Energy analysis of Internet of things data mining algorithm for smart green communication networks

Ziping Du

PII: S0140-3664(19)31964-4
DOI: https://doi.org/10.1016/j.comcom.2020.01.046
Reference: COMCOM 6167

To appear in: *Computer Communications*

Received date : 13 December 2019
Revised date : 9 January 2020
Accepted date : 22 January 2020

Please cite this article as: Z. Du, Energy analysis of Internet of things data mining algorithm for smart green communication networks, *Computer Communications* (2020), doi: https://doi.org/10.1016/j.comcom.2020.01.046.

Energy analysis of Internet of things data mining algorithm for smart green communication networks

Ziping Du

Information Engineering College, Suzhou Industrial Park Institute Services Outsourcing, JiangSu, Suzhou 215123, China

Corresponding author: Ziping Du (e-mail: zipingdu@126.com).

ABSTRACT

With the continuous development of Internet technology and electronic information technology, big data technology and cloud computing technology also rise and develop, and have a positive impact on people's lives. Data mining system can deeply mine the value information contained in big data, so as to assist users to solve practical problems and help users to make correct decisions and judgments. This paper presents an energy analysis of data mining algorithm based on cloud platform for Internet of things (IoT). First of all, an improved Apriori algorithm is proposed, which is based on Boolean matrix and sorting index rules. Then Boolean matrix is obtained after scanning the dataset and the Boolean matrix is preprocessed to delete the useless transactions and the item set, which are combined with sorting index to produce other item sets, effectively improving the efficiency of frequent item mining, which effectively reduce the memory usage. Secondly, the density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm needs human intervention in the global parameter selection, and the process of regional query is complex and the query is easy to lose objects. An improved parameter adaptive and regional query density clustering algorithm is proposed, which can effectively delete the redundant data in the high-level complex data space on the premise of retaining the internal nonlinear structure of the IoT data. The efficiency of clustering is also improved accordingly. Finally, the simulation based on cloud platform verifies the effectiveness and superiority of the algorithm.

Keywords: Energy analysis; Internet of Things; Cloud platform; Hadoop; Data mining algorithm

1. Introduction

With the continuous development of Internet and the intelligent terminals, the amount of data has become more and more, and the types of data have become extremely rich, which represents that human beings have been big data age [1]. In recent years, data analysis of the Internet of things (IoT) has attracted more and more [2]. Through effective data analysis methods, the characteristics and rules of the Internet of things can be mined to reduce the loss in each life cycle of the Internet of things [3]. At present, the traditional Internet of things data analysis method needs to establish a high-precision mathematical model of the IoT data system, and the system is limited to the linear and time invariant system [4]. At the same time, it has high requirements for knowledge in the professional field and low precision shortcomings, so it has great limitations in application. At present, the collection mechanism of Internet of things data in many areas is becoming more and more perfect [5]. The Internet of things database will collect Internet of things data, which contains a lot of undiscovered massages, and the Internet of things data has many disadvantages, but it is difficult for traditional Internet of things data analysis methods to find and summarize the content of these Internet of things data [6]. Knowledge discovery, machine learning and data mining technology are just new disciplines born under this background. Because data mining technology has the ability to deal with large amount of data and find potential massages, it is a way to mine important massages in IoT data [7]. Data mining method does not need too much expert domain knowledge, and effectively finds useful information between data, which has achieved good application results in many fields [8]. Therefore, research on the construction of data analysis model of the Internet of things and how to apply different data mining methods to the field of data analysis of the Internet of things, use data mining technology to conduct in-depth analysis of the Internet of things data, mainly study the clustering method and association rule method in the data mining algorithm, realize the effective analysis and evaluation of the Internet of things data, and pass the theoretical analysis through the data mining model in the way of software [9]. Because the Internet of things technology can realize the interconnection between various terminal devices in the way of wireless or wired communication, use the characteristics of IoT to transmit the Internet of things collected by IoT sensors in real time, and it uses the IoT technology to obtain the

online monitoring, early warning report and plan management of the IoT data in the mobile terminal [10]. Management and decision support and other management and service functions help administrators to manage and guide energy-saving transformation of public buildings more conveniently and accurately[11].

Because more and more people put forward the mining algorithm suitable for cloud computing according to the shortcomings of the current data mining algorithm [12]. Reference proposed an improved Apriori method based on transaction set array, which improved the time and space load [13]; Reference proposed a new pruning method of candidate set, which deleted useless candidate options and improved the mining efficiency [14]; Reference proposed to classify the value of data set into 0-n, and improved the generation of frequent 2-term set, which effectively improved the efficiency of association rule mining [15]; Reference combined the optimal features of matrix and apriori, reduced the space complexity and realized the purpose of mining frequent item set rapidly [16]; Reference proposed a SIM Apriori algorithm based on similarity measurement, which is a cosine similarity calculation method [17]; Reference proposed a Apriori method based on characteristics, each transaction item is assigned with its own characteristics to carry on With more information, experiments show that the proposed method reduces the computation time and gets more reasonable association rules [18]; Reference combined the advantages of Apriori simplicity and FP growth effectiveness, proposed a new frequent itemset PIETM algorithm [19]. The disadvantage of Apriori algorithm is that producing the suitable candidate item sets is required to scan IoT data center for multiple times, which leads to the increase of computation [21, 20]. The above association rule improvement method is mainly aimed at improving the efficiency for reducing the number of database scanning process [22]. The matrix converts the data to Boolean type, which improves the efficiency of the algorithm [23]. However, there is no preprocessing operation before the calculation, and redundant transactions and items are not deleted, which increases the calculation amount of the algorithm to a certain extent [24].

Data clustering mining algorithm is also a common form. Reference put forward an IDBSCAN method to significantly deal with big-scale spatial data when DBSCAN processes the whole database [25]. Reference proposed a fast clustering method based on density [26].

After sorting according to the coordinates of specific dimensions, he selected the orderly unmarked points outside the neighborhood of the core object as the seed extended cluster. In this way, the frequency of region query execution can be reduced, and the conversion of objects into kernel functions improves the clustering accuracy and reduces the dependence on density threshold to some extent; Reference proposes a DBSCAN algorithm based on probability distribution [27]. The improved points based on density clustering algorithm are mainly based on the idea of core object extended class, global input parameter sensitivity, sampling and layering. The selection of DBSCAN global parameters is manually determined according to k-dist curve, which is cumbersome and not practical [28]. In other methods based on mathematical statistical method, some of the internal parameters are determined by specific data distribution, which making the calculated global parameters inaccurate [29, 30].

For the above problems, this paper presents an energy analysis of data mining algorithm based on cloud platform for Internet of things (IoT). First of all, an improved Apriori algorithm is proposed, which is based on Boolean matrix and sorting index rules. Then Boolean matrix is obtained after scanning the dataset and the Boolean matrix is preprocessed to delete the useless transactions and the item set, which are combined with sorting index to produce other item sets, effectively improving the efficiency of frequent item mining, which effectively reduce the memory usage. Secondly, the DBSCAN density clustering algorithm needs human intervention in the global parameter selection, and the process of regional query is complex and the query is easy to lose objects. An improved parameter adaptive and regional query density clustering algorithm is proposed, which can effectively delete the redundant data in the high-level complex data space on the premise of retaining the internal nonlinear structure of the IoT data. Effectively improve the efficiency of clustering. Finally, the simulation based on cloud platform verifies the effectiveness and superiority of the algorithm.

2. Model design of data processing platform based on Cloud Computing

2.1 Architecture design of Internet of things data mining algorithm based on Cloud Computing

As a new technology, IoT is the inevitable result of social development [31, 32]. Architecture design of Internet of things data mining algorithm based on Cloud Computing as shown in Figure 1, and it is also a mining link in the data processing of the IoT. In the model, the parallel operation and distributed operation of mining algorithm and recommendation algorithm are fully considered. The model divides the data processing platform into three basic levels. The idea of hierarchical design makes the whole Internet of things data processing more effective, and the processing efficiency has been greatly improved. From bottom to top: cloud computing support platform layer, data mining capability layer, data mining cloud service layer.

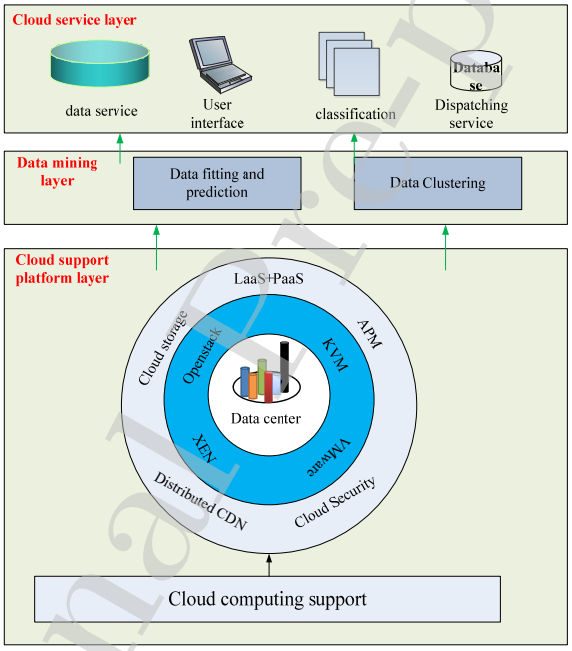


Figure 1. Architecture design of Internet of things data mining algorithm based on Cloud Computing

(1) Cloud computing support platform layer: provides file or data storage space and data computing capacity, which is the cornerstone of the data processing platform. In this platform, the third-party mining algorithm service is integrated. The business operation can be independently developed by the enterprise or provided by the third party, which is also the convenience of the IoT.

(2) Data mining capability layer: it provides the basic capability of data mining for the whole platform. In this level, there must be basic service. At the same time, it also needs to provide necessary support for the ability of data mining cloud service layer. To some extent, the level of data mining ability directly affects the service ability of cloud computing, and the service ability of the whole Internet of things will be affected.

(3) Data mining cloud service layer: it provides data mining cloud services externally, and the interface forms of its service capability encapsulation are diverse externally. Simple object access protocol can become the external interface forms of cloud services. The basic function of the Internet of things is to provide users with more convenient services by using information technology. Cloud service exists to strengthen its service ability. In fact, the cloud service layer is a kind of user demand that integrates the data processing of the next two layers. In addition, the cloud service layer can also support the access of structured query language statements, which makes the language transformation more convenient in data processing.

2.2 Structure design of real time data processing system based on spark

The operation process of real-time data processing system on account of spark framework is shown in Figure 2. The information in the IoT data set is arranged in the message queue in chronological order. In the process of processing, there are four kinds of business processing logic, one kind of spin and three kinds of bolt. The output reads the data message queue; bolt splits each original data and processes it into a standardized data processing structure. The geohash partition in the sample area is completed, and finally the aggregation operation is implemented. In the specific application, the speed of data sensing is very high, if using the traditional database processing, its efficiency is very low. Therefore, memory data is used to store and reduce data processing delay. In the process of real-time data processing, five desktop computers are distributed, and they form a cluster running environment. The Ubuntu server operating system is installed on the node, and the data is presented through the web page to achieve user interaction. Using the above algorithm design and distributed computing framework, it can also meet the requirements of real-time data processing in the current

computer equipment environment. In the process of data interaction and access, the delay of computing unit is kept at millisecond level, which can satisfy the needs of real-time data.

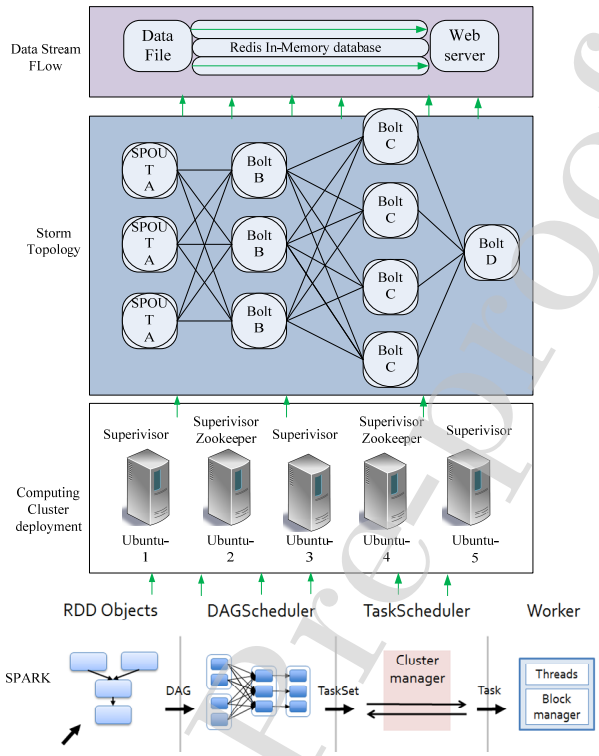


Figure 2. Data processing platform model based on Cloud Computing platform

Data mining architecture has four main functions, each of which satisfies the needs of big data mining. The first function is data preprocessing. Data preprocessing refers to the synchronous processing of historical data and data flow. Compared with traditional data mining, it has better flexibility, stronger operability and higher efficiency of data timely processing. The second function is data storage. The function of data storage mainly depends on database. The big data mining system adopts distributed storage, which can save disk space by storing data in columns. The third function is data calculation and analysis. Data computing and analysis is the core function of big data mining system. It mainly deals with data through probability dimension index. The fourth function is data display. The data display function mainly provides data display services, which can not only display text, reports, graphics, but also display animation. The visualization effect is very good, which can

help users understand the data and deepen their impression on the data. The model builds on spark is shown in figure 3.

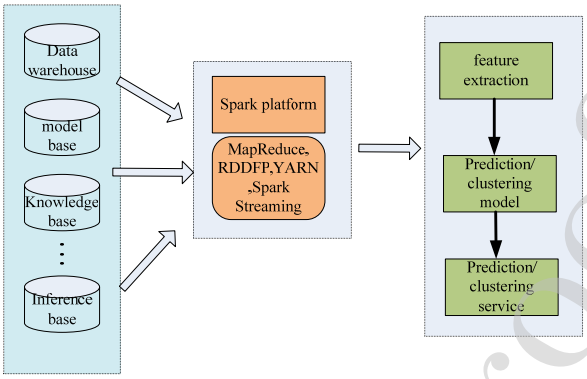


Figure 3. The model builds on spark

As a late distributed framework [33], Apache spark is not only compatible with traditional Hadoop cluster, but also has strong scalability [34]. Spark puts a large number of operations into memory for execution, which greatly improves the efficiency of operations. Spark streaming, the subsystem under spark streaming, has a unique operation mode of converting streaming data into batch processing [35, 36].

As can be seen from Figure 3, based on the IoT data mining, this paper puts the Internet of things data feature extraction, data prediction and clustering algorithm analysis, data service on the spark platform to improve the efficiency and algorithm performance. In addition, structured data, unstructured data or mixed data need to be processed through MapReduce, which is based on large-scale data mining results evaluation. Through the development of multi spark flow processing, parameter optimization and variable selection, parallel computing is realized and redundancy is reduced.

3. Implementation and Energy Analysis of Internet of Things Data Mining Algorithms on Cloud Platform

3.1 Data mining method for association rules of matrix sort index

The IoT data in the database server of the IoT supervision platform hides the use characteristics of data. In this chapter, through the data mining method of association rules, we analyze the linkage between each data item, so as to provide reference for further understanding the data mining results of the whole Internet of things. This section mainly

discusses how to effectively find the strong association rules of the entire IoT data, reduce the spatiotemporal complexity of algorithm, and improve the efficiency of association rule data mining.

Apriori algorithm is one of the typical algorithms of association rules. Apriori algorithm first traverses the database once, calculates the support of each item set data item, generates candidate item set C_1, and uses the support threshold to produce frequent item set L_1. Connect L_1 to obtain reserved set C_2, then scan the database for a second time. In this way, the algorithm search ends until frequent item set cannot be found.

When the matrix association rule algorithm directly uses the data set transformed matrix to calculate the association rules, it does not prune, which increases the unnecessary calculation. In this chapter, the transaction matrix is pruned before the calculation to produce the generation of association rules. The basic idea of data mining algorithm for matrix association rules is to scan database D and transform transaction database d into binary Boolean matrix, assuming that transaction database d contains n items and M transactions. The whole database is as shown in formula (1):

$$A_{mn} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{1}$$

Where definition of internal element $\{a_{ij}\}$ in transaction matrix A_{mn} is shown in formula (2):

$$a_{ij} = \begin{Bmatrix} 1, if & i_j \in D, i = 1,2,\cdots,m \\ 0, if & i_j \notin D, i = 1,2,\cdots,n \end{Bmatrix} \tag{2}$$

Where Array A_{mn} is matrix, where each column indicates an item and each row indicates a transaction. If transaction D_i contains item j in item set I, the value corresponding to row m and column j in the matrix is 1, $a_{ij} = 1$. $a_1, a_2, \cdots, a_m$ is called a transaction vector.

The frequent item set generated by association rules of Boolean matrix is operated according to the "and" of elements. Each set is compared with each column of elements in the

corresponding Boolean matrix after being accumulated and operated with the minimum support minsup to determine whether it is a frequent item set, as shown in formula (3).

$$\sup port(I_i, I_j) = \sum_{k=1}^{m} I_i^k I_j^k \geq \min \sup \tag{3}$$

The main steps of setting the transaction matrix and generating the frequent two item set are: querying the transaction database d, initializing the transaction matrix A, and generating the marker sequences L_r and L_c. According to the properties 1 and 2 of association rules, delete the useless transactions and items in transaction matrix A, get the pruning matrix B, then update the marking sequence of transactions and items to L_c^* and L_r^*, multiply the pruning matrix and its transpose matrix to obtain matrix s, as shown in formula (4):

$$S = B * B^T \tag{4}$$

Compare the value of the upper triangular matrix IJs of matrix s with the minimum support minsup, as shown in formula (5):

$$S_{ij} \geq \min \sup(i < j) \tag{5}$$

If formula (5) is satisfied, it is a frequent item set. Query the transactions in the tag sequence and the tags L_c^* and L_r^* of the items to get all frequent 2 item sets.

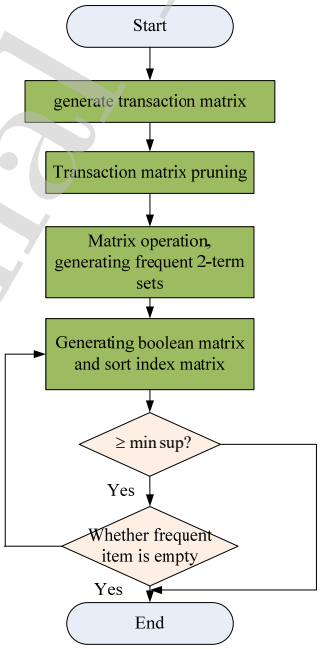


Figure 4. The flowchart of data mining method for association rules of matrix sort index

3.2 Research on adaptive fast DBSCAN density clustering method

Density clustering algorithm is a typical clustering method. Other clustering methods include partition, model and hierarchy.

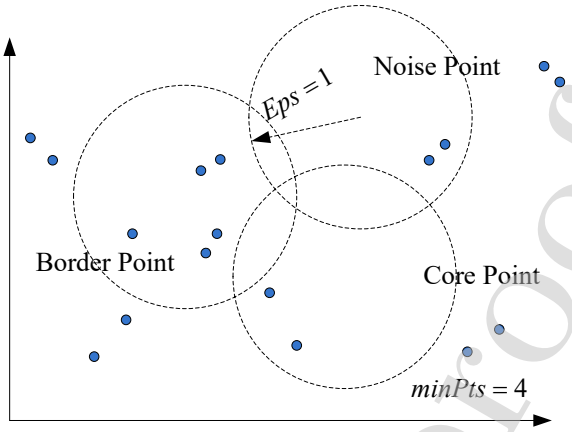


Figure 5. Schematic diagram of core point, and boundary object point in DBSCAN algorithm

The nearest neighbor of *Eps* represents the nearest neighbor within the radius of *Eps* of a given object, which is called the nearest neighbor of EPS of the object, which is expressed as $NEps(p)$:

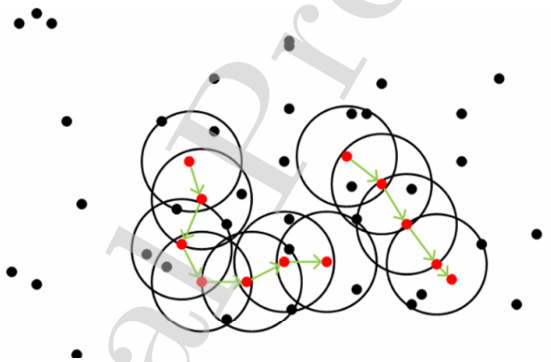
$$NEps(p) = \{q \in D \mid dist(p,q) \le Eps\} \tag{6}$$

Direct density reachability refers to that for *Minpts* and *Eps* with given global parameters, the direct density from object q can reach object p. The following conditions need to be met:

$$p \in NEps(q), \left|NEps(q)\right| \ge Minpts \tag{7}$$

Schematic diagram of Noise point, core point, and boundary object point class data point is shown in Figure 5. According to *Eps* scanning radius and minimum *Minpts*, when included points in *Eps* radius of data point is greater than *Minpts*, the data point is the core point, such as p_1 point; when included points in *Eps* radius of data point is not greater than minpts, the data point is the boundary object point, such as p_2 point; if the data point does not meet both p_1 point and p_2 point, the data point is noise. The detailed figure is shown in figure 5.

Because the selection of *Eps* and *Minpts* in DBSCAN algorithm depends on human experiences, after the data with uniform density distribution are arranged in ascending order according to k-dist curve, the points with sharp increase of curve change amplitude are selected as *Eps* parameters artificially, and the parameter of *Minpts* is determined as fixed constant 4, the implementation process is tedious and depends on human intervention. In this section, a global parameter self-adjusting selection method is presented. According to the statistical distribution characteristics of the distance space of the data, the distribution of k-dist value is calculated, and then the curve fitting is carried out to fit the distribution curve. By calculating the corresponding value at the inflection factor of the fitting profile, the *Eps* parameters are adaptively determined, and the distribution of points in the *Eps* field of each point in the data is calculated. The implmented process of *Eps* in proposed algorithm is represented in figure 6.



AF-DBSCAN

Figure 6. The adjustment process of parameters *Eps* and *Minpts* in AF-DBSCAN

Due to the single density measurement index, the data set mainly aims at the data with insignificant cluster density difference, that is, the data with uniform density distribution. Calculate the distance distribution matrix $DIST_{n\times n}$ according to the input data set D.

$$DIST_{n\times n} = \{dist(i,j) \mid 1 < i < n, 1 < j < n\} \tag{8}$$

In this section, the density clustering algorithm is conneted to the statistical model. According to the mathematical statistics methods, the data set is assumed to be generated by

the statistical process. Futhermore the optimal global variables *Eps* and *Minpts* are calculated adaptively.

The selection method of representative objects proposed in this paper is as follows: draw a circle with the core object P as the center, *Eps* as the radius, draw a circle with the object P as the origin, draw a circle at points A, Y, P and Q, then drawing two diameters with angles of 45° and 135° with the x-axis at points B, C, G and M. In the first scene of selecting target objects, the closest target points to point A, Y, P and Q are selected as representative objects in the *Eps* region of P, with points a, C, E and G as references. When the closest point is selected, this point can only be selected once and belongs to the representative object of first target point.

In this way, the number of selected representative objects for any one object in 2D space is at most 8. Generally speaking, there are $3^n - 1$ target points and 2n quadrants. So the maximum number of seeds selected is $3^n - 1$. According to the above methods, regional query is implemented, which effectively improves efficiency and solves troubles f object loss.

3.3 Framework design of data prediction and data clusteringalgorithm

The implementation process of the unified integration algorithm is to integrate the above-mentioned improved imsia algorithm and af-dbscan algorithm into the system and integrate the existing two more mature data mining methods Apriori method and DBSCAN method into the system and compare them with the two improved methods to verify the correctness. On this basis, two kinds of data mining models, association rules and density clustering, are formed. The Framework design of data prediction and data clusteringalgorithm based on cloud computing platform is shown in figure 7.

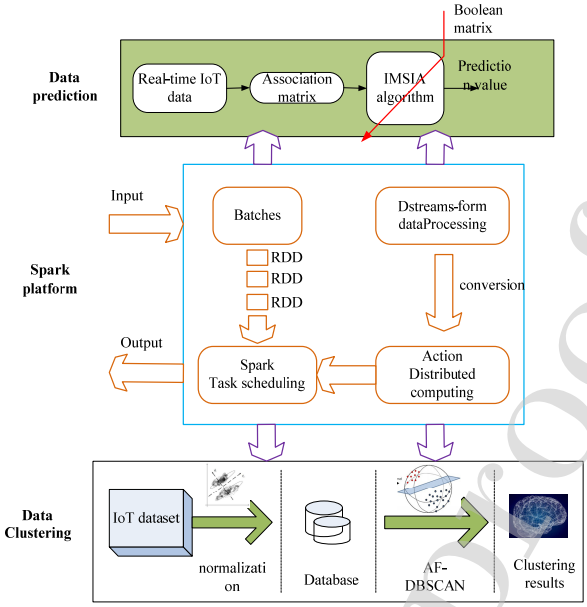


Figure 7. Comparative analysis of three clustering algorithms

As shown in Figure 7, the bottom layer of the system is the data source, mainly including the energy consumption data in the front view, the real-time data in the Internet of things system, as well as the related personnel data, scientific research data, etc. The first layer is the data storage layer, which is divided into two layers: one is the basic data from other systems; the other is the intermediate data of the data mining system itself. The Spark platform layer mainly realizes various functions and businesses of the system, manages the database, and establishes interface services for the display layer to call. Among them, the workflow engine is responsible for the realization of the business process of the system, mainly including data collection and data mining. The data access engine implements data read and write operations and database management tasks. This part is the interface between the functional component layer and the database platform. The model engine is to obtain the extracted data and transform it into an instance object to prepare for the subsequent data mining work. The model engine divides the data into memory cache and persistent storage, and chooses the appropriate way to store according to the specific data type. The memory cache data should not be too much, and when the data fails, the memory space should be recovered in time to improve the efficiency and performance of the platform.

3.4 Evaluation criteria of algorithm

F-measure evaluation standard: F-measure is according to the exaction rate and recall rate in information retrieval. The definition of accuracy rate and recall rate is shown in formula (9).

$$p(m,n)=\frac{l_{mn}}{l_n}, r(m,n)=\frac{l_{mn}}{l_m} \tag{9}$$

Where l_n means the amout of data sets in each class n, l_m indicates the amout of data sets in cluster m. l_{mn} means the amout of points in class m and cluster n. for a class n and cluster m, F-measure is expressed as shown in formula (10):

$$F(i,j)=\frac{(b^2+1)\times p(m,n)\times r(m,n)}{b^2\times p(m,n)+r(m,n)} \tag{10}$$

Let $b=1$, so that $p(m,n)$ and $r(m,n)$ get equal weights. For the whole F-measure with X data set, as shown in formula (11), the larger the merit of F-measure is, the better the effect is.

$$F=\sum_i \frac{n_i}{n}\max\{F(i,j)\} \tag{11}$$

4. Experiments and results

4.1 Database description

The algorithm is implemented in Java language and debugged in Windows XP system and eclipse environment. The experiment is realized by MATLAB language. The hardware configuration of PC is Pentium (R) CPU, 3G memory and 300gb hard disk. 11 typical data sets of UCI data set are selected for simulation experiments, which include category number information, sample number information and dimension information. In order to judge the result of data dimensionality reduction, this chapter introduces the validity evaluation index, siloette index, which is used to express the cluster structure's intra cluster compactness and inter cluster separability. It can be used to determine the optimal cluster number, and also can be used to evaluate and analyze the cluster results. To evaluate the accuracy, this paper uses F-measure external evaluation to estimate the effect.

In the experiment, 11 typical UCI datasets were standardized and dimensionality was reduced. By adjusting the bias parameter P, the dimensionality was determined by the silhouette index.

The dimensionality of glass, hearts, WBC and German datasets after dimensionality reduction was fixed 2. The dimensionality of air, vowel and German datasets after dimensionality reduction was in two dimensions and three dimensions. The dimension selection of four kinds of data sets is more than three dimensions. It can be seen that the dimension selection of the data with dimension higher than 30 dimensions is more, while the dimension selection of the data with dimension higher than 10 dimensions is less. The final dimension reduction of each data set is determined according to the corresponding value of the silhouette index, and the dimension corresponding to the maximum value of the silhouette index. The detailed results are shown in table 1.

Table 1. detailed table of test dataset

Datasets	Number	Samples	Dimensions
Glass	6	214	10
Hearts	2	270	13
Ionosphere	2	351	35
Sonar	2	208	61
air	3	359	65
WBC	2	683	10
Vote	2	435	17
Vowel	11	528	11
German	2	1000	25
X8D5K	5	1000	9
Dnatest	3	1186	81

4.2 Accuracy Verification of proposed Clusterting algorithms under same dataset

To evaluate the parameter adaptive selection of the improved method and the effectiveness of the regional query method, database selection is carried out according to the three standards of data set dimension, data amount and data density distribution. Five typical data sets in the UCI database are selected, as shown in Table 2.

Table 2. Comparison of experimental results of different clustering algorithms

Datasets	Algorithm	Minpts	Eps	Time	F-measure	Accuracy (%)
Iris	DBSCAN	5	0.437	0.413	0.6959	89.321
	I-DBSCAN	5	0.398	0.401	0.7821	91.356
	AF- DBSCAN	5	0.369	0.210	0.7623	90.988
Wine	DBSCAN	5	24.689	0.397	0.4615	87.623
	I-DBSCAN	7	21.465	0.362	0.4165	88.648

	AF- DBSCAN	7	27.385	0.279	0.3968	88.561
Glass	DBSCAN	3	2.890	0.406	0.7532	96.732
	I-DBSCAN	3	2.671	0.435	0.6998	97.847
	AF- DBSCAN	3	2.549	0.201	0.6849	97.589
Cme	DBSCAN	5	2.764	4.239	0.3956	89.362
	I-DBSCAN	7	2.364	4.145	0.3665	91.064
	AF-DBSCAN	6	1.987	4.266	0.3564	90.874
Normal7	DBSCAN	7	12.176	4.947	0.7602	88.645
	I-DBSCAN	7	11.855	4.536	0.8529	89.045
	AF-DBSCAN	7	11.615	3.249	0.9968	89.013

The performance of DBSCAN, I-DBSCAN and AF-DBSCAN are compared. F-measure is used for clustering accuracy. EPS value in DBSCAN is set according to k-dist curve, and dist4 curve is selected for parameter determination according to the two indexes of accuracy and time characteristic analysis, as shown in Figure 8.

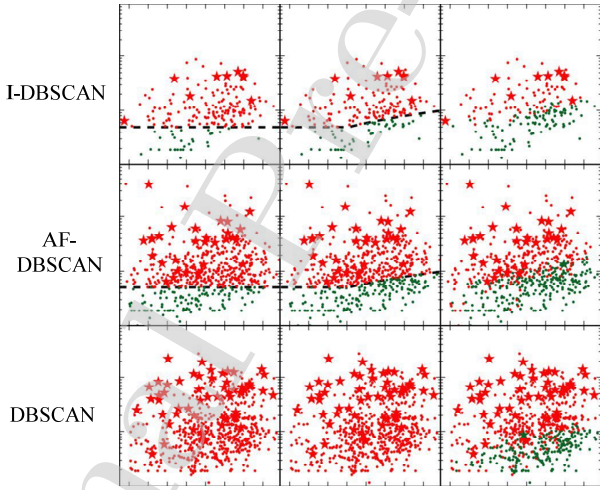


Figure 8. Comparative analysis of three clustering algorithms

According to the k-dist value corresponding to the sharp rise after the gentle change in the figure, it is taken as the value of the global parameter EPS, and the Min PTS value is set to 4. The (min PTS, EPS) of iris, wine, glass and CMC are (4, 0.436), (4, 27.330), (4, 3.700) and (4, 1.732), respectively. The (min PTS, EPS) of af-dbscan proposed in this chapter are (7, 0.389), (6, 29.870), (4, 2.695) and (5, 1.646), respectively. It can be seen from the table that the af-dbscan algorithm proposed in this chapter adaptively determines the global parameters, and avoids traversing all objects through the selection of seed representative objects. Under the

premise of ensuring the authentication efficiency, it effectively improves the accuracy of density clustering.

From the above analysis, we can see that the af-dbscan algorithm proposed in this chapter can effectively improve the efficiency of clustering, reduce the occurrence of missing points, reduce the error of normal data points as noise points, and effectively improve the accuracy of clustering by using adaptive global parameter determination and seed representative object selection. It can be seen that the global parameters adaptively calculated by the f-dbscan algorithm proposed in this chapter reduce the error and workload of determining the global parameter EPS according to the k-dist curve, and set min PTS as a fixed value of 4, so that the clustering results can not reach the global optimal effect. However, i-dbscan algorithm stipulates that the data conforms to Poisson distribution, and its F-measure value is unstable for different data sets, which can not adapt to data sets with different statistical characteristics. Because of the single density measurement index, af-dbscan algorithm is suitable for data sets with no obvious cluster density difference. The af-dbscan algorithm improved by region query runs faster than DBSCAN and i-dbscan algorithm, which reduces the time of density clustering significantly.

4.3 Performance verification of data mining method for association rules of matrix sort index

The spatial complexity of Apriori algorithm is mainly reflected in the memory occupation of candidate item set, AMBM algorithm is mainly reflected in the memory occupation of Boolean matrix and item set connection, while the memory occupation of IMISA algorithm is mainly for the storage of sorting index matrix, which effectively saves the computing space. In the real data set, random functions are ultilized to produce the data set and standard UCI data set randomly and automatically respectively. The data set generation method can accurately generate the required data set without preprocessing, and the size of the data set can be generated at any time according to the experimental needs. The standard UCI data set is used to simulate the Internet of things data with real data, reflecting the practical feasibility of theoretical research.

Firstly, the running time of each algorithm is compared and analyzed under the condition that the minimum support is different and the data set is the same. While $k = 3$, the

altogether opration time of proposed method is the time of sorting index matrix jumping search term set, while the total running time of Apriori algorithm is used to scan whole IoT matrix, producing candidate term set and pruning time. By comparing the running time under different support degrees, as shown in Figure 9, it shows that in the same data set ($|T|=1000,|I|=6$), the operation time of each presented algorithm is smaller than that of Apriori, which indicates that the calculation frequency 2-term set of IMSIA algorithm is calculated on the premise of this, and the useless items and transactions are deleted first to reduce unnecessary. Second, the frequency 2-term set is obtained by matrix multiplication and marking sequence, and then other frequent term sets are calculated by sorting marking matrix, which improves the calculation efficiency and is superior to Apriori algorithm and matrix algorithm.

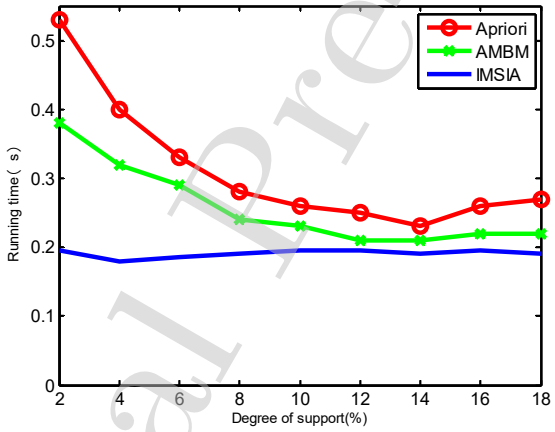


Figure 9. Opration time of three methods under the same data set for support degrees

Contrast of three algorithms operation time under the same support degree, as shown in Figure 10, wherein the data sets are randomly generated 9 kinds of data sets, and transactions in the data sets basically rises exponentially; the minimum support count is set as 2.

The algorithm is less, and the operation time of presented algorithm is always shorter than that of AMBM method. It is verified that the proposed IMSIA algorithm is the most efficient compared with the other two methods.

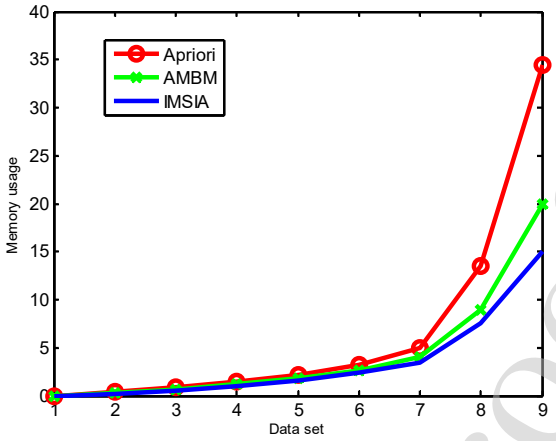


Figure 10. Running time comparison of different algorithms under the same support degree of different data sets

Considering the space cost of the algorithm, the proposed imsia algorithm reads the database only once, and does not produce other items set. It verifies the characteristics of sorting index matrix and spanning search item set, which takes up less memory. In different data sets, the memory usage of the algorithms is shown in Figure 11. In the case of the same data set ($|T|=1000, |I|=6$), and set the minimum support count to 2, and compare the memory usage of the three algorithms, as shown in figure, the memory usage of Apriori algorithm is 8.64%; that of ambm algorithm is 7.25%; that of imsia algorithm is 6.22%. The results show that the memory usage of ambm association rule algorithm is 1.39% lower than that of Apriori algorithm, and the memory usage of improved imsia algorithm is the lowest. Compared with Apriori algorithm, the memory usage of ambm association rule algorithm is reduced by 2.42%, which further proves that the improved algorithm has the characteristics of fast operation time and low space complexity.

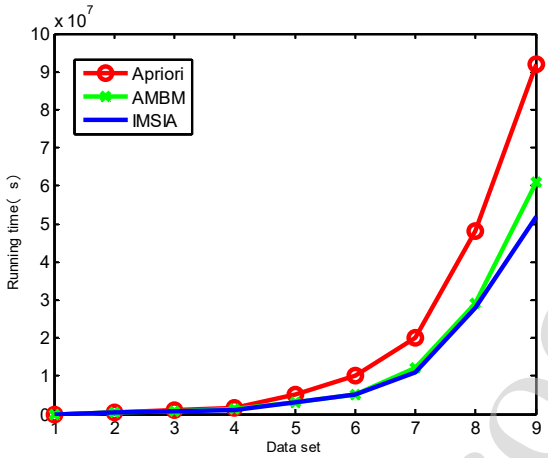


Figure 11. Running time comparison of different algorithms under the same support degree of different data sets

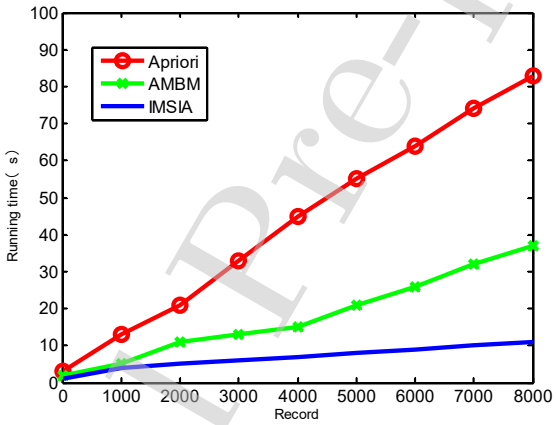


Figure 12. The opration time of three algorithms with different records in the mushroom dataset

Figure 12 shows the opration time of Apriori algorithm, AMBM algorithm and the algorithm proposed in this chapter in generating association rules on the mushroom dataset with large data volume. It can be expressed from the figure that with the addition of records, the time consumption of Apriori algorithm increases significantly, while the algorithm proposed in this chapter grows slowly in generating association rules, and with the increase of the amount of recorded data. In this chapter, the time consumption of the method shows a relatively stable growth trend. Compared with the other two methods, the imsia method proposed in this chapter has low time complexity and can effectively mine association rules among data sets.

Through the above data comparison can be obtained. The connect data set has 67557 records and 42 attributes. Because the Apriori algorithm requirs to read IoT frequently every time it generates frequent item sets, and produces other item sets. At the same time, it needs to prune the candidate option sets, which leads to the addiciton of the method's cost and affects the algorithm's operation efficiency. The algorithm in this chapter improves the efficiency of data mining effectively because of the advantages of using reduction matrix, scanning database only once and jumping search of sorting index matrix. The experimental results expressed that the algorithm in this section is superior to the other two typical algorithms under the data set generated by random numbers and the typical UCI data set.

4.4 Energy and efficiency analysis of Internet of things data mining algorithm based on cloud platform

The proposed IMSIA algorithm is compared with Apriori algorithm in terms of time complexity: assuming that the transaction database scale is m, n represents the number of attributes, C_k represents the candidate frequent k-item set, L_k represents frequent k-item set, and the maximum length of frequent item set is t.

(1) Apriori algorithm not only requires to read the IoT data many times, but also needs to produce the other option set C_{k-1} by combination, and then obtains the frequent item set L_k by pruning the other option set C_{k-1} according to the support threshold, while imsia algorithm generates the frequent item set L_k by matrix multiplication and index matrix jump connection, avoiding the tedious generation process of candidate option set.

The DBSCAN density clustering algorithm needs human intervention in the global parameter selection, and the process of regional query is complex and the query is easy to lose objects. An improved parameter adaptive and regional query density clustering algorithm is proposed, which can effectively delete the redundant data in the high-level complex data space on the premise of retaining the internal nonlinear structure of the IoT data.

(2) For the generation rule of frequent k-term set with $k=3$, the IMSIA algorithm proposed in this paper uses L_2 pruning matrix multiplication to analyze frequent 2-term set

generated by upper triangular matrix, which saves pruning operation of candidate term set and improves the efficiency of frequent term set generation. For the generation rule of frequent k-term set of $k=3$, this paper establishes 0-1 matrix and sorting index matrix according to frequent $(k-1)$ term set, and generates frequent k-term set according to jumping down join operation.

For the data set with m transaction database and N attributes, the frequent time complexity of Apriori algorithm is $O(m^2)$, mainly for the generation of database scanning and candidate item set, the time complexity of ambm algorithm is $O(mn)$, mainly for the generation of Boolean matrix and the calculation of combination and iteration of each item set, the time complexity of proposed IMSIA method to read IoT data is $O(nn)$. The main calculation amount is the generation of 0-1 matrix and sorting target matrix, which effectively improves the calculation efficiency of the algorithm. The detailed effciency comparison figure is shown in the Figure 13. As is presented, the opration time of IMSIA method increases linearly with the amount of data, while the opration time of Apriori method increases exponentially, this is consistent with our above analysis. So our algorithm is more efficient?

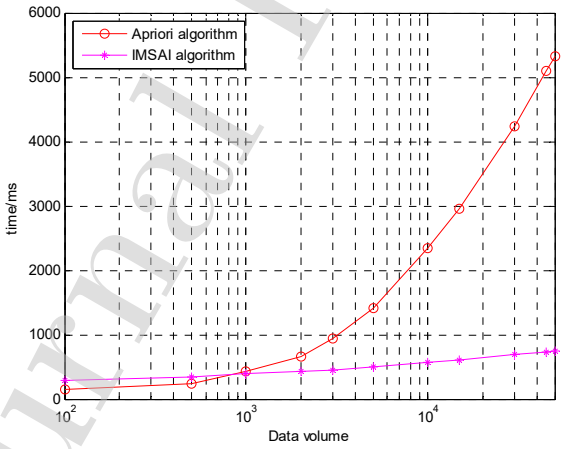


Figure 13. The comparison of training efficiency under different deep neural network model

5. Conclusions

With the continuous development of Internet technology and electronic information technology, big data technology and cloud computing technology also rise and develop, and have a positive impact on people's lives. Data mining system can deeply mine the value

information contained in big data, so as to assist users to solve practical problems and help users to make correct decisions and judgments. This paper presents an energy analysis of data mining algorithm based on cloud platform for Internet of things (IoT). First of all, an improved Apriori algorithm is proposed, which is based on Boolean matrix and sorting index rules. Then Boolean matrix is obtained after scanning the dataset and the Boolean matrix is preprocessed to delete the useless transactions and the item set, which are combined with sorting index to produce other item sets, effectively improving the efficiency of frequent item mining, which effectively reduce the memory usage. Secondly, the DBSCAN density clustering algorithm needs human intervention in the global parameter selection, and the process of regional query is complex and the query is easy to lose objects. An improved parameter adaptive and regional query density clustering algorithm is proposed, which can effectively delete the redundant data in the high-level complex data space on the premise of retaining the internal nonlinear structure of the IoT data. Effectively improve the efficiency of clustering. Finally, the simulation based on cloud platform verifies the effectiveness and superiority of the algorithm.

References

[1] Ferraro, U Petrillo, et al. "Informational and Linguistic Analysis of Large Genomic Sequence Collections via Efficient Hadoop Cluster Algorithms. " *Bioinformatics,* vol.34, no.11, pp. 1-11, 2018.

[2] Kesavaraja, D., and A. Shenbagavalli. "Framework for Fast and Efficient Cloud Video Transcoding System Using Intelligent Splitter and Hadoop MapReduce." *Wireless Personal Communications* vol. 2, pp. 1-16, 2018.

[3] Kong, B., et al. "Demonstration of application-driven network slicing and orchestration in optical/packet domains: on-demand vDC expansion for Hadoop MapReduce optimization." *Optics Express* , vol. 26, no. 11, pp. 14066-, 2018.

[4] Chaudhary, Rajat, et al. "Optimized Big Data Management across Multi-Cloud Data Centers: Software-Defined-Network-Based Analysis." *IEEE Communications Magazine* , vol. 56, no. 2, pp. 118-126, 2018.

[5] Luna, J. M., et al. "Apriori Versions Based on MapReduce for Mining Frequent Patterns

on Big Data. " *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 1-15, 2018.

[6] Khan, Murad, et al. "Context-aware low power intelligent SmartHome based on the Internet of things." *Computers & Electrical Engineering vol.* 52, no. C, pp. 208-222, 2016.

[7] Ruiz, M. Carmen, et al. "Formal performance evaluation of the Map/Reduce framework within cloud computing." *Journal of Supercomputing,* vol. 72, no. 8, pp. 3136-3155, 2016.

[8] Um, Jung Ho, et al. "Distributed RDF store for efficient searching billions of triples based on Hadoop." *Journal of Supercomputing, vol.* 72, no. 5, pp. 1825-1840, 2016.

[9] Chen, Dequan, et al. "Real-Time or Near Real-Time Persisting Daily Healthcare Data Into HDFS and ElasticSearch Index Inside a Big Data Platform." *IEEE Transactions on Industrial Informatics* , vol 13, no. 2, pp. 595 – 606, 2017..

[10] Zhang, Chuanting, et al. "Deep Transfer Learning for Intelligent Cellular Traffic Prediction Based on Cross-Domain Big Data." *IEEE Journal on Selected Areas in Communications* vol.37, no.26, pp. 1389 - 1401, 2019.

[11] Fang, Yuling, et al. "RGCA: A Reliable GPU Cluster Architecture for Large-Scale Internet of Things Computing Based on Effective Performance-Energy Optimization." *Sensors* vol.17, no.8, pp. 1799 -, 2017.

[12] Mahapatra, Chinmaya, A. K. Moharana, and V. C. M. Leung. "Energy Management in Smart Cities Based on Internet of Things: Peak Demand Reduction and Energy Savings." *Sensors* vol.17, no.2, pp. 2812 -, 2017.

[13] Pau, Marco, et al. "Design and Accuracy Analysis of Multilevel State Estimation Based on Smart Metering Infrastructure." *IEEE Transactions on Instrumentation and Measurement* vol.68, no.11, pp. 4300 - 4312, 2019.

[14] Ji, Cun, et al. "A Fast Shapelet Discovery Algorithm Based on Important Data Points." *International Journal of Web Services Research* vol.14, no.2, pp. 67-80, 2017..

[15] Muhammad, Zahid, et al. "Hybrid Artificial Bee Colony Algorithm for an Energy Efficient Internet of Things based on Wireless Sensor Network." *Iete Technical Review* vol.34, no. sup1, pp. 39-51, 2017.

[16] Ying, Zuo, T. Fei, and A. Y. C. Nee. "An Internet of things and cloud-based approach for

energy consumption evaluation and analysis for a product." *International Journal of Computer Integrated Manufacturing* vol.31, no. 4, pp. 1-12, 2017.

[17] Kaur, Kuljeet, et al. "Edge Computing in the Industrial Internet of Things Environment: Software-Defined-Networks-Based Edge-Cloud Interplay." *IEEE Communications Magazine* vol. 56, no. 2, pp. 44-51, 2018.

[18] Piccialli, Francesco, S. Cuomo, and G. Jeon. "Parallel Approaches for Data Mining in the Internet of Things Realm." *International Journal of Parallel Programming* vol. 46, no. 5, pp. 1-5, 2018.

[19] Rashid, Md. Mamunur, I. Gondal, and J. Kamruzzaman. "Dependable large scale behavioral patterns mining from sensor data using Hadoop platform." *Information Sciences* vol.379, no. 1, pp. 128-145, 2017.

[20] Xiao, Y., et al. "A Study of Pattern Prediction in the Monitoring Data of Earthen Ruins with the Internet of Things:." *Sensors* vol.17, no. 5, pp. 1076-, 2017.

[21] Sidibé, Abdoulaye, and S. Gao. "Study of Automatic Anomalous Behaviour Detection Techniques for Maritime Vessels." *Journal of Navigation* vol.70, no. 4, pp. 847-858, 2017.

[22] Laska, Marius, et al. "A Scalable Architecture for Real-Time Stream Processing of Spatiotemporal IoT Stream Data—Performance Analysis on the Example of Map Matching." *ISPRS International Journal of Geo-Information* vol.7, no. 7, pp. 238-, 2018.

[23] Lin, Hsueh Yuan, and S. Y. Yang. "A Smart Cloud-Based Energy Data Mining Agent Using Big Data Analysis Technology." *Microelectronics Reliability* vol. 97, no. 6, pp. 66-78, 2019.

[24] Xu, Xiaowei, et al. "MDA: A Reconfigurable Memristor-based Distance Accelerator for Time Series Mining on Data Centers." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* vol. 38, no. 5, pp. 785 - 797, 2018.

[25] Zhu, Jinkang, et al. "Foundation study on wireless big data: Concept, mining, learning and practices." *China Communications* vol. 15, no. 12, pp. 1-15, 2018.

[26] Guo, Kehua, Y. Tang, and P. Zhang. "CSF: Crowdsourcing semantic fusion for heterogeneous media big data in the internet of things." *Information Fusion* vol. 37, no. 1, pp. 77-85, 2017.

[27] Farhan M, Jabbar S, Aslam M, et al. A Real-Time Data Mining Approach for Interaction Analytics Assessment: IoT Based Student Interaction Framework[J]. International Journal of Parallel Programming, vol. 17, no. 1, pp. 1-18, 2017.

[28] Yang, Jinfei, J. Li, and S. Liu. "A new algorithm of stock data mining in Internet of Multimedia Things." *Journal of Supercomputing* vol. 9, no. 1, pp. 1-16, 2017.

[29] Wang, Dan, et al. "From IoT to 5G I-IoT: The Next Generation IoT-Based Intelligent Algorithms and 5G Technologies." *IEEE Communications Magazine* vol. 56, no. 10, pp. 114-120, 2018.

[30] Fu, Junsong, et al. "Secure Data Storage and Searching for Industrial IoT by Integrating Fog Computing and Cloud Computing." IEEE Transactions on Industrial Informatics vol. 14, no. 10, pp. 4519 - 4528, 2018.

[31] Said, Omar. "Analysis, design and simulation of Internet of Things routing algorithm based on ant colony optimization." *International Journal of Communication Systems* vol. 30, no. 8, pp. 3174-, 2016.

[32] Liu, Mingyang, M. Qu, and B. Zhao. "Research and Citation Analysis of Data Mining Technology Based on Bayes Algorithm." *Mobile Networks & Applications* vol. 22, no. 3, pp. 1-9, 2016.

[33] Zhao, Jianyang, et al. "Wide-area smart grids with new smart units synchronized measurement analysis and control based on cloud computing platform." *International Journal of Energy Research* vol. 40, no. 3, pp. 362-378, 2016.

[34] Xiu, Li, J. Song, and B. Huang. "A scientific workflow management system architecture and its scheduling based on cloud service platform for manufacturing big data analytics." International Journal of Advanced Manufacturing Technology vol. 84, no. 1, pp. 119-131, 2016.

[35] Laghari, Samreen, and M. A. Niazi. "Modeling the Internet of Things, Self-Organizing and Other Complex Adaptive Communication Networks: A Cognitive Agent-Based Computing Approach." Plos One vol. 11, no. 1, pp. e0146760-, 2016.

[36] Qiang, Yi, and N. S. N. Lam. "The impact of Hurricane Katrina on urban growth in Louisiana: an analysis using data mining and simulation approaches." International Journal of Geographical Information Science vol. 30, no. 9, pp. 1-21, 2016.

Ziping Du was born in JiangSu, China, in 1980. From 1999 to 2003, he studied in Nanjing Normal University and received his bachelor's degree in 2003. From 2007 to 2009, he studied in Beijing University of Posts and Telecommunications and received his Master’s degree in 2009. Currently, he works in Industrial Park Institute Services Outsourcing. He has published a total of fifteen papers. His research interests are included IT、WSN and IoT. Email: zipingdu@126.com

Author statement

Ziping Du is the sole author of this article.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: