Null Value Estimation of Uncertainty Database Based on Artificial Intelligence

Shuang-cheng Jia(✉) and Feng-ping Yang

Alibaba Network Technology Co., Ltd., Beijing 100102, China
xindine30@163.com

Abstract. Due to the complexity of the objective world, information loss and uncertainty are common. As a tool to express the real world, database uses null values to express the problem of information missing. Aiming at the problem of null value in uncertain database, an artificial intelligence based null value estimation algorithm is proposed. Firstly, the characteristics of uncertain database are analyzed, then the lost information retrieval model is constructed, and the empty value estimation of database is completed by feature selection and data transformation, artificial intelligence clustering, influence degree calculation, empty value step estimation and other methods. Finally, it analyses the time complexity of the algorithm, and improves the problem of poor evaluation effect of traditional algorithms. Supported by experimental data and environment, the results show that the proposed algorithm has higher accuracy than the traditional algorithm. It shows that this algorithm can effectively estimate the null value in the uncertain database, and has high practical application value, and can provide theoretical reference value for related research.

Keywords: Artificial intelligence · Uncertainty · Database · Null value estimation · Time complexity

1 Introduction

In recent years, with the vigorous promotion of the rapid development of computer technology and network information technology, various information systems have been growing, carrying more and more data storage and collection tasks. Especially in the big data environment, with the rapid growth of business data of organizations, all kinds of data are generated and processed at an unprecedented speed [1]. Therefore, how to mine and extract effective information from accumulated massive data has become a hot issue in academic circles.

With the maturity of relational database theory model, various relational database systems are widely used in various fields of social life, especially in the field of data mining. However, data in real databases often contain noise, default and ambiguity, which will affect the validity of data mining. Therefore, how to accurately estimate the null value in the process of data preprocessing is an important research topic [2]. In the face of this problem, there are usually several ways to deal with it: (1) Discarding records with null values; (2) Replace null value with a constant value; (3) Take an average value instead of the null value in the range of null value; (4) In the range of null

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 283–294, 2021.
https://doi.org/10.1007/978-3-030-67871-5_26

value, a random value is used instead of null value. (5) Statistical distribution function of the original data, and then according to the distribution function to generate the replacement value of null value. However, the above methods can not deal with all the null value problems perfectly, the calculation process is complex, and the tendency of original data clustering is neglected, and the null value estimation effect can not be given very well [3, 4].

To this end, relevant personnel have proposed some database null value estimation methods, Reference [5] proposes a general boundary value estimation method for uncertain data model indicators. According to the characteristics of uncertain transaction databases with weights, a general boundary value estimation framework for commonly used model indicators is first designed, and then a quick estimation method for the upper bounds of model indicators under this framework is presented. Finally, the upper bounds of two typical model indicators are estimated to illustrate their feasibility. Experimental results show that although this method can realize the estimation of data null values, but the estimation effect is not good when facing the same local search time and the same risk measurement time. Reference [6] proposes an effective method for estimating the null value in relational databases. The method firstly mines the data in the data table to find the attribute set associated with the estimated attribute. This process only uses the data provided by the data itself. Information, avoiding the error caused by subjectivity when the expert determines the conditional attributes. Secondly, fuzzy clustering is performed according to the obtained attribute set to obtain a division of the original data, and then an estimated empty value in the relationship table is given based on the scored cluster and linear regression The method. Finally, the average absolute error rate is used to measure the accuracy of the algorithm estimation. The experimental results show that the result of this method has a high accuracy rate, but there is a problem of poor effect.

To solve the above problems, an artificial intelligence-based null value estimation algorithm for uncertain databases is proposed. This method introduces the principle of error to determine the order of estimating null values for each column. Through data mining, the attribute set associated with the estimated attributes is found. The original data is divided into fuzzy clusters by artificial intelligence method, and the null value is estimated by linear regression method within each cluster.

2 Characteristic Analysis of Uncertain Database

Uncertainty data is a general term for data that does not have complete confidence in data model. In reality, data is deterministic. The reason for producing uncertain data is due to its own knowledge limitation, which leads to the existence of uncertain data in data model. As a result, the following factors will lead to uncertain data:

(1) When describing the real world in the data model;
(2) When modifying or transforming data in the data model;
(3) When manipulating the data in the data model.

The term “uncertain data” is used to represent data that do not have full confidence in all data models. Overall, uncertain data mainly include the following categories:

(1) Probabilistic data: those data which are judged to be true or false by a certain probability value are called probabilistic data.
(2) Inaccurate data: This kind of data is available in data models, but not very clear. For example, the data may be a range or a non-value.
(3) Fuzzy data: In a data model, such data is expressed as vague in quantity or unit.
(4) Inconsistent data: Data with different true and false attributes at different times may change over time.
(5) Ambiguity data: Some data in the data model may lead to ambiguity or ambiguity.

The purpose of database system is to provide users with the information they want, and the information they interact with is the result of operation transformation of the description of real world information [7]. That is to say, interactive information with users is the focus of database system.

Since its birth, relational database has been widely used in various fields because of its simple and clear data structure, flexibility, independence, integrity, less redundancy and convenient application. But in practical application, it also reflects some shortcomings of relational database, such as the introduction of null value to solve the problem of uncertain data processing in the real world [8].

A null value in a relational database represents an unknown value, which is neither a number 0, nor an empty string, nor any other meaningful value. In the early database, there was no concept of null value. All values were determined and knowable. With null value, we can express data that we don't know, undefined data and, of course, human error data. The introduction of null value makes the data representation of relational database more complete, and to some extent, it deals with the uncertainty problem [9].

3 Building Lost Data Retrieval Model

Building the lost data retrieval model is mainly to describe the simulation and abstract process of lost data retrieval. First of all, TIR technology is used to obtain database information, and the retrieval target is set to obtain information closely related to keywords in a certain period of time.

In order to better achieve the needs of lost information retrieval, the main objectives of lost data retrieval model are to define lost data retrieval, define retrieval results, calculate the relevance of retrieval results, etc. According to the characteristics of database, the lost data retrieval model is defined as four tuple form, which is represented by $[Q, D, R, S]$, where, Q represents query information; D represents data model; R represents lost data retrieval; S represents the scoring mechanism of query information and retrieval results.

With the development of network technology, data is mainly stored in the database, but for the data, it has the time attribute. In order to better express the time attribute of the data, the data in the database is represented by the temporal data graph.

The temporal data graph is represented by $G = (V_t, E_t)$, where, V_t represents the set of temporal nodes and E_t represents the set of temporal edges.

The temporal node is v_t, expressed as $v_t = [v, (ts_{vt}, te_{vt})]$, where, v represents the identification of the temporal node, $[ts_{vt}, te_{vt}]$ represents the semi open time interval, and E is the effective time of the data.

Temporal edge e_t is expressed as $e_t = [u_t, v_t, (ts', te')]$; $[ts', te']$ is the retrieval effective time.

The temporal data figure is shown in Fig. 1. As shown in Fig. 1, information has effective time and transaction time. In the process of lost data retrieval, the effective time of information is mainly considered.

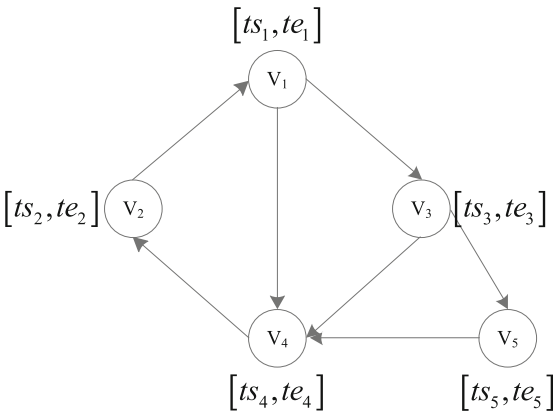

Fig. 1. Temporal data graph.

Based on the lost data retrieval model, according to the time constraints and key words in the query information, the result subtree is retrieved in the temporal data graph. In general, in the case of missing data query, we will get a lot of query result subtrees. In order to get the most similar missing data information, we sort the query result subtrees by similarity calculation, and the first one is similar missing data retrieval subtrees. Therefore, it is very important to calculate the similarity of lost data information.

In general, the smaller the temporal edge weight is, the greater the similarity is. The specific calculation method of temporal edge weight is as follows.

In order to ensure that as many keyword nodes are retrieved as possible, and ensure that the retrieval results are highly related to the lost data information of the query, calculate the node structure weight. The weight of node structure represents the importance of node in temporal data graph, which is introduced into the calculation formula of edge weight, and the calculation formula of edge weight is obtained as follows:

$$W(Q, e_t) = \frac{1}{IR_{(k,u)} + IR_{(k,v)}} \times W_e(u, v) \tag{1}$$

Where, $W_e(u, v)$ represents the node structure weight.

The temporal edge has timeliness, and the weights of different temporal edges are also different. Therefore, in the process of lost data retrieval, the temporal edge finite time needs to be consistent with the lost data query time. The temporal edge weight is set according to the lost data query time, and its calculation formula is:

$$W(Q, e_t)' = 1 - \frac{|I_c \cap I_e|}{|I_c|} \tag{2}$$

Where, I_c represents the query time of lost data; I_e represents the effective time of temporal edge.

Through the above formula, the temporal edge comprehensive weight value is obtained, so as to judge the similarity of the lost data and provide data support for the following uncertainty database null value estimation.

4 Space Value Estimation Algorithms Based on Artificial Intelligence

In the actual application of database, the problem of data missing is almost unavoidable, resulting in the problem of null value. Null value estimation has become the mainstream research direction of null value processing, and a large number of database null value estimation methods emerge [10]. Most of these methods use part of the complete data in the database table as training set, learn knowledge from the training set through machine learning or some theory of soft computing, derive decision rules or models, and finally estimate the null value according to the rules or models.

There are many commonly used null value estimation algorithms, such as rough set method, cloud model method and genetic algorithm based method. These algorithms have their own advantages, but there are also some obvious shortcomings. Rough set method is mainly based on the compatibility relationship between data, which is filled by compatible tuple values. However, if a tuple is not compatible with other tuples or the attribute values corresponding to compatible tuples are missing, then an estimate can not be given. The method of cloud model is mainly based on the generation of random points near the equilibrium position by the subordinate cloud generator to fit the original distribution of data, which will cause some "randomness" of the estimated value and affect the results of the algorithm [11]. The main disadvantage of null value estimation based on genetic algorithm is that it needs to analyze natural language semantics into effective coding, and the algorithm needs a long iteration time, and has poor scalability when the amount of data is large. In order to estimate the empty value in the relational database more accurately, based on the lost data retrieval model and the lost data similarity, a method of estimating the empty value in the uncertain database based on artificial intelligence is proposed. The specific implementation process is as follows:

Step 1: Feature selection and data conversion.

(1) Feature selection, attribute reduction algorithm based on rough set is used to reduce the attributes of the original data table and get the key attribute set after reduction.
(2) Data conversion, mainly refers to data preprocessing, making it easy to use data form. Firstly, the natural language semantic attributes are numeralized, so that the attributes can be conveniently used for data mining. Then the formula of fuzzy number is used to normalize the numerical information and simplify the calculation [12].

Step 2: Artificial intelligence clustering.

The non-null attribute sets associated with the attributes with null values obtained in step 1 are used for clustering. Make similar data together, different data are divided into different clusters. Figure 2 shows the compatibility of objects in different attribute sets.

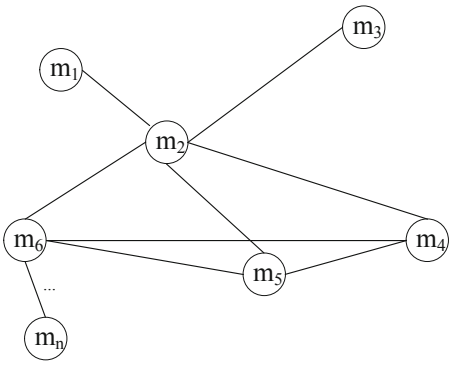

Fig. 2. Compatibility relationship among objects in different attribute sets

As shown in Fig. 2, considering that different attributes have different influence weights on columns with null values, the relevant weights are introduced:

$$w = \frac{r^2 - W(Q, e_t)'}{\sum_{k=1}^{m} r_1^2} \tag{3}$$

In the formula, m is the number of attributes in the set of non-empty attributes related to attributes with null values; r represents the correlation coefficients of attributes with null values; w is the ratio of the correlation coefficients of attributes with null values and the sum of the correlation coefficients of all related attributes and attributes with null values, which reflects the weight of the influence of attributes with null values. After artificial intelligence clustering, the clustering center is obtained.

Step 3: Calculate the impact [13, 14].

After clustering the data into several clusters, for each cluster, the influence of different independent variables on dependent variables is different. Artificial intelligence regression coefficient is used to calculate the influence of different independent

variables on dependent variables [15]. Firstly, the fuzzy correlation coefficient is used to represent the correlation degree between attributes, then the independent variable coefficient is determined, and finally the influence degree of attributes is obtained.

The formula for calculating the degree of correlation is as follows:

$$z_{a,b} = \frac{\sum_{i=1}^{n} (a_i - \overline{a}) \cdot (b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n} (a_i - \overline{a})^2 \cdot \sum_{i=1}^{n} (b_i - \overline{b})^2}} \tag{4}$$

In the formula, $\overline{a}$ and $\overline{b}$ represent the sample mean of a, b of the fuzzy set.

The formula for determining the coefficient of independent variable is as follows:

$$COD = \pm \frac{r^2}{\sum_{k=1}^{m} r_1^2} \tag{5}$$

Step 4: Estimate null values.

Firstly, the Euclidean distance between the tuple and each cluster center is calculated, and the null value estimation algorithm is used to obtain the estimated value [16].

The null value estimation algorithm of uncertain database based on artificial intelligence is described. If the number of records in the data table is N, the number of records containing null values is N_{null}. The key attribute number after attribute reduction is m and the clustering number after partition is C. Then the time complexity of the algorithm is analyzed as follows:

(1) Feature selection and data conversion. In this step, an attribute reduction algorithm based on discernible matrix in rough set is used, and the time complexity of the algorithm is $O(N^2)$.
(2) Clustering using artificial intelligence algorithm of null value estimation in uncertain database, the time complexity of the algorithm is $O(N)$;
(3) Calculate the correlation degree to obtain the overall time consumption, i.e. linear complexity [17];
(4) Estimating null values and evaluating. This step estimates a small number of null values contained in the database tables with a time complexity of $O(N_{null})$, it is a high order infinitesimal of $O(N)$, which can be neglected [18].

In summary, the algorithm of null value estimation in uncertain database of artificial intelligence has high estimation accuracy.

5 Experimental Analysis

In order to verify the rationality of null value estimation algorithm in uncertain database based on artificial intelligence, experimental verification and analysis are carried out.

5.1 Experimental Data and Environment

Experimental environment: The operating system is Windows 7, using 3.40 GHz Intel Core i7-3770 CPU, using 8G memory bars, using C+ language on Microsoft Visual Studio 2012 development platform to realize the uncertainty database null value estimation algorithm based on artificial intelligence.

The algae dataset, a classical data set in data mining, is used in the experiment. The attributes are shown in Table 1.

Table 1. Independent attributes of Algae datasets.

Variable	Range
Season	{Spring, Summer, Autumn, Winter}
Size	{Small, medium, large}
Speed	{Low, medium, high}
PH	Positive real number
CL	Positive real number
NO_3	Positive real number
NH_4	Positive real number
PO_4	Positive real number

The table describes that under the influence of independent variables, the data set contains 184 pieces of data, 164 of which are used as training data set, and the remaining 20 as test set.

5.2 Experimental Steps

Experiments are carried out on algae datasets. The specific process is as follows:

Step 1: Data conversion and feature selection: because season, size and speed are text variables, they can not be directly processed, changing season's "spring, summer, autumn, winter" to "1, 2, 3, 4"; size's "small, medium, large" to "1, 2, 3"; speed's "low, medium, high" to "1, 2, 3". Then we use feature selection algorithm to eliminate the four attributes of season, CL, NH4 and PO4. The remaining attributes represent more than 95% of the data features.

Step 2: Clustering the remaining attributes, using clustering indicators to get a better clustering effect, dividing the original data set, and finding the clustering center.

Step 3: For each type of data, calculate the influence of independent variables on dependent variables.

Step 4: Estimate the null value.

Figure 3 is the flow chart of the experimental steps.

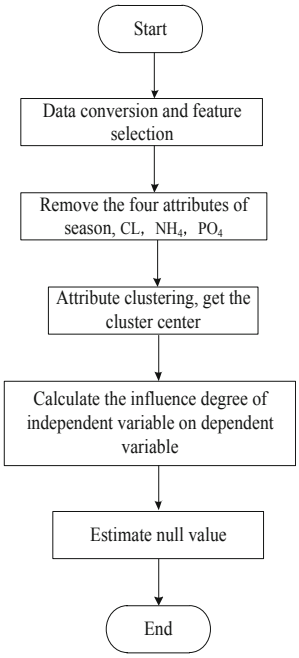

Fig. 3. Experimental flow chart

5.3 Experimental Results and Analysis

In order to study the validity of AI-based null value estimation algorithm for uncertain databases, local search time and risk measurement time are taken as criteria to compare the evaluation results of traditional references [5] algorithms and AI-based algorithms.

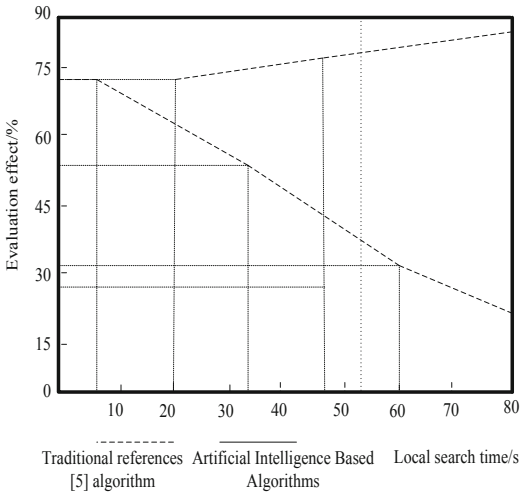

Fig. 4. Evaluating the effect of two methods using the same local search time.

(1) Local search time

When the local search time is consistent, the traditional references [5] algorithm is compared with the evaluation effect based on artificial intelligence algorithm, and the results are shown in Fig. 4.

Figure 4 shows that the AI-based algorithm is 2% better than the traditional references [5] algorithm when the local search time is 10 s; the AI-based algorithm is 8% higher than the traditional references [5] algorithm when the local search time is 20 s; the AI-based algorithm is 20% higher than the traditional references [5] algorithm when the local search time is 30 s; and the AI-based algorithm is 40 s higher than the traditional references [5] algorithm when the local search time is 40 s. Compared with the traditional references [5] algorithm, the AI-based algorithm has 32% higher evaluation effect; the AI-based algorithm has 39% higher evaluation effect when the local search time is 50 s; the AI-based algorithm has 53% higher evaluation effect when the local search time is 60 s; the AI-based algorithm has 58% higher evaluation effect when the local search time is 70 s; and the AI-based algorithm has 80 s higher evaluation effect when the local search time is 80 s. The AI-based algorithm is 60% more effective than the traditional references [5] algorithm. Therefore, under the same local search time, the AI-based algorithm is better than the traditional references [5] algorithm in evaluating the effect.

(2) Risk measurement time

When the time of risk measurement is consistent, the traditional references [5] algorithm is compared with the evaluation effect based on artificial intelligence algorithm, and the result is shown in Fig. 5.

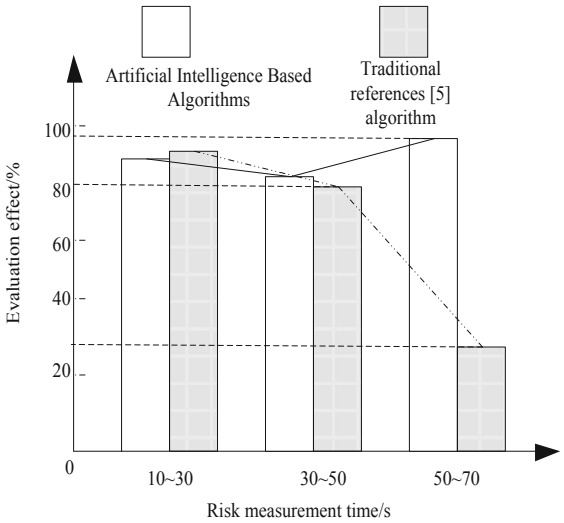

Fig. 5. The effect of two methods for evaluating the same risk measurement time.

Figure 5 shows that when the time of risk measurement is 10–30 s, the evaluation effect of traditional references [5] algorithm is 90%, and that of AI-based algorithm is 86%. When the time of risk measurement is 30–50 s, the evaluation effect of traditional references [5] algorithm is 80%, and that of AI-based algorithm is 83%. When the time of risk measurement is 50–70 s, the evaluation effect of traditional references [5] algorithm is 28%, and that of AI-based algorithm is 97%. It can be concluded that under the same risk measurement time condition, the AI-based algorithm is better than the traditional references [5] algorithm in evaluating the effect. This is because the method through the analysis of the characteristics of the uncertain database, data conversion, and the use of artificial intelligence technology for data clustering, and calculate the impact degree, and ultimately achieve the null value estimation of the database, through the above steps to enhance the effect of null value estimation.

To sum up, the null estimation effect of the proposed algorithm is better than that of the traditional references [5] algorithm under the same local search time or the same risk measurement condition, indicating that the proposed algorithm has high application value.

5.4 Experimental Conclusions

To sum up, the null value estimation algorithm of uncertain database based on artificial intelligence is effective. Under the same local search time, the maximum evaluation effect of artificial intelligence algorithm is 86%, and under the same risk measurement time, the maximum evaluation effect of artificial intelligence algorithm is 97%.

6 Concluding Remarks

Non-deterministic database is based on strict mathematical concepts. It has a single concept and simple and clear data structure. Its greatest advantage is that the relationship between entities can be expressed by relationship, that is, the indefinite database can describe itself. Many advantages make the indefinite database occupy the dominant position in the market and has been widely used.

With the gradual expansion of the application field of uncertain database, the data processing scope and ability of uncertain database are demanded in various fields. For example, in the fields of scientific computing, sensor application and knowledge learning system, the database is required to deal with uncertain data. However, most uncertain databases can only deal with accurate data, lacking a comprehensive method for dealing with uncertain data. Now the only way to solve this problem is to use the artificial intelligence algorithm for estimating the null value of uncertain databases.

This method uses artificial intelligence algorithm to classify the initial data, taking into account the fuzzy nature of data classification, and introduces weighted values according to the different dimensions of sample data to the contribution of clustering, which makes the clustering results more accurate. Then, the null value estimation model is constructed by multiple linear regression, which makes the null value estimation method more effective and accurate.

References

1. Alam, M.K., Aziz, A.A., Latif, S.A., et al.: Error-aware data clustering for in-network data reduction in wireless sensor networks. Sensors **20**(4), 1011 (2020)
2. Zhang, T.A.: Dynamic threats assessment based on intuitionistic fuzzy set under missing data condition. Fire Control Command Control **43**(8), 93–97 (2018)
3. Kim, K.: Identifying the structure of cities by clustering using a new similarity measure based on smart card data. IEEE Trans. Intell. Transp. Syst. **21**(5), 2002–2011 (2020)
4. Wang, J., Liu, F.X., Jin, C.J.: General bound estimation method for pattern measures over uncertain datasets. J. Comput. Appl. **38**(01), 165–170 (2018)
5. Liu, L., Wang, L.S., Wu, F.: An efficient method for estimating null values in relational database. Comput. Technol. Autom. **35**(03), 110–114 (2016)
6. Li, H.: RFID tag number estimation algorithm based on sequential linear Bayes method. J. Comput. Appl. **38**(11), 3287–3292 (2018)
7. Gao, J.M.: Adaptive deduplication simulation in privacy protection database. Comput. Simul. **36**(01), 239–242 (2019)
8. Song, X.P.: Global estimates of ecosystem service value and change: taking into account uncertainties in satellite-based land cover data. Ecol. Econ. **143**(1), 227–235 (2018)
9. Faghih, M., Mirzaei, M., Adamowski, J., et al.: Uncertainty estimation in flood inundation mapping: an application of non-parametric bootstrapping. River Res. Appl. **33**(4), 611–619 (2017)
10. Wang, Y., Tao, W., Yan, Z., Wei, R.: Uncertainty analysis of dynamic thermal rating based on environmental parameter estimation. EURASIP J. Wirel. Commun. Netw. **2018**(1), 1–10 (2018). https://doi.org/10.1186/s13638-018-1181-7
11. Ju, H., Zhang, G., Cui, J., et al.: A novel algorithm for pose estimation based on generalized orthogonal iteration with uncertainty-weighted measuring error of feature points. J. Mod. Opt. **65**(3), 331–341 (2018)
12. Frédérique, S., Bernard, C., Paolo, D.G.: High-resolution humidity profiles retrieved from wind profiler radar measurements. Atmos. Meas. Tech. **11**(3), 1669–1688 (2018)
13. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
14. Forsberg, E.M., Huan, T., Rinehart, D., et al.: Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat. Protoc. **13**(4), 633–651 (2018)
15. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. (2020). https://doi.org/10.1007/s40747-020-00161-4
16. Koch, D.C.L., Jean-Paul, G., Xue, M., et al.: Terahertz frequency modulated continuous wave imaging advanced data processing for art painting analysis. Opt. Express **26**(5), 5358 (2018)
17. Yang, Y.G., Guo, X.P., Xu, G., et al.: Reducing the communication complexity of quantum private database queries by subtle classical post-processing with relaxed quantum ability. Comput. Secur. **81**(3), 15–24 (2019)
18. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mob. Netw. Appl. **24**(1), 1–4 (2019)