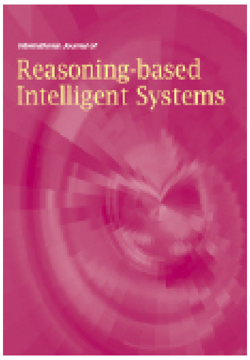
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

A network big data classification method based on decision tree algorithm

Nian Xiao, Siguang Dai

DOI: 10.1504/IJRIS.2023.10053641

Article History:

Received:	15 August 2022
Last revised:	02 October 2022
Accepted:	07 November 2022
Published online:	19 March 2024

Copyright © 2024 Inderscience Enterprises Ltd.

A network big data classification method based on decision tree algorithm

Nian Xiao*

Department of Information Engineering,
Wuhan Vocational College of Communications and Publishing,
Wuhan 430223, China
Email: nianxiao6@mls.sinanet.com
*Corresponding author

Siguang Dai

School of Management,
Hubei University of Education,
Wuhan 430205, China
Email: daisg@hue.edu.cn

Abstract: Aiming at the problem of low accuracy and low efficiency of network big data classification, a new network big data classification method based on decision tree algorithm is designed. First, the crawler manager circularly collects network big data, sets the collection threshold and randomly generates crawler signatures, so as to continuously collect and update data. Then, the directed graph of network big data is constructed that automatically select and extract the key feature attributes of network big data, and the interference factors of feature data are extracted. Finally, the network big data classification decision tree is constructed to obtain the optimal gain data, the node attributes of the data are determined, and the classification algorithm design combined with recursive call rules and classification termination conditions is completed. Experimental results show that the algorithm can improve the accuracy and efficiency of data classification.

Keywords: decision tree algorithm; network big data; classification; crawler signature; digraph; information gain; recursive call.

Reference to this paper should be made as follows: Xiao, N. and Dai, S. (2024) 'A network big data classification method based on decision tree algorithm', *Int. J. Reasoning-based Intelligent Systems*, Vol. 16, No. 1, pp.66–73.

Biographical notes: Nian Xiao received her MAEng from the School of Computer Science and Engineering, Wuhan Institute of Technology in 2009. She is currently an Associate Professor in the Department of Information Engineering of Wuhan Vocational College of Communications and Publishing. Her research interests include software development, data mining and artificial intelligence algorithms.

Siguang Dai received her PhD in Management Science and Engineering from the Wuhan University of Technology in 2020. She is currently an Associate Professor in the School of Management of Hubei University of Education. Her research interests include business intelligence, data analysis and service management.

1 Introduction

The internet provides a huge amount of relevant data, entertainment, learning and other methods, and also provides the key media and ways for the whole society to obtain data conveniently. According to relevant research, the number of users using the internet has exceeded 900 million in 2019. The internet has become an indispensable medium for people's life and an important role that affects people's life (Wang et al., 2020). With the continuous development and deepening of this trend, the amount of data generated in the network is also very considerable. There are many kinds of network data generated through different forms, and there is many key information in the data, which has become a key resource. Data in the network is a kind of data generated in a specific connection form according to the network traffic and the whole network, which has many characteristics (Cgab et al., 2020). Each kind of data contains different information due to its different characteristics. The classification of a large number of network data is very difficult due to the problems of the data itself (Tsai and Lee, 2020), which leads to the key data information cannot be fully mined and applied.

Copyright © 2024 Inderscience Enterprises Ltd.

The classification of network big data is of great significance (Yang et al., 2022). However, the current network big data classification has some limitations. The classification accuracy and classification effect of big data classification methods are low, leading to low practicability, therefore, relevant scholars have deeply studied the classification of network big data.

Sleeman and Krawczyk (2021) designed a classification of multi category unbalanced big data on Spark. This method constructs a composite framework for dealing with multi class big data problems, and embeds big data information into the popular resampling algorithm to achieve information balance of multi class big data. The data features determined in the classification process of this method are relatively fine, but the amount of data studied is small, which still needs further selection, which is not enough to prove the effectiveness of classification, and the classification efficiency is low. BenSaid and Alimi (2020) designed an online feature selection method for big data classification based on multi-objective automatic negotiation. MOANOFS uses two decision levels. In the first level, decide which learners are trustworthy. These selected k-level learners will participate in the second level learning, build an OFS method based on multilateral automatic negotiation, and achieve big data classification. This method is fast in classification and can effectively divide different data in the same set, but the division error is large, that is, the classification accuracy is low. Gong and Zhang (2021) proposed a data classification method to improve the difference information tree. Design big data through depth attribute weighted Bayesian algorithm, and vaguely distinguish big data. According to the DIT algorithm, the fuzzy data set is simplified, and the shuffle algorithm is introduced to construct a fuzzy classifier, which inputs the sample data to complete the effective classification of the data. The classifier designed by this method can quickly realise the rapid classification of large quantities of data, but the classification process is prone to the problem of repeated classification of large data, which requires more processing of repeated data to improve classification effect.

This paper designs a network big data classification method based on decision tree algorithm in order to improve the accuracy and efficiency of network big data classification and solve the problems of the above methods based on above classification algorithm design. This method mainly introduces the decision tree algorithm, and combines other algorithms to optimise the classification method to achieve classification of network big data. The main research process:

Step 1 Network big data collection. With the help of the crawler manager, the network big data is collected circularly within the set range, and the collection threshold is set to randomly generate the crawler signature, constantly update the data, and realise the collection of network big data.

Step 2 Feature extractions and pre-processing of network big data. By constructing the directed graph of network big data, automatic selection of key feature attributes in the model, extract the linear and nonlinear features, and extract the interference factors of characteristic data by transforming the big data format, complete the feature extraction and pre-processing of network big data.

Step 3 Classification method based on decision tree algorithm. Build a network big data classification decision tree, calculate the expected gain and information gain of data, determine the node attributes of big data based on the optimal data of gain results, set recursive call rules and classification termination conditions, and complete the design of classification algorithm.

Step 4 Experimental analysis. Design the experimental process, construct the classified data set, clarify the experimental indicators, and complete the experimental analysis.

Step 5 Conclusion. Summarise the full text and clarify the research results.

2 Network big data collection

The data generated in the internet is the data with the same characteristics formed by the connection between different network nodes under a certain drive. Nodes have a certain impact on the data, and the related impact between nodes also has a certain influence in this process. The data information generated in the network, such as personal information, social information, economic information, etc. have different attributes and characteristics that is network big data. Therefore, first extract the characteristics of network big data (Yuliana and Chang, 2020). Before extracting big data features, network big data needs to be obtained. This paper uses web crawlers to randomly obtain data; web crawler refers to the technology that automatically grabs information according to certain rules. In the process of web crawler, we need to collect the web node page data and its expanded node page network data. In the collection of network node page data, only account verification, page analysis and content acquisition are required. It is necessary to deeply optimise the data collection method in the expanded node page network data extraction, which is also one of the focuses of this paper. In the process of getting big data, web crawlers mainly use the crawl controller. When it determines the task of big data collection, it uses its controller to collect network big data. The specific collection process is shown in Figure 1.

Figure 1 Schematic diagram of network big data crawler collection process

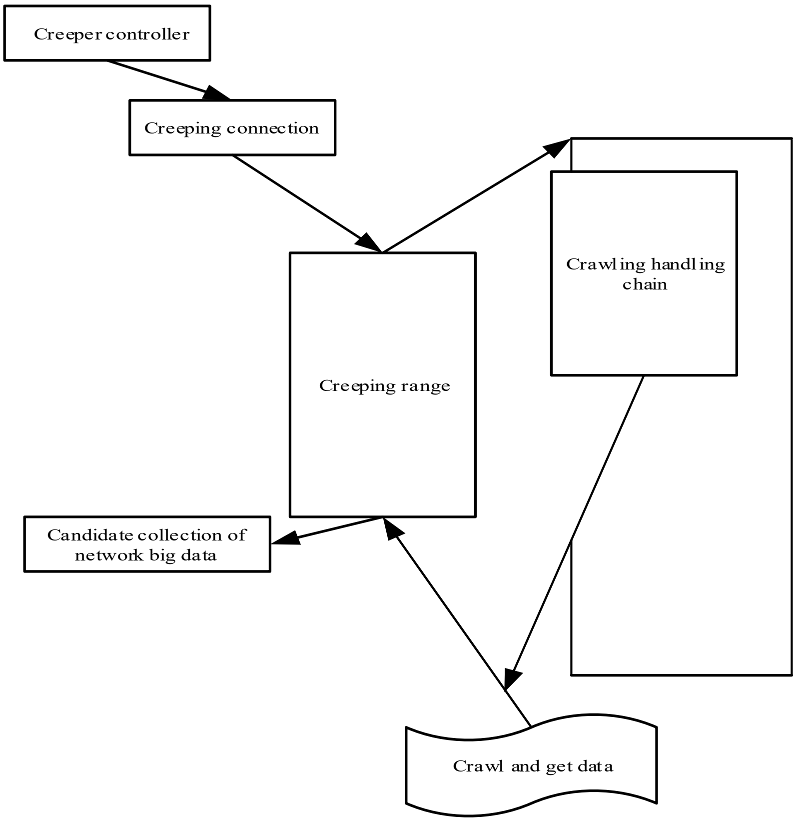
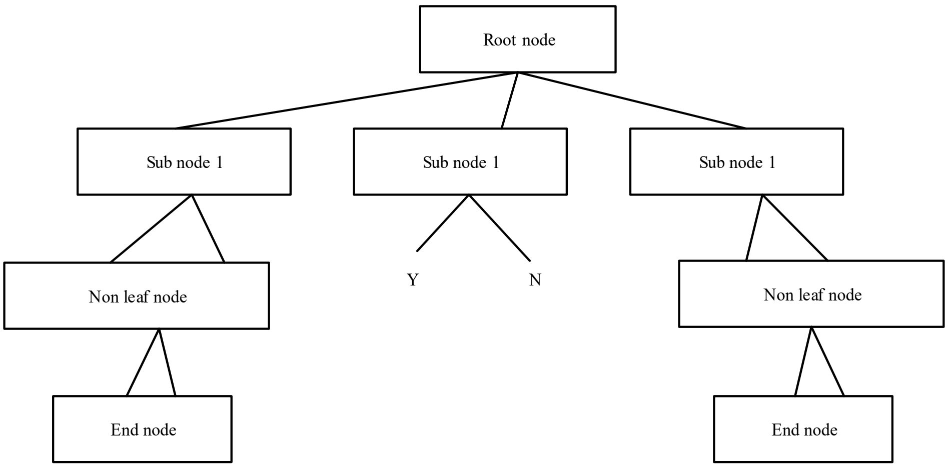


Figure 2 Schematic diagram of decision tree model



The crawler controller determines the boundary connection of various data within the set range in the process of collecting network big data, effectively crawls them, and places the crawled big data in the key data storage set, which is regarded as the network big data set studied in this paper. Crawler technology can continuously collect data in this range, and automatically update it to the network big data set after finding new data. In the process of crawling, set the threshold of network big data collection, randomly generate the signature of network big data crawler, and constantly update the newly generated data into the big data set. The network big data set obtained after crawling is:

$$V_i = \{A; e, v_1, v_2, \dots v_t; H_{v1}, H_{v2}, \dots H_{vt}\} \tag{1}$$

Among them, V_i represents the candidate set of network big data after crawling, and the data in this set is constantly updated automatically; A represents the automatic signature in the process of crawler data, e represents the set crawler threshold, v_1, v_1, …, v_t represents the initial crawler network big data set, and H_{v1}, H_{v2}, … H_{vt} represents the network big data set after continuous iteration and update.

The manager circularly collects the network big data within the set range with the help of the crawler technology in the network big data collection, the collection threshold is seted, randomly generates the network big data crawler signature, constantly updates the collected data, and realises the collection of key data set.

3 Feature extraction and pre-processing of network big data

Network big data is increasingly diversified according to the protocols and services it supports, which also leads to the continuous emergence of new attributes and characteristics of network big data (Shen et al., 2021; Fei et al., 2020). Therefore, based on the above collection of network big data, in order to improve the accuracy of classification algorithm, it is necessary to extract the characteristics of network big data and pre-process (Shao, 2020). The format and semantic information of big data is the key elements that affect the formation of its characteristics. Therefore, in this feature extraction, we first use K association graph to design the directed graph of network big data, and then extract the key features according to the directed graph, a digraph is an ordered triplet.

Randomly select some big data from the network big data candidate set as the training data set, namely:

$$X_t = \{(x_1, c_1), \dots (x_n, c_n)\} \tag{2}$$

Among them, X_t represents the training set in the directed graph, x_i and c_i represent arbitrarily selected big data, and n represents the total amount of selected network big data.

To get the distance between the data, and according to the measured distance, construct the directed graph of any network big data with the nearest K nodes. Formula is:

$$\text{DIS} = \sqrt{(x_i - x_j)^2 + (c_i - c_j)^2 + (x_{in} - x_{jn})^2} \tag{3}$$

Among them, DIS represents the distance value between data.

Combined with the above contents, a directed graph model of network big data is constructed:

$$R_i = \text{DIS} \sum X_t \sum_{i=1}^{t} (x_{in} - x_{jn})^2 (a_i - c_j)^2 \tag{4}$$

Among them, R_i represents the description of directed graph model.

According to the designed directed graph, the key feature attributes in the model are automatically selected to improve the effectiveness of feature extraction. Network big data is composed of many field attributes with different meanings. The data fields of any attribute feature are awakened by the nodes generated by them in the changes of value size, length and so on. Therefore, in feature extraction, affected by the structural characteristics of network big data, the extracted data features have a certain randomness. Suppose the given big data feature set is:

$$y_t = \{y_1, y_2, \dots y_n\} \tag{5}$$

All data attribute features in formula (5) are known. Assuming that there is an attribute feature whose feasible solution is β_i, the network data set is linearly separable, that is, at this time, the feature of the network data is a linear feature, and its extraction formula is:

$$h_i = c \sum_{i=1}^{n} s_i + \sum_{i=1}^{m} \|\beta_i\| \tag{6}$$

Among them, h_i represents the linear characteristic result of network big data, c represents the linear characteristic coefficient, and s_i represents the key points of data characteristics.

There are also many nonlinear feature data in its data after extracting the linear features of network big data. Therefore, it is necessary to further extract the nonlinear characteristics of the extractor. Assuming that there is a feature with a feasible solution of β_j, the network data set is nonlinear. At this time, the nonlinear feature is determined according to the judgement, that is:

$$g_i = -\sum_{i=1}^{n} s_i + \sum_{i=1}^{m} \|\beta_j\| \tag{7}$$

Among them, g_i represents the nonlinear characteristic result of network big data.

After extracting the linear and nonlinear features of network big data, there are some differences in the data protocols of network data, resulting in more interference data in the features. Therefore, interfering data needs to be handled. Firstly, the interference data set of the feature sample set is extracted, namely:

$$\text{sam} = \{\text{b}_\text{i} = [0, 1, \dots 1], \text{b}_\text{m} = [1, 0, \dots 0]\} \tag{8}$$

Among them, sam represents the interference dataset in the feature sample set, and b_i represents the interference data. The format of the data in formula (8) is transformed to lay the foundation for the subsequent removal of interference factors. The transformed characteristic data is expressed as:

$$\text{F} = \{\text{b}_\text{i} = [\text{b}_1^1, \text{b}_2^1, \dots \text{b}_{\text{n}1}^1], \text{b}_\text{m} = [\text{b}_1^\text{m}, \text{b}_2^\text{m}, \dots \text{b}_{\text{nm}}^\text{m}]\} \tag{9}$$

Among them, F represents the converted feature data set, and b_{nm}^m represents the converted byte vector.

Under the influence of a large number of interference factors, the interference factors of all target data are extracted separately, namely:

$$G_t = \{\text{b}_\text{i}, y_j\}_{j-1}^{n} \tag{10}$$

Among them, G_t represents the target domain data of characteristic interference factors, and y_j represents interference labels.

Accordingly, a series of processing is carried out on the determined interference data to improve the accuracy of feature extraction.

In the feature extraction and pre-processing of network big data (Xia et al., 2021), the key feature attributes in this model are automatically selected by constructing the directed graph, the linear and nonlinear features are extracted, and the interference factors of characteristic data are extracted by transforming the big data format to complete the feature and pre-processing of network big data. The difference between this method and the traditional method is that it uses the directed graph to extract the network big data features. On this basis, it further clarifies the nonlinear features and linear feature data, and deeply removes interference factors. This step is to prepare for the introduction of decision tree algorithm to classify network big data and provide more accurate data for classification methods.

4 Classification method based on decision tree algorithm

Decision tree is a commonly used classification algorithm (Javed et al., 2020; Amgad et al., 2021), which presents classifiers in tree form, so it is called decision tree. This method can realise the rapid classification of research objectives through a variety of processing algorithms for different parts of the tree (Nancy et al., 2020), it has many advantages, such as being able to deal with multiple output problems, simple algorithm and high effect. The algorithm consists of nodes and branches. It is mainly divided into leaf nodes and non-leaf nodes. The attribute of the research object corresponding to each leaf node is a kind of attribute. Due to the advantages of this node, this paper selects this algorithm to realise the classification of network data. The basic tree form of decision tree classification is shown in Figure 2.

The construction process of decision tree is a top-down process. The growth process of tree is the process of continuously subdividing network big data, and each subdivision produces corresponding data set nodes. Start from the root node, classify network big data from top to bottom until the generation meets the specific classification conditions, and the classification is then terminated and completed. The specific decision tree classification network big data process is shown below.

Step 1 In the process of decision tree classification, first take its information gain as the standard for continuous subdivision of the tree (Amgad et al., 2021). At this time, it is necessary to take the attribute network big data with the largest gain value as the best classification point, which can reduce the loss of entropy and reduce the amount of information when the classified network big data is divided again. Setting z contains n network big data templates, and the sample contains a number of different categories of characteristic data D_i. Set z_i as the sample data in D_i network big data feature, and the information gain result of this data is:

$$\tau_i = -\sum_{i=1}^{n} f_i \log u_i \tag{11}$$

Among them, τ_i represents the information gain result of the sample data in the network big data feature, f_i represents the probability of the sample data in the D_i network big data feature, and u_i represents the proportional value of the information gain.

Step 2 Set the attribute set E to contain n different values $e_1, e_2, \dots e_n$. at this time, E can divide the set z into n subsets, and set s_i to represent the attribute values in the big data set. At this time, the expected information gain of the network big data is:

$$INFO(E, S) = -\sum_{i=1}^{n} \frac{|s_i|}{|S|} \times INFO(z) \tag{12}$$

Among them, $INFO(E, S)$ represents the expected information gain result of network big data.

Step 3 Set decision tree classification network big data rules (Nancy et al., 2020). After determining the information gain and expected gain of network big data to be classified, set the decision rule set. Set a tree node as W. if the node is not in Z, the data is marked as non-leaf node data. If the node is in it, it is marked as leaf node data, that is:

$$\begin{cases} IF\ W \nexists Z, W \to P_{non} \\ IF\ W \exists Z, W \to P_{exit} \end{cases} \tag{13}$$

Among them, P_{non} represents non-leaf node data, and P_{exit} represents leaf node data.

A decision tree to be classified will be formed at this time after all the network big data are marked. Most of the root nodes in the tree are leaf nodes, and the data with the greatest gain will be selected as the classification standard.

Step 4 Set the value of network big data in the decision tree to be classified as:

$$a_i = |i = 1| 1, 2 \dots n \tag{14}$$

Among them, a_i represents the attribute value of network big data in the classification decision tree.

Then the network big data with this attribute is divided into multiple subsets to obtain:

$$z_i = |i = 1| 1, 2 \dots m \tag{15}$$

Among them, z_i represents the subset of network big data segmentation, and m represents the m subset of network big data segmentation.

Step 5 Set the recursive call conditions of network big data classification to ensure that all data in the decision tree has been classified, and:

$$Q = \text{decision\_tree}(a_1, z_1), \text{tree}(a_2, z_2), \dots \text{tree}(a_i, z_i) \tag{16}$$

Among them, Q represents the recursive call rule.

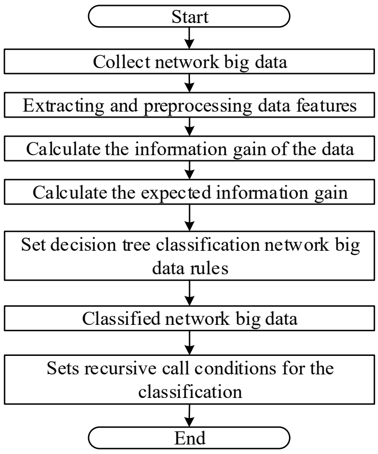
The classification termination conditions is seted on this basis, the final classification of network big data is completed, and get:

$$\sigma(a_i, z_i) = -Q\sum_{i=1}^{n} \frac{|a_i|}{A} \log \frac{|z_i|}{Z} \tag{17}$$

Among them, $\sigma(a_i, z_i)$ represents the classification result of network big data, and A and Z represent the classification termination coefficient respectively.

Based on this, the design of network big data classification method based on decision tree algorithm can be completed according to the above steps. The specific implementation process of this method is shown in Figure 3.

Figure 3 Implementation process of network big data classification method based on decision tree algorithm



The proposed method is superior to traditional methods, because the network big data classification decision tree is constructed, and the recursive call rule link is added to ensure the classification of all network big data in the decision tree. Before classification, the crawler manager and digraph are applied, and the collected network big data are optimised combined with this algorithm, it can provide more accurate data for the classification of decision tree, in addition, the decision tree classification method constructed has less time complexity, is insensitive to missing values, and can handle big data with irrelevant features, Therefore, the network big data classification method of decision tree algorithm has higher classification accuracy and efficiency.

5 Experimental analysis

5.1 Experimental scheme design

Design experiments to verify the performance of network big data classification method based on decision tree algorithm. The experimental scheme is designed as follows:

First, experiment preparation. In the stage of experiment preparation, it is necessary to determine the research object and build a training sample data set;

Secondly, the experimental indicators are set. According to the actual needs, with the goal of effectively reflecting the performance of the classification method, set the experimental indicators and the experimental indicators are classification accuracy and classification efficiency.

Finally, the results are analysed. After the completion of the experimental preparation, statistics of experimental results, analysis of experimental indicators, verify the feasibility and effectiveness of the classification method.

The experiment verifies the performance of classification methods by means of comparative analysis, which adopts the methods of this paper, BenSaid and Alimi (2020) and Gong and Zhang (2021).

5.2 Experimental preparation

Based on MySQL database, select 5,000 pieces of basic data of network big data for network big data classification. Training sample data. For each training data set, cross validation was carried out, and the distance between the data was measured using Euclidean distance, identify sample data. The training sample dataset is as follows:

Table 1 Training sample data set content

Data set serial number	*Data volume/piece*	*Number of attributes*	*Number of categories*
1	1,500	5	3
2	500	62	15
3	1,300	25	52
4	700	14	14
5	1,000	52	23

5.3 Experimental index setting

The experiment adopts the comparison method; the BenSaid and Alimi (2020) method and the Gong and Zhang (2021) method according to the above training sample data set, and takes the accuracy and efficiency of classification in the sample training data as the experimental indicators for analysis. Among them, the accuracy calculation formula of classification is:

$$\text{right} = r_i / r_{all} \times 100\% \tag{18}$$

Among them, right represents the accuracy of classification, r_i represents the number of big data in the sample network, r_{all} represents the total amount of all classified data.

The time cost reflects the efficiency of classification methods. The method with the least time cost represents the method with higher efficiency. The higher the accuracy of classification, the better the accuracy and feasibility of the validation method.

5.4 Result analysis

First, verify the accuracy of analysis and classification, and compare the accuracy of classifying sample training data with this method, BenSaid and Alimi (2020) method and Gong and Zhang (2021) method. The experimental results are shown in Figure 4.

We can see that the change of sample data size affects the accuracy of the three methods by analysing the results in Figure 4. Among them, the BenSaid and Alimi (2020) method and Gong and Zhang (2021) method show a downward trend in classification accuracy as the amount of classification data increases, so they cannot accurately classify network big data. However, the classification accuracy curve of this method fluctuates slightly, and the classification accuracy has no downward trend, and the classification accuracy remains at about 90%, when the data volume is 3,000, the classification readiness of the literature method is 57.3% and 29.5% respectively. Compared with the three methods, the method in this paper has improved by 32.7% and 60.5%. Therefore, this method is a classification method with high feasibility and effectiveness.

This paper further analyses the classification time cost of sample network big data classified by this method on the basis of ensuring the accuracy of classification, the BenSaid and Alimi (2020) method and the Gong and Zhang (2021) method. The results are shown in Table 2.

Figure 4 Accuracy result analysis of sample training data classification

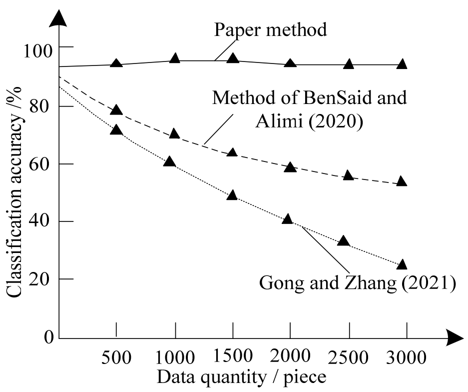


Table 2 Comparison results of classification time cost of sample network big data (s)

Data volume/piece	*Method in this paper*	*BenSaid and Alimi (2020) method*	*Gong and Zhang (2021) method*
1,000	5.3	6.8	6.9
2,000	5.6	7.5	7.9
3,000	5.5	8.2	8.5
4,000	5.4	8.5	8.9
5,000	5.5	8.7	9.3

By analysing the above data show that the classification time cost of the sample network big data classified by the method in this paper, the BenSaid and Alimi (2020) method and the Gong and Zhang (2021) method changes continuously with the change of the sample data volume. Among them, the time cost of the classification method in this paper is relatively stable, and it is always maintained at about 5S. The lowest time cost of the literature method is 6.8 s and 6.9 s respectively. Compared with the BenSaid and Alimi (2020) method and the Gong and Zhang (2021) method, the classification time cost of the method in this paper is shorter to classify the sample network big data, it is reduced by 1.8 s and 1.9 s, and the reduction range is more than 1.5 s. From this, the classification speed of the paper method is faster. Therefore, this method improves the efficiency of classification, higher application value.

6 Concluding remarks

Network big data has shown a blowout growth. How to effectively classify massive network data has become a key issue in this field. In order to improve the feasibility of classification algorithm, this paper designs a network big data classification method based on decision tree algorithm. Collect network big data with the help of crawler manager; by constructing a directed graph of network big data, the key feature attributes in the model are automatically selected to complete feature extraction and pre-processing. Build a network big data classification decision tree, calculate the expected gain and information gain of data, set recursive call rules and classification termination conditions, and complete the design of classification algorithm. The experiment proved that the proposed algorithm can improve the accuracy and efficiency of data classification.

References

Amgad, M., Atteya, L., Hussein, H. et al. (2021) 'Explainable nucleus classification using decision tree approximation of learned embeddings', *Bioinformatics (Oxford, England)*, Vol. 38, No. 2, pp.213–219.

BenSaid, F. and Alimi, A.M. (2020) 'Online feature selection system for big data classification based on multi-objective automated negotiation', *Pattern Recognition*, Vol. 110, No. 1, pp.107629–107640.

Cgab, C., Jhab, C. and Hogab, C. (2020) 'Tectonic discrimination and application based on convolution neural network and incomplete big data', *Journal of Geochemical Exploration*, Vol. 22, No. 23, pp.358–365.

Fei, W., Quan, W., Nie, F.P. et al. (2020) 'A linear multivariate binary decision tree classifier based on K-means splitting', *Pattern Recognition*, Vol. 107, No. 29, pp.107521–107521.

Gong, J-H. and Zhang, Y-J. (2021) 'Efficient big data classification method using deep AWB algorithm and improved DIT', *Computer Engineering and Design*, Vol. 42, No. 2, pp.468–474.

Javed, A., Malik, K.M., Irtaza, A. et al. (2020) 'A decision tree framework for shot classification of field sports videos', *Journal of Supercomputing*, Vol. 76, No. 9, pp.7242–7267.

Nancy, P., Muthurajkumar, S. and Ganapathy, S. (2020) 'Intrusion detection using dynamic feature selection and fuzzy temporal decision tree classification for wireless sensor networks', *IET Communications*, Vol. 14, No. 5, pp.888–895.

Shao, C. (2020) 'Data classification by quantum radial-basis-function networks', *Physical Review A*, Vol. 102, No. 27, pp.234–239.

Shen, H., Zhang, M., Wang, H. et al. (2021) 'A cloud-aided privacy-preserving multi-dimensional data comparison protocol', *Information Sciences*, Vol. 545, No. 1, pp.739–752.

Sleeman, W.C. and Krawczyk, B. (2021) 'Multi-class imbalanced big data classification on Spark', *Knowledge-Based Systems*, 5 January, Vol. 212, No. 17, pp.106598.1–106598.15.

Tsai, T.H. and Lee, Y.C.A. (2020) 'Light-weight neural network for wafer map classification based on data augmentation', *IEEE Transactions on Semiconductor Manufacturing*, Vol. 36, No. 99, p.1.

Wang, X., Du, J., Zou, R.C. et al. (2020) 'Key node identification of wireless sensor networks based on cascade failure', *Modern Physics Letters, B. Condensed Matter Physics, Statistical Physics, Applied Physics*, Vol. 34, No. 34, pp.2050394–2050410.

Xia, Y., Li, W., Zhuang, Q. et al. (2021) 'Quantum-enhanced data classification with a variational entangled sensor network', *Physical Review X*, Vol. 11, No. 2, pp.967–974.

Yang, H., He, H., Zhang, W. et al. (2022) 'MTGK: multi-source cross-network node classification via transferable graph knowledge', *Information Sciences*, Vol. 589, No. 1, pp.359–415.

Yuliana, O.Y. and Chang, C.H. (2020) 'DCADE: divide and conquer alignment with dynamic encoding for full page data extraction', *Applied Intelligence*, Vol. 41, No. 10, pp.271–295.