor and - 20% is very poor (expressed in 1–5 respectively). It can be seen that "education accounts for the proportion of GDP" "The information value of this attribute for training set classification is the largest, so this index is selected as the basic attribute for division, and the decision tree is generated by analogy, as shown in Fig. 3.

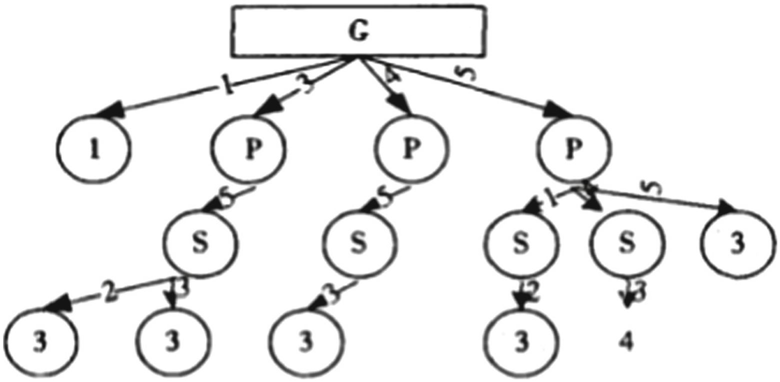

Fig. 3. Education management performance index evaluation decision tree

4.2 Optimization and Pruning of Decision Tree

Decision tree is one of the most important classification methods in data mining. The training data sets are combined repeatedly to construct the decision tree. If the data of the training data set can accurately reflect the character of the analysis object, the decision tree of the training data set can correctly classify the problem. However, there are many uncertain factors in practical problems. By using the decision tree structure algorithm to classify these data, the obtained decision tree becomes huge and complex, and the generated knowledge rule set becomes huge and complex. Therefore, it is necessary to prune the decision tree. The purpose of pruning is to reduce the fluctuation caused by the noise in the training set. Because many branches may reflect the noise or isolated points in the training data, pruning can improve the accuracy of classification of unknown data sets, avoid the systematic error caused by such points, and improve the mining accuracy. There are two kinds of pruning methods:

(1) Pruning afterwards. It allows the decision tree to get the most full growth, and then according to certain rules, it cuts out the leaf nodes or branches which are not generally representative in the decision tree. After pruning, the pruned branch node becomes a leaf node, and it is marked as the class with the largest number of categories in the sample it contains. This is a process of pruning while checking. Of course, this process may be at the expense of prediction accuracy. When the standard deviation of classification data increases rapidly, pruning should be abandoned, otherwise the classification results will be affected.

(2) This deliberative method determines whether the point generation process should be stopped as soon as possible and continue to provide the training sample set of the current node. If the blanking key stops, the current node becomes an inversion node. Leaf nodes can include different types of training samples. Hiccups are formed in front of the brand, so the general way to hiccup in advance is to set the maximum height (layer) of the crystal tree to limit the growth of the tree. Another method is to set the minimum number of records to include in each node. If the number of records of the node is less than this value, the segment stops. But it is often difficult

to determine such a reasonable threshold. If the threshold is too large, the decision tree will be too simple, and if the threshold is too small, the redundant branches will not be pruned. The decision tree generated in this paper adopts this kind of method. Of course, pre pruning and post pruning can be combined with each other. Using the advantages of the two methods, it can save time and improve the accuracy of the tree, thus forming a hybrid pruning method to obtain a more reliable decision tree.

5 Conclusion

Based on the three indicators of educational management performance, this paper puts forward a kind of data mining method for public management performance evaluation. Its practicability lies in that it can analyze the results of management performance by substituting the corresponding case data under its generated decision tree evaluation rules. At the same time, by calculating the information entropy of each performance index, find out the main factors that affect performance, so as to find measures to improve performance.

Of course, the performance of public utilities management is affected by various factors. The structure of evaluation index is complex, non market, non discrete and non deterministic, which makes the performance itself fuzzy. Although there is a great correlation between the satisfaction degree of service objects and the administrative efforts of government organs, in fact, it is not entirely determined by the administrative efforts of the government, but the result of the comprehensive action of many factors, among which the social environment factors have an important influence on the satisfaction degree. In addition, the relationship between government cost and management performance of public utilities is not completely linear. Therefore, as a basic data mining technology, decision tree is used to describe the performance evaluation of public utilities with a large number of data. It is a very useful fuzzy mathematical evaluation tool to provide a wide range of evaluation and reasonable quantitative space.

References

1. Astratova, G.V., Shklyaeva, N.A.: On application of occupational standards in training bachelors and masters in economics and housing and public utilities management. J. Manag. Syst. **158**, 107704 (2018)
2. Bresciani, S., Del Giudice, M., Papa, A.: Public control and strategic governance in state-owned public utilities: empirical evidence from Italian listed firms. Sinergie Ital. J. Manag. **74**(18), 3823–3831 (2018)
3. Koliopoulos, T., et al.: A project management utility for agricultural facilities and public health. J. Eng. Stud. Res. **155**, 23–36 (2018)
4. Lukmanova, O., Volkova, E., Zabolotnyi, A., Gorelik, A.: Blockchain technology for public utilities. In: 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), pp. 9–16 (2019)
5. Golladay, S.W., Craig, L.S., DePalma-Dow, A.D., Emanuel, B.N., Rogers, S.G.: Building resilience into water management through public engagement. Freshw. Sci. **29**(2), 300–312 (2020)

6. Liu, X.: Forecasting utilities management performance based on discriminant technology and BP neural network. J. Intell. Fuzzy Syst. **20**(4), 609–616 (2021)
7. Yu, Q.: Application of big data information system in the field of public utilities management. J. Inf. 785–788 (2021)
8. Xie, X., Li, H.: Analysis on the evaluation system of emergency response ability to public health emergencies. J. Health Manag. **20**(4), 609–616 (2020)
9. Spellman, F.R.: Fundamentals of public utilities management. Int. Public Util. Manag. **220**, 106918 (2020)
10. Spellman, F.R.: What is public utility management? Int. Public Util. Manag. **220**, 106877 (2020)