IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 15, 2019, accepted November 25, 2019, date of publication December 9, 2019, date of current version December 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958377

Energy Analysis and Application of Data Mining Algorithms for Internet of Things Based on Hadoop Cloud Platform

YUANPAN ZHENG AND GUANGYU CHEN
School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450000, China
Corresponding author: Yuanpan Zheng (ypzheng@zzuli.edu.cn)

ABSTRACT The paper analyses and studies the classification and characteristics of Internet of Things (IoT) information, and discusses the construction and application of Hadoop Cloud Platform. This paper mainly carries out from two aspects. One is to design the system architecture of the Open Platform for Data Simulation Resources of the IoT and design the key modules. A platform for data simulation resources of the IoT is built to provide the running environment and external services for the sensor data simulation model established. On the other hand, it is the key method to study the simulation data model based on IoT sensors. That is, based on the research environment of the IoT built by the existing laboratories, collect the data of sensors, analyze and study the characteristics of sensors in the IoT, and design the key algorithms for data simulation. This paper presents two key models for sensor data modeling: the Long Short-term Memory (LSTM) prediction model and the Support Vector Machine (SVM) model based on IoT data, which are suitable for different data volumes. Extensive simulations are executed to validate the remarkable nature in Hadoop platform, in terms of prediction accuracy and training efficiency under different working condition.

INDEX TERMS Energy analysis, Internet of Things, cloud platform, Hadoop.

I. INTRODUCTION

Cloud computing and Internet of Things (IoT) are two hot research directions in IT field and scientific research institutes in recent two years. Cloud computing is an industry model and technology system that integrates and abstracts IT resources to users [1], [2]. It is a distributed basic platform that integrates computing capacity, storage capacity, broadband capacity and service capacity. IoT means a network that connects everything by RFID technology, infrared sensors, GPS information [3]. Nowadays, chip technology, sensors chips, communication technology and global positioning technology are all relatively mature. How to integrate and link these technologies will undoubtedly be used in the Internet, and the key to the development of the IoT is that each node of the Internet needs a platform to connect them. The emergence of cloud computing has just solved the connection problem in the IoT [4]. Only in cloud computing can the IoT achieve the effect of interconnection of things and networks [5]. Communication between sensors in the IoT will inevitably produce huge amounts of data. How to find the desired massages is a biggest problem [6]. Data mining used to be done on high-performance computers [7]. Only high-performance computers can achieve massive data processing. Of course, there are also problems that high-performance computers can't handle. Now with cloud computing, it has low cost, low energy consumption and flexible computing capabilities [8]. Nowadays, the dimensionality and complexity of data are also increasing. Data mining has some special requirements, so data mining under cloud platform is worth studying and promoting, some PDE models are widely used in this field [9], [10].

At present, the discovery of available information data from heterogeneous data in the IoT system plays a fundamental role in the implementation of upper application of intelligent decision-making. However, in the IoT system, the data processing and mining methods for data characteristics must be different from traditional data mining methods [11]. The mass data mining method of the Internet of things should be oriented to specific applications, using improved methods

The associate editor coordinating the review of this manuscript and approving it for publication was Honghao Gao.

to filter all kinds of data, clean up clustering, classification, frequent patterns and other aspects [12]. In order to provide effective health data information for the application staff, Kesavaraja and Shenbagavalli proposed a unique model for storing RFID data, which can provide important compression and path dependence aggregate while protecting object transformation [13]. Kesavaraja and Shenbagavalli proposed a mechanism for compressing probability king flow, which can capture motion and special RF melon flow anomalies. The effective method is to filter and preprocess large data to decrease the dimension of big data [14]. About the research of filtering and mining available information in the IoT big data environment, the related literatures are elaborated [15]. Especially, the reference proposes a data crossover big data filtering and mining algorithm based on particle filter algorithm [16]. The literature adopts genetic algorithm to mine large data in the IoT [17]. Yang *et al.* studied the outlier mining algorithm for RFID data stream [18]. Rathore *et al.* proposed a frequent closed loop mining algorithm for RFID [19]. Chaudhary *et al.* proposed a framework for trajectory aggregation of moving objects [20]. For data mining of sensor data, general probability architecture for supervisory learning according to computation and memory is proposed [2], [21], [22].

The IoT system is an integrated platform that gathers data collection, data exchange, data processing and specific business applications [23], [24]. The key problem of its application is also integration problem. Only through effective technology integration can the above technologies be integrated to form a complete data collection, data exchange, data processing and data response [25]–[27]. Only by using the platform can the real application of the IoT be realized. In the data conversion and filtering cleaning problem, because the Internet of Things data are many high-dimensional data, unstructured data processing, referring to XML data cleaning technology and dimension reduction processing direction for research. Reference proposed a method of calculating tree editing distance in $O(n^3)$ time [28]. Tree editing distance is a method to measure the comparability between trees. Representing the minimum number of nodes that need to be added, deleted or modified when transforming from one tree to another. The similarity of XML data is described by Bayesian method in XML Dup system. The Bayesian structure in this model can also be expressed as a tree structure [29]. The phase between the values on the leaves of two XML data is used as a prior probability on the leaves of the Bayesian structure. Document discussed the optimization strategy in describing the similarity of XML documents by Bayesian network [30], [31]. The strategy is to victories the XML documents and to determine the new structure of documents by training and learning [32].

This paper designs the cloud platform of IoT system based on Hadoop framework. On this basis, the prediction model based on deep regression network and the prediction classification model based on support vector machine (SVM) and Long Short-Term Memory (LSTM) models are established. Running the two data analysis models on the cloud platform can effectively improve the efficiency of the algorithm. Extensive simulations are executed to validate the remarkable nature in Hadoop platform, in terms of prediction accuracy and training efficiency under different working condition.

The rest of the paper was organized as follows. Architecture design of IoT based on Hadoop cloud platform was expressed in Section II. Section III described data mining of big data in IoT based on SVM and deep learning. Experimental results were discussed and analyzed in detail in Section IV. Finally, Section V concluded the work and proposed the outlook.

II. ARCHITECTURE DESIGN OF IoT DATA MINING SYSTEM BASED ON HADOOP CLOUD PLATFORM

A. ARCHITECTURE DESIGN OF FOUR-LAYER DATA MINING SYSTEM BASED ON HADOOP CLOUD PLATFORM

In the application of data in the IoT, physical modeling needs to understand the relationship between things first, and then establish the old mathematical model describing the quantitative relationship. However, the problems of data loss or error, high data complexity in the IoT, the use of traditional physical modeling methods will have great limitations.

In order to make servers serve well, it is necessary to increase the number of servers, or to restrict the amount of access. In reality, this is not feasible. Because the amount of access is constantly changing, increasing the number of servers will cause waste of equipment. If access is restricted, there will be some drawbacks, so using cloud computing in the IoT can help the IoT system solve this problem.

As shown in Figure 1, an intelligent data mining system based on Hadoop Cloud Platform is composed of sensor perception layer, transmission layer, data mining layer and application layer.

In the sensor perception layer, the user data, transportation data, GPS statistical data, wireless data and GIS data are collected to achieve IoT data acquisition of in real world. It is the base of the optimization system.

The transmission layer is one of the key layers in the whole network architecture. It is mainly responsible for providing services for the communication between processes in two hosts. Because a host runs multiple processes at the same time, the transport layer has the functions of reuse and sharing.

The data mining layer is a technology to find its rules from big data by analyzing each data; it is the bridge of the information and services. It can use the advanced machine learning data mining and analysis, get the data we need, and carry out visualization.

The application layer relies on the hidden data relations and key information mined by the data mining layer to provide users with analysis results, visualization and so on. The application layer directly serves the application process. Its function is to accomplish a series of services needed for

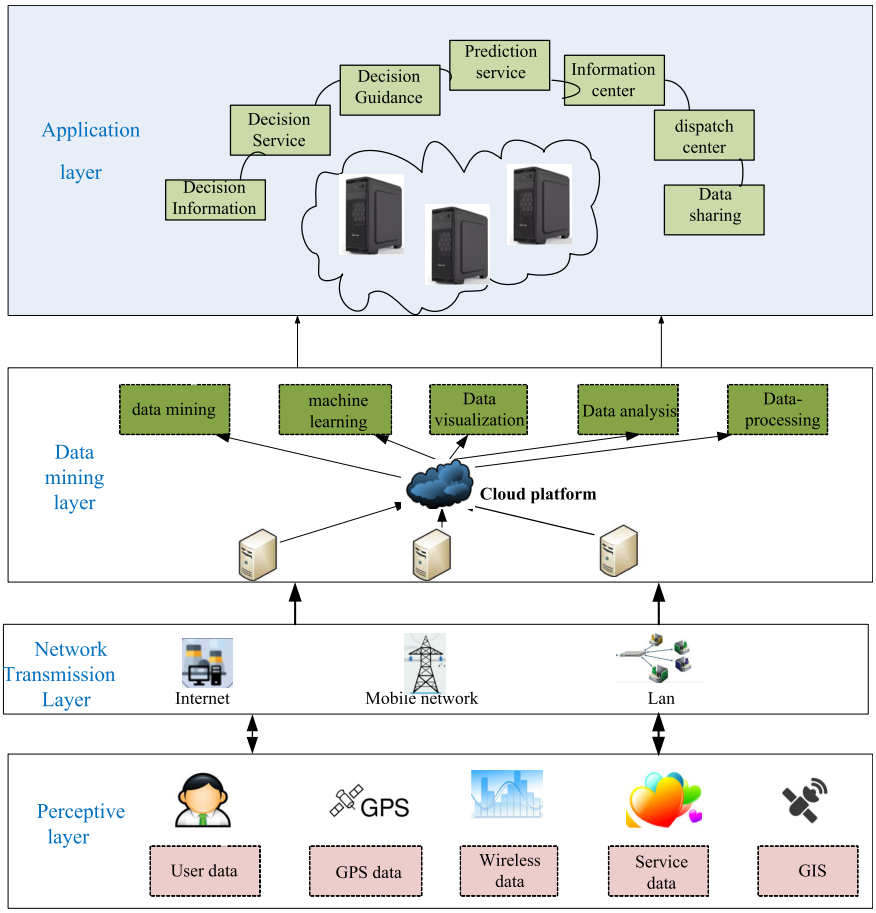

FIGURE 1. Architecture design of four-layer data mining system based on Hadoop Cloud Platform.

business processing while communicating with each other in multiple application processes. Its service elements are divided into two categories: CASE, a common application service element, and SASE, a specific application service element.

B. THE SYSTEM DESIGN OF USER TERMINAL AND DATA MANAGEMENT CENTER

The system consists of data acquisition sensor, data acquisition terminal, transmission network, data management center and user terminal. As shown Figure 2,

(1) Data acquisition sensor of IoT: According to the various data acquisition modules installed in the field, various effective data related to the equipment are collected. The parameters of the data system and the environment parameters outside the system are obtained by sensors.

(2) IoT data acquisition terminal: The data acquisition terminal of IoT is a multi-functional device, which is now available to users. IoT data collected by the field generally have a variety of upstream and downstream data interfaces and communication protocols, which can effectively process data and increase the data transmission efficiency; it also can achieve communication with intelligent communication devices with different protocols.

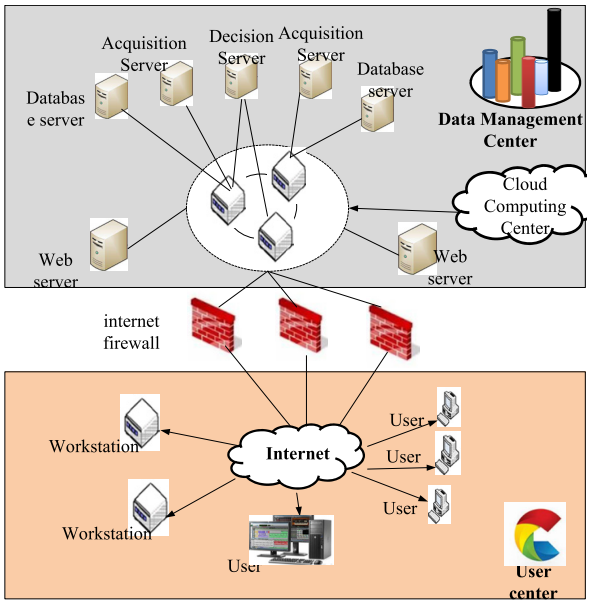

FIGURE 2. The system design between data management center and user center.

(3) Transmission network: Transmission network can upload monitoring data to management center, which is the link between user center and data management center,

and providing transmission methods to meet the requirements of existing transmission channels, such as Ethernet, GPRS/CDMA and so on.

(4) Data management center: The index data center can store and manage the data of the IoT in the background of the monitoring system. Then the data from the IoT is collected, and the data center uses distributed technology to realize storage. The data center stores a large amount of data by using disk array technology. Then it balances the transmission and storage of data through multi-server technology, and backs up the collected data of the IoT regularly. It uses the distributed processing algorithm (such as MapReduce) module to distribute the operation of the system, such as semantic analysis and reasoning. Mining algorithm for optimal control in large-scale cluster system with parallel computing ability is obtained from data.

(5) User terminal. User terminal is to analyze and compare the data of data center, and display it through Web site. It can realize the energy-saving management of users in the authorization login system, as convenient as browsing ordinary websites.

As the source of the Internet of things, cloud data processing platform is supported by cloud computing and pervasive computing. Through the calculation and analysis of massive information in the network, a large-scale intelligent Internet of things network is formed, which is efficient and scalable, and can provide reliable technical support platform for the upper service management and industry application. Using cloud computing technology to build the Internet of things database can better provide accurate and comprehensive diagnosis information and preventive measures for digital and remote control schemes. In the data processing platform, data mining for the complex information of the Internet of things is very important. The key to promote the application of Internet of things is to preprocess the Internet of things information, integrate advanced deep learning network model, and develop a fast and robust mining algorithm. In data processing, it is very important to design and study efficient data processing algorithms.

C. SPATIAL-TEMPORAL RELATIONSHIP OF DATA IN DATA MINING SYSTEM OF IoT

The data mining mode depends on the environment of the IoT system. Because of the different characteristics of the IoT, such as the complexity of data and the correlation between the modeling mode of the IoT and the traditional mode, the data mining mode of IoT is to analyze the data characteristics of IoT firstly, then proposes appropriate solutions, and then summarize the appropriate physical model. The characteristics of data in the IoT are as follows: relevance, large amount of information (mass), poor quality, spatiotemporal and non-structural, which is totally different from the traditional data mining field.

The spatiotemporal nature of IoT data, the original data is usually collected from a four-dimensional IoT network. As shown in Figure 3, each point in the abstract sketch can represent an individual in the IoT, and each edge can represent the interconnection of things in the IoT.

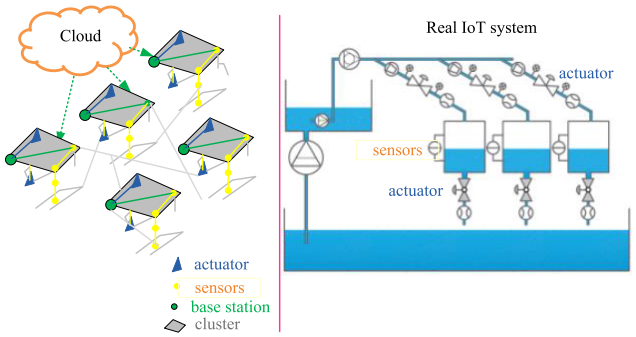

FIGURE 3. The basic physical structure of IoT system.

Cloud computing technology is used in file data storage of IoT data processing platform, and distributed file system HDFS based on Hadoop cluster is used for storage. Because there may be redundancy and abnormal data in the data collected by a large number of heterogeneous intelligent collection terminals, distributed data cleaning is needed for the data. The data after cleaning can be based on Hadoop number. The distributed data processing algorithm is designed based on the data processing platform, and the distributed data mining algorithm is mainly studied to get the information needed for the upper application.

In the application of the IoT, the data of the IoT will be lost and errors in batches. For this kind of data errors and losses of the IoT, they may be random or systematic. The data mining mode of the IoT based on cloud computing should take into account the problem of data loss and errors, and we give the solution. The case should be able to tolerate data errors and loss. In data mining application modeling based on the IoT, we should also fully consider how to express the relationship between physical individuals. If physical individuals are indirect relationships, we can derive them by Laplace transform model or SVD model. The direct relationship is very important. The data mining model of the IoT should have the ability to fully express the direct relationship, which will facilitate the inference of the indirect relationship.

III. EFFECTIVE INFORMATION FILTERING MINING OF BIG DATA IN IoT BASED ON SVM AND DEEP LEARNING

To realize the intelligent data mining and decision-making of the IoT, this paper applies deep learning algorithm and SVM to the data warehouse system. SVM is suitable for small quantity and low precision, while LSTM is suitable for complex nonlinear big data state. Each data mining algorithm has its own advantages. We must choose the corresponding algorithm to face different data types and data volumes. It uses principal component analysis (PCA) to implement reduction of data dimensions, extracts data features, and normalizes the data. The architecture of data mining system of IoT based on SVM and deep learning is designed, as shown in Figure 4.

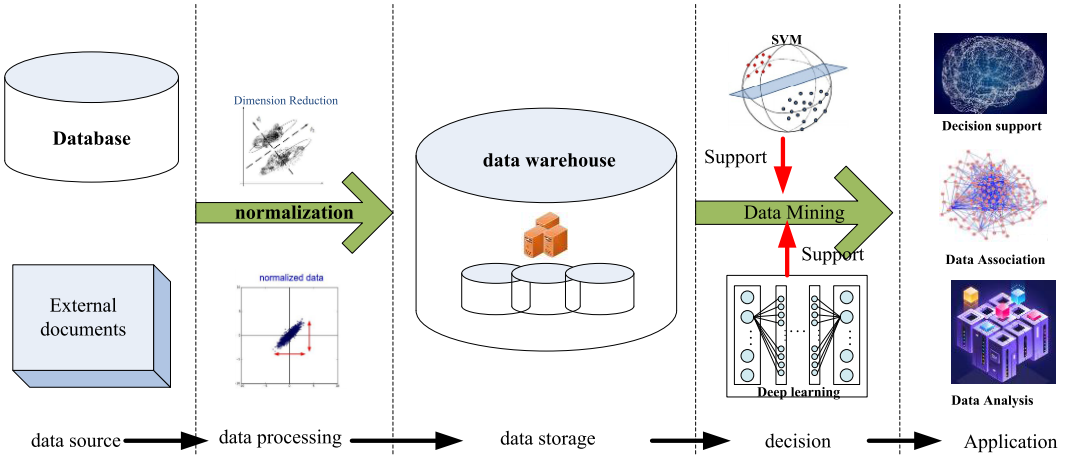

FIGURE 4. Architecture of data mining based on SVM and Deep learning.

As it is revealed in Figure 4, the data mining methods based on SVM and deep learning include data collection, data processing, data storage, and decision application section. Decision section, which applies SVM and deep learning neural network to achieve the data prediction and data classification, is the core algorithm in data mining system.

A. INFORMATION FEATURE EXTRACTION AND PROCESSING OF BIG DATA IN IoT

Based on the above-mentioned cloud platform system framework, this paper designs and processes the information feature extraction algorithm for the big data of the IoT system. According to the spatial information discreteness of the big data of the IoT, discrete points are considered, we need to design reduction method for the feature dimension. The features contain various IoT data model categories, assuming that the big data of the IoT is discrete. The starting time of the feature partition is t_0, and the effective information of the big data in the IoT is distributed at the *i-th* level, so the prediction system state $x(t_k)$ is expressed as:

$$\dot{Y} = |AY + B[f(Y) + u] \tag{1}$$

For the cleaning of XML data, it is an important problem to detect similar duplicate data. Moreover, the corresponding business standard XML can be represented as an ordered number tree by W, and each node corresponds to the corresponding data node. Therefore, tree editing distance algorithm or other related intelligent algorithm can be used to determine similarity. Data vector X divides as $\{X_v, v = 1, 2, \cdots V\}$. MapReduce is the decomposition of tasks and the summary of results. Data features can be shown as following form:

$$g_{mn}(t) = g(t - mT)e^{j2\pi(nF)t} \quad m, n = 0, \pm 1, \pm 2 \cdots \tag{2}$$

Supposing that observational values of IoT sensor S_k is θ, The result of data feature segmentation satisfies the following expression:

$$x_{id}^{t+1} = wx_{id}^{t} + c_1 r_1(p_{id} - x_{id}^{t}) + c_2 r_2(p_{gd} - x_{id}^{t}) \tag{3}$$

When data is transferring, others are disallowed to send real time value of sensors at the same time. Because of the diversity of intelligent collection terminals in the Internet of Things, most of the collected information may be heterogeneous, so the data loading and conversion module of middleware data processing should have the function of data receiving and conversion. Therefore, it is necessary to write XML conversion configuration files for data sent by different data protocol devices, and establish data representation specifications for different applications for upper data processing. The feature extraction is defined as $P(X, T_d)$.

$$P(X, T_d) = \prod_{i_u \in X \wedge X \subset T_d} p(i_u, T_d) \tag{4}$$

The cumulative interference of IoT node is computing as follows

$$\begin{aligned} \Re(\rho, \theta) &= \frac{p(z_1|x_t)p(x_t|u_{t-1,\cdots z_0})}{p(z_t|u_{t-1}, d_{0,\cdots t-1})} \\ &= \eta p(z_1|x_t) \int p(x_t|x_{t-1}, u_{t-1}) Bel(x_{t-1}) dx_{t-1} \end{aligned} \tag{5}$$

Assuming $x(t)$ is the sequence of the samples, filtering the non-associated message of the big data of the IoT reasonably, extracting the main characteristic quantity of the correlation degree, and obtaining the probability density of node j by node i is shown as follows:

$$\begin{aligned} \Re(\rho, \theta) &= \frac{p(z_1|x_t)p(x_t|u_{t-1,\cdots z_0})}{p(z_t|u_{t-1}, d_{0,\cdots t-1})} \\ &= \eta p(z_1|x_t) \int p(x_t|x_{t-1}, u_{t-1}) Bel(x_{t-1}) dx_{t-1} \end{aligned} \tag{6}$$

Firstly, data cleaning is carried out by MapReduce program. After the cleaned files are stored in HDFS, in order to carry out more targeted data control and mining research, W continues to import data into Curve data warehouse and carry out data cleaning based on Skillful, which lays a good foundation for the following data mining work.

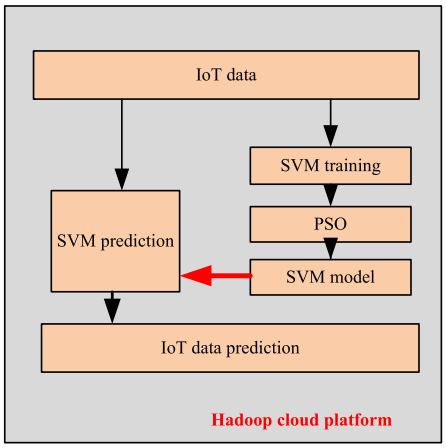

FIGURE 5. The SVM algorithm based on Hadoop Cloud Platform.

B. IMPLEMENTATION OF DATA MINING ALGORITHMS BASED ON SVM

On the foundation of the architecture of correlation dimension extracted features, we design an effective data mining algorithm in the IoT. The algorithm has high requirements for the initial trajectory of particle filter and high data noise. An efficient information filtering and mining algorithm for big data in IoT based on SVM is proposed, as shown in Figure 5. PSO is used to optimize the weight of SVM and improve the convergence speed and accuracy of the algorithm. It always improves the efficiency of the algorithm and reduces energy consumption and improve practicability.

Firstly, SVM local mining is carried out on the data sets of each sub-site; then, multi-tree construction algorithm is used to map the support vectors extracted locally into local feature multi-tree, and support vectors and information are loaded into the next site by mobile agent. Then, the new samples (shell vectors of the first few sites) and the existing samples (samples of the next site) are merged to mine. With the accumulation of sample sets (the movement of each site), the learning accuracy is gradually improved, and the global mining of SVM in distributed environment is finally realized. And then the number is counted. Finally, data specification, data mining rules are formulated, and data valid information. Then, under some heuristic constraints, super-rectangular rules are constructed based on the obtained support vector and clustering center. It is very easy to control the support degree and quantity of rules in SVM, and the obtained rules have higher quality. According to the above improved thinking, the core technologies for the implementation of the methods are described as follows:

Define: $\left\{S_j^{(n)}, j = 0, 1, 2, \cdots N-1\right\}$ of SVM method and seek the minimum range of SVM node N_{j^*}, $d_{j^*} = \min\{d_j\}$. The Flow chart of SVM algorithm is expressed in Figure 6.

Defining unbiased risk estimate is exp $SN(X)$:

$$\exp SN(X) = \sum\nolimits_{T_d \supseteq X \wedge T_d \in D} P(X, T_d) \tag{7}$$

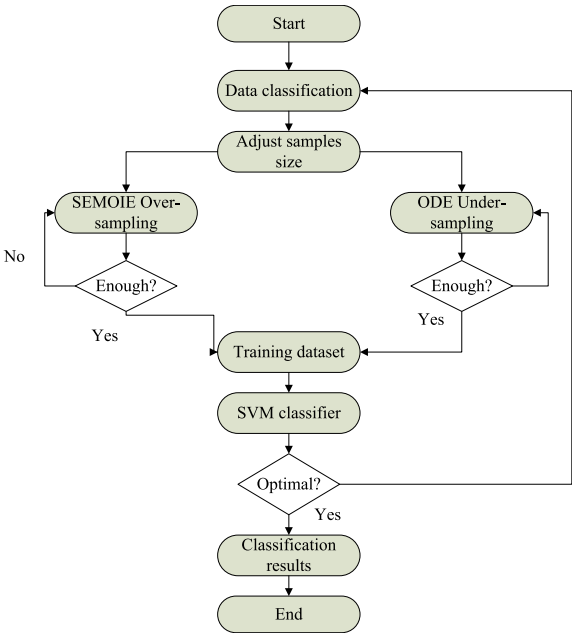

FIGURE 6. The Flow chart of SVM algorithm.

where T_d is the association massages from a single observation. Assuming that the standard support vector machine solution is expressed as:

$$\min W = \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K\left(x_i, x_j\right) - \sum_{i=1}^{l} \alpha_i + b\left(\sum_{i=1}^{l} y_j \alpha\right) \tag{8}$$

where $\left(x_i, x_j\right)$ are samples. The weight vector α_c is used to adjusting, and information dataset S_s are obtained.

$$Q' = \begin{bmatrix} 0 & y_1 & \cdots & y_n \\ y_1 & Q_{11} & \cdots & Q_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & Q_{n1} & \cdots & Q_{nn} \end{bmatrix} \stackrel{def}{=} \begin{bmatrix} 0 & y^T \\ y & Q \end{bmatrix} \tag{9}$$

Extracted features of the big data in the IoT system is gotten as follows:

$$\det\left(Q'\right) = \det(Q) \cdot \left(-y^T Q^{-1}\right) \neq 0 \tag{10}$$

$$\tau = \alpha(1+2^{-\alpha/2})(\alpha-1)^{-1} + \frac{\pi 2^{-\alpha/2}(\alpha-2)^{-1}}{2} \tag{11}$$

Then, the unbiased phase characteristics of the parent nodes on the data aggregation tree TDAG of the IoT are obtained as follows:

$$K = \left[(4\beta\tau P l^{-\alpha}) \cdot (2^{-\alpha/2} P l^{-\alpha} - \beta N_0)^{-1/\alpha} + 1 + \sqrt{2}\right] \tag{12}$$

According to the data transfer theory between dominant nodes, the time required for the topology center of the IoT sensor to transmit the results to sink is R in each round of effective information mining for big data in the IoT. Through the above analysis, the big data validity of the IoT based on SVM is realized for improvement of information filtering mining algorithm.

C. IMPLEMENTATION OF DATA MINING ALGORITHMS BASED ON LSTM

In Hadoop-based processing platform, effective data cleaning algorithm is compiled to filter the redundant data, and the process of cloud data cleaning is designed. According to the source code of MapReduce process, based on this programming model, the algorithms of anomaly filtering, business filtering, time filtering and similar filtering based on specific business are designed and studied. In order to further simplify the data and combine it with traditional database, the Hive data cleaning based on platform is analyzed and studied.

From the macro visual angle, the sensor data of the IoT has the characteristics of polymorphism and heterogeneity, massive data and fluctuation of data. From the perspective of data prediction, this type of data is better handled. However, for wind sensor data and relative humidity sensor data, the overall trend is around a certain mean line up and down vibration. Because the change of vibration frequency and amplitude is random, it will have some influence on the regression prediction model with common data.

In this paper, sensor data is abstracted into inter-inch sequence for processing and analysis. Time series has four elements: long-term trend, cyclic change, seasonal change, irregular change and so on. Among them, the long-term trend factor refers to the general change trend of the sequence data which is influenced by some fundamental factor for a long time. Cyclic variation factor refers to the regular change of the wave shape of the sequence data in a certain period of time. Seasonal variation factors refer to the regular periodic variation of sequence data with seasonal variation. Irregular change elements refer to change of sequence data in an irregular way, which includes strict random change and irregular sudden change with great impact. For sensor data, it satisfies the above four elements very well. The concepts of cyclist, volatility and data trend of the previous analysis are in good agreement with those of the four elements.

LSTM neural network was first proposed in 1997 [33]. It is a kind of perfection based on Recurrent Neural Network (RNN) to solve the problem of gradient extinction in RNN. Compared with ordinary neural networks, the biggest difference of RNN is related to the received data. This feature is very helpful for processing time-related data.

Figure 7 shows the internal structure of LSTM. Gated state is used to control the transmission state, remember the information that needs long memory and forget the unimportant information; unlike ordinary RNN, only one way of memory superposition can be achieved. This is especially useful for many tasks requiring "long-term memory".

The principle of LSTM can be shown in the following formulas (13)-(15):

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{13}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{14}$$

$$\tilde{C}_t = \tanh\left(W_C\,[h_{t-1}, x_t] + b_C\right) \tag{15}$$

$$o_t = \sigma\left(W_o\,[h_{t-1}, x_t] + b_o\right) \tag{16}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{17}$$

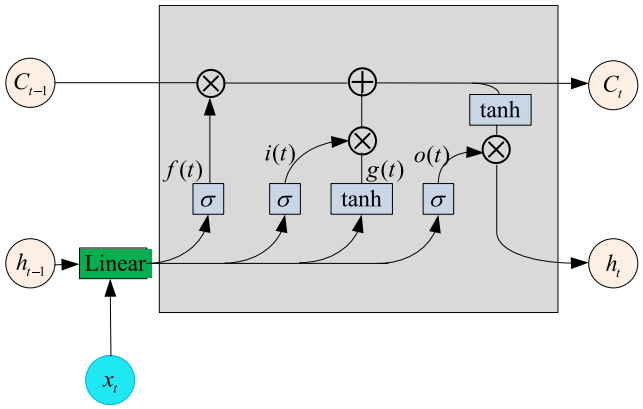

FIGURE 7. The internal structure of LSTM neural network.

Data representation specification is the unified representation format of data in data exchange platform. The formulation of data representation specification determines the efficiency of data conversion task. Data exchange specification should fully consider the scalability of application and the possibility of practical application, laying the foundation for the formation of relevant industry standards, so as to fully meet the needs of users and services.

The training model structure based on LSTM is shown in Figure 8. As we can see from figure 8, the LSTM neural network provides the prediction support for our model. In order to apply LSTM network, we describe the input and output data as sequential data to meet the input and output requirements of LSTM network.

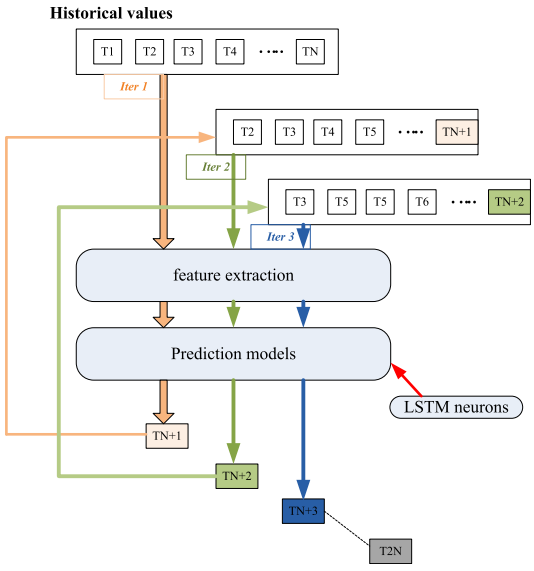

FIGURE 8. The prediction model structure based on LSTM.

IV. EXPERIMENTS AND DISCUSSIONS

A. CONSTRUCTION OF EXPERIMENTAL ENVIRONMENT AND DATA ACQUISITION

Due to the limitation of experimental conditions, the experimental platform used in this paper is a Hadoop distributed

cluster composed of four machines as a data processing platform for testing IoT system. One node is used as the Master nodes of HDFS and MapReduce, namely NameNode and MapReduce. JobTracker node is mainly responsible for metadata management and task scheduling of distributed data processing platform. Other nodes are Slave working nodes, namely DataNode and TaskTracker, which are mainly responsible for data storage and specific distributed computing processes. NameNode and DataNode are planned in the following table. After the deployment of three nodes, the firewall service ip-tables stop of three nodes should be closed first. The planned deployment allocation is shown in Table 1.

TABLE 1. Cluster deployment allocation table of data processing platform.

Node name	name	usage
192.168.200.92	Master	Namenode
192.168.200.86	Slave1	Datanode
192.168.200.87	Slave2	Datanode
192.168.200.89	Slave3	Datanode

To validate the performance of the algorithm in the implementation in the IoT systems, experiments are carried out on MATLAB 2010 in computers (Windows 7).

The structure model of IoT is set up by sensors network. When the sensors are used for data acquisition, the input energy is continuously supplied. In order to parse the information of the XML file effectively and quickly and find the data needed for sending to the upper application, the juicy XML should be rationalized without too many sub-tags nested; the tag name of the XML file can clearly reflect the content information it represents for easy maintenance after W; the storage content of the XML file should be as complete as possible, and the information should be reflected. Relevant analytical method for nodes uses some technology.

User parameters are defined as L = 500, weight vector W of ontology feature, information parameter of standard dataset in data mining is shown in Table 2.

TABLE 2. Standard data sets for experiment acquisition.

Dataset name	Dataset size	Attribute set size
Breast cancer	212	64
Ship radiated noise	426	35
temperature	812	43

For distributed data processing, this paper mainly studies distributed clustering and classification algorithm. the accuracy, average recall, iteration times and running time are mainly used for analysis and comparison. The accuracy rate is based on correctly clustered data to the total data. The average recall rate is the average value of the ratio of the number of classes correctly classified to the data objects that the cluster should have. Iterations required for the main algorithm is to run. Specific accuracy and average recall formulas are given as:

$$precision = \frac{N_r}{N_a} * 100\% \tag{18}$$

$$Recall = \frac{1}{k}\sum_{a=1}^{k}\frac{N_{a\to a}}{N_{a\to a}+N_{a\to b}} * 100\% \tag{19}$$

B. DISTRIBUTED CLUSTERING RESULTS ANALYSIS FOR IMPROVED SVM IN HADOOP CLOUD PLATFORM

Breast Cancer_dataset is selected from UCI Machine Learning. Each sample contains nine columns of attribute values and is divided into two categories: benign and malignant. Pima-indians-diabetes data set has 8 attribute values per sample. Due to the limitation of experimental conditions, the clustering effect of parallel SVM algorithm is tested by replication processing for each data set. The size of data is about 1GB, and the number of samples is tens of millions. Using the same clustering parameters and the same initial method, the accuracy, average recall, number of generations and running time of two data sets in clustering algorithm are unified by cluster experiments. The results are exhibition in Tables 3 and 4.

TABLE 3. Comparing tables of clustering results for breast cancer datasets.

algorithm	K-means	SVM	SVM+cloud
accuracy	96.49%	97.07%	97.36%
recall	96.61%	97.48%	97.97%
Iterations	8	6	5
time	733	711	699

TABLE 4. Comparing tables of clustering results for Pima-indians-diabetes datasets.

algorithm	K-means	SVM	SVM+cloud
accuracy	89.75%	92.71%	92.84%
recall	90.46%	92.95%	93.18%
Iterations	9	7	7
time	592	629	562

The experimental results indicate that under the H-node test cluster, the distributed clustering algorithm runs faster than the same serial memory algorithm. In terms of accuracy and average recall, the SVM clustering method proposed is better than parallel K-means based clustering algorithm under the same conditions. And SVM based on Hadoop platform is superior to SVM algorithms. However, we can also see some shortcomings from the test results. For example, the results are greatly influenced by the original physiological data set. The next step is to improve this aspect by combining experience and knowledge.

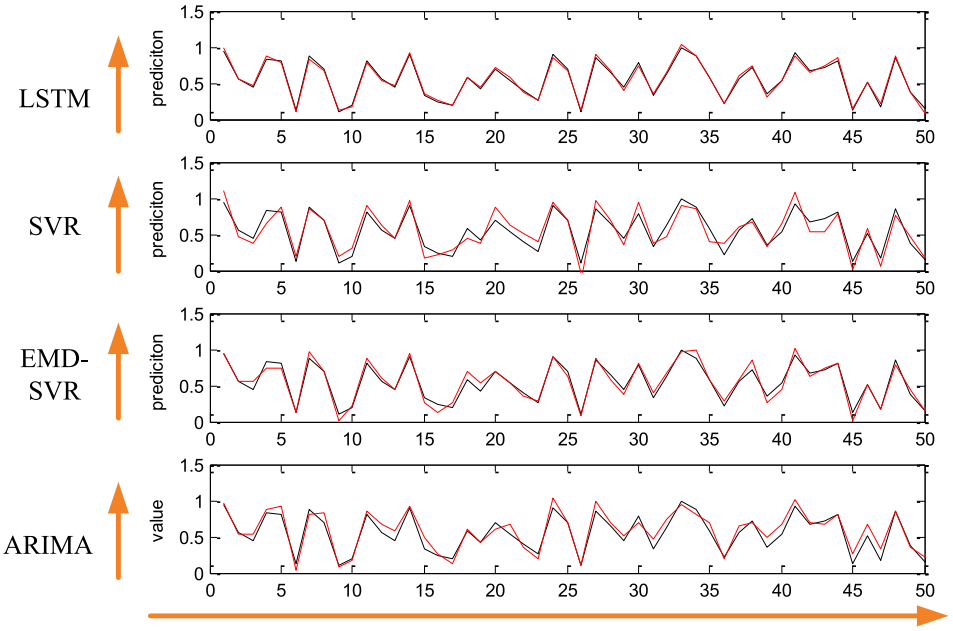

FIGURE 9. The prediction results based on LSTM compared with ARIMA, EMD-SVR, and SVR method (the black line is real value collected from IoT system, the red line is prediction value generated by ARIMA, SVR, EMD-SVR and LSTM neural network).

C. DISTRIBUTED PREDICTION RESULTS ANALYSIS FOR LSTM IN HADOOP CLOUD PLATFORM

MapReduce based on Hadoop studies distributed cluster data mining algorithm. The problem that can be solved has a common feature: tasks can be designed and decomposed into several sub-problems, and these sub-problems are relatively independent of each other, and will not be too constrained. After each node has processed these sub-problems in parallel, the task will be solved. And the distributed data mining algorithm designed needs to be aimed at the fragmented data sets, to ensure its independence, not to split the interrelated data, mining results and traditional serial algorithms to maintain consistency, and to be able to prove the consistency and validity of the experimental results.

In the simulation, the wind sensor data, air pressure sensor data, temperature data and relative humidity data mentioned above are selected as the main analysis targets. There is a data point in each sensor data. This paper divides it into two parts: training set and test set. The first data point is training data point, and the last data point is test data point. In testing, this paper will test predictions as the content of the test. At the same time, other existing better models are used to verify our model, compared with SVR, EMD-SVR and ARIMA three models.

Firstly, the sensor data decomposition part is processed by the ensemble empirical mode decomposition algorithm. Through many experiments, the size of the set in the ensemble empirical mode decomposition algorithm is set to 200. The temperature sensing data are decomposed into 8 sequences data and one remainder. Because the sequence data and remainder of IMF5 to IMF8 are relatively smooth, this paper combines these parts into a regression prediction. In this way, under the premise of guaranteeing the prediction accuracy, it can reduce part of the calculation overhead and improve the efficiency of the model. By calculating the correlation coefficients of each component and the original data, the threshold is set to 0.6, and the IMF1 with low correlation with the original data is removed. Secondly, in the data regression prediction part, According to the number of IMF sequence data selected, In the support vector regression model, the basic regression equation is $y = f(X^n)$. X^n is the input vector of n dimension and y is the regression prediction data. For variables in X, hoof selection is done by means of autocorrelation.

In this study, some mainstream regression prediction algorithms are selected for comparison and verification. The native LSTM is used to validate the generality and generalization ability of the short-term regression prediction model for IoT sensor data. The simulation results are shown in Figure 9.

As we can see from Figure 9, the black line is real value collected from IoT system, the red line is prediction value generated by ARIMA, SVR, EMD-SVR and LSTM neural network. Obviously, the prediction value of our LSTM network can better predict the real-time Internet of things data. The relative tracking error of LSTM network is small.

From the experimental data, we can see the following phenomena. First, it is obvious that LSTM performs much better than the other three hybrid algorithms. This is the rule for the data of one-step prediction, two-step prediction and three-step prediction. The results show that the composite model is more versatile and generalization than the single model for sensor data with complex and different modes. Secondly, the performance of LSTM model is better than

TABLE 5. Comparison and analysis of sensor data prediction results.

Step	Relative humidity sensors data forecasting results by different models							
	Single method				Hybrid method		Deep learning method	
	SVR		ARIMA		EMD-SVR		LSTM	
	MAPE	MSE	MAPE	MSE	MAPE	MSE	MAPE	MSE
1	0.0038	0.0614	0.3915	0.1105	0.2504	0.0310	0.2772	0.0247
2	0.3586	0.0651	0.4312	0.1194	0.2742	0.0439	0.2954	0.0110
3	0.3677	0.0724	0.4537	0.1310	0.2871	0.0487	0.3010	0.01546
	Relative temperature sensors data forecasting results by different models							
step	MAPE	MSE	MAPE	MSE	MAPE	MSE	MAPE	MSE
1	0.2735	0.1046	0.3075	0.1520	0.2107	0.02648	0.2347	0.0212
2	0.2809	0.1306	0.3242	0.1634	0.2255	0.0318	0.2576	0.0287
3	0.2817	0.1437	0.3518	0.1818	0.2108	0.0302	0.2849	0.0161

that of SVR model, which is due to eliminating the pattern aliasing phenomenon and further improving the performance of LSTM. Thirdly, compared with the traditional SVR method, the performance of the LSTM is improved. This shows that the linear regression prediction model is not suitable for sensor data with strong nonlinearity, high noise and non-stationary.

A summary of the above simulation results is shown in Table 5. The main reason is that the relative humidity sensor data are more volatile and less regular than the first two data, which affects the accuracy of prediction.

Based on the above analysis and experimental data, the following conclusions are drawn in this paper. Firstly, the sensor data affect the regression prediction performance of short-term regression prediction model. The weaker the fluctuation and the stronger the regularity is, the higher the prediction accuracy is. Secondly, compared with single regression prediction model, short-term regression prediction model has stronger adaptability and better generalization ability, so it can be better applied to regression prediction of IoT data. Thirdly, for dealing with the problem of non-linear regression prediction, the non-linear model has better performance than the linear regression prediction model.

For exception filters, Map stage only needs to determine whether the value of partition meets the length of business requirements. For business filters, Map stage calls the corresponding attribute value judgment function according to specific services to determine whether the value of value partition meets business requirements. For time filters, the same ID number is judged in Map stage. In the Reduce stage, the data are analyzed by comparing the same Key values, and in the similarity filter, the similarity judgment function can be invoked according to the specific value partition values (attribute values). In the overall design, data fragmentation in Map stage is mainly divided into the same Key values. Data with the same Key values are filtered and cleaned in Combiner and Reduce stages according to various principles.

We divide the temperature data set into verification machine and test set, and conduct a series of experiments

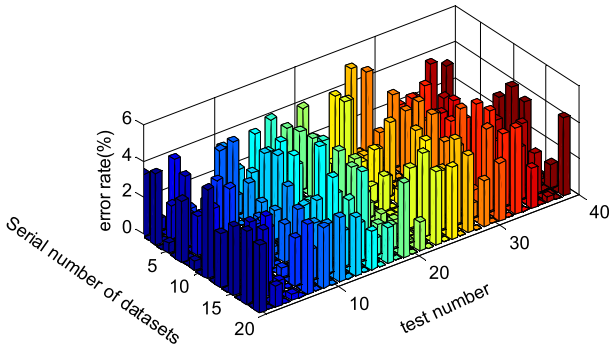

FIGURE 10. The comparison results under different dataset number and test number.

on different datasets that collected form IoT system. The experimental error rate is shown in Figure 10; we validated 20 datasets, each of which was tested 40 times. The overall error rate in each test shall not exceed 10%.

As we can see form Figure 10, Data mining system is used to get useful information that hidden in incomplete, noisy and chaotic data. SVM and LSTM prediction model based on cloud platform can use IoT data to classify and predict. This trained machine learning model can also be regarded as the potential and meaningful information and knowledge we extracted from the data, which has always improved the accuracy of the algorithm prediction.

D. ENERGY ANALYSIS AND APPLICATION OF HADOOP CLOUD PLATFORM

According to Shannon's information entropy in information theory, the bigger the information entropy of a system is, the higher the ability to realize implicit parallelism is. The calculation process of implicit parallelism is reversible, and its entropy value is invariable in the transition process, and it does not need to consume energy. Improving implicit parallelism and the calculation process is green.

The energy needed to change 1 bit data in memory is S = K (ln2-ln1) = Kln2. In the process of deep learning,

sorting attributes, it is necessary to change 1 bit content in memory, so a work process is needed, and energy needs to be obtained from the outside world. The computational complexity and communication complexity of the LSTM algorithm are related to the implementation complexity of the specific algorithm. The lower limit of the algorithm complexity is proportional to the required energy, so reducing the complexity of the algorithm can reduce the energy consumption.

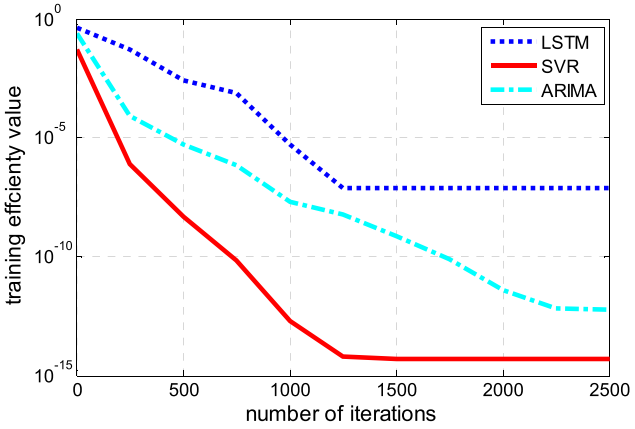

FIGURE 11. The comparison results of training efficiency value.

As shown in Figure 11, LSTM has a larger descent gradient and better training efficiency in the process of data prediction and regression. Compared with other corresponding SVR and ARIMA regression models, the training effect is better. Especially, by adding these thresholds, there are only a few linear operations and no multiplication operations in the process of CT-1 to CT propagation. The advantage is that its gradient flow is not interfered too much, that is, the gradient propagation is very smooth throughout the process, which solves the problem of gradient explosion or gradient disappearance in RNN.

To implement the application of data mining system in the IoT, we need to compare the running time under cloud platform and conventional conditions, as shown in Figure 12.

As shown in Figure 12, with the increase of IoT data volume, training time increases correspondingly, but neither of them increases linearly. When the amount of data is less than 1000, the cloud platform consumes more time than the algorithm without cloud platform; when the data volume is more than 1000, the cloud platform consumes less time than the algorithm without cloud platform; the time difference between the two consumes is larger, so the cloud platform is more suitable for large-scale data management. Data mining system of the Internet of Things uses cloud platform as a long-term storage database.

In Figure 12, through analysis of information entropy and information gain, the intrinsic relationship between algorithm complexity and energy is obtained. It is pointed out that the computational complexity and communication complexity of parallel data mining affect the efficiency. It is concluded that

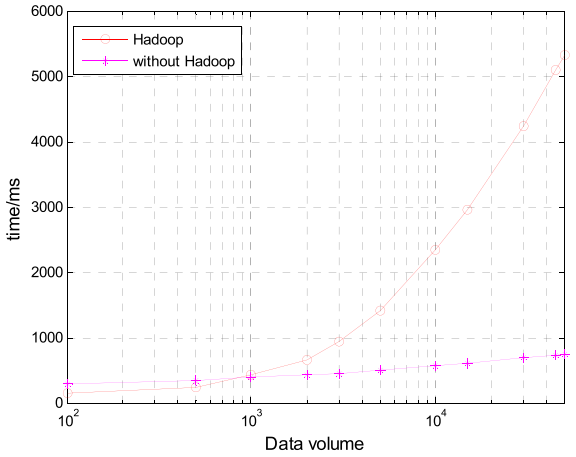

FIGURE 12. The comparison results with Hadoop and without Hadoop.

improving the implicit parallelism of the system is to reduce the running time under the cloud platform.

Parallel data mining can be flexibly migrated to data mining under cloud platform in many aspects. Improving the implicit parallelism of the system can improve efficiency of the algorithm; meet the requirements of low-carbon life and the needs of a conservation-oriented society. From the point of view of physical essence, it is concluded that the higher the complexity, the greater difference of entropy; the greater difference of entropy, the greater change of energy.

V. CONCLUSION

This paper analyses and studies the classification and characteristics of Internet of Things information, and discusses the construction and application of Hadoop cloud platform. This paper mainly carries out from two aspects. One is to design the system architecture of the Open Platform for Data Simulation Resources of the IoT and design the key modules. A platform for data simulation resources of the IoT is built to provide the running environment and external services for the sensor data simulation model established. On the other hand, it is the key method to study the simulation data model based on IoT sensors. That is, based on the research environment of the Internet of Things built by the existing laboratories, collect the data of sensors, analyze and study the characteristics of sensors in the IoT, and design the key algorithms for data simulation according to the characteristics of sensors in the IoT. Extensive simulations are executed to validate the remarkable nature of the proposed model and in Hadoop platform, in terms of prediction accuracy and training efficiency under different working condition. The energy analysis and application of the IoT algorithm will be conducted in the future.

REFERENCES

[1] J. H. Ku, "A study on prediction model of equipment failure through analysis of big data based on RHadoop," *Wireless Pers. Commun.*, vol. 98, no. 4, pp. 3163–3176, 2018.

[2] M. M. Rashid and I. K. J. Gondal, "Dependable large scale behavioral patterns mining from sensor data using Hadoop platform," *Inf. Sci.*, vol. 379, pp. 128–145, Feb. 2017.
[3] C. Yuan, L. Peng, and Z. Yuzhuo, "Parallel processing algorithm for railway signal fault diagnosis data based on cloud computing," *Future Gener. Comput. Syst.*, vol. 88, pp. 279–283, Nov. 2018.
[4] L. Yu, X. Wu, and Y. Yang, "An online education data classification model based on Tr_MAdaBoost algorithm," *Chin. J. Electron.*, vol. 28, no. 1, pp. 21–28, Jan. 2019.
[5] M. Gohar, S. H. Ahmed, M. Khan, N. Guizani, A. Ahmad, and A. U. Rahman, "A big data analytics architecture for the Internet of small things," *IEEE Commun. Mag.*, vol. 56, no. 2, pp. 128–133, Feb. 2018.
[6] P. U. Ferraro, G. Roscigno, G. Cattaneo, and R. Giancarlo, "Informational and linguistic analysis of large genomic sequence collections via efficient Hadoop cluster algorithms," *Bioinformatics*, vol. 34, no. 11, pp. 1826–1833, 2018.
[7] J. B. Bibal and D. Dejey, "An auto-scaling framework for heterogeneous Hadoop systems," *Int. J. Cooperat. Inf. Syst.*, vol. 26, no. 4, 2018, Art. no. 1750004.
[8] S. Rashmi and A. Basu, "Resource optimised workflow scheduling in Hadoop using stochastic hill climbing technique," *IET Softw.*, vol. 11, no. 5, pp. 239–244, 2017.
[9] M. C. Chen, S. Q. Lu, and Q. L. Liu, "Global regularity for a 2D model of electro-kinetic fluid in a bounded domain," *Acta Mathematicae Applicatae Sinica, English Ser.*, vol. 34, no. 2, pp. 398–403, 2018.
[10] S. Lu, M. Chen, and Q. Chen, "On regularity for an Ericksen–Leslie's parabolic-hyperbolic liquid crystals model," *ZAMM-J. Appl. Math. Mech./Zeitschrift Für Angewandte Mathematik Und Mechanik*, vol. 98, no. 9, pp. 1574–1584, 2018.
[11] I. Ganchev, Z. Ji, M. O'Droma, and L. Zhao, "Smart recommendation of mobile services to consumers," *IEEE Trans. Consum. Electron.*, vol. 63, no. 4, pp. 499–508, Nov. 2017.
[12] R. R. Expósito, J. González-Domínguez, and J. Touriño, "HSRA: Hadoop-based spliced read aligner for RNA sequencing data," *PLoS ONE*, vol. 13, no. 7, 2018, Art. no. e0201483.
[13] D. Kesavaraja and A. Shenbagavalli, "Framework for fast and efficient cloud video transcoding system using intelligent splitter and Hadoop MapReduce," *Wireless Pers. Commun.*, vol. 102, no. 3, pp. 2117–2132, 2018.
[14] J. Jin, J. Luo, Y. Li, and R. Xiong, "COAST: A cooperative storage framework for mobile transparent computing using device-to-device data sharing," *IEEE Netw.*, vol. 32, no. 1, pp. 133–139, Jan./Feb. 2018.
[15] Z. Dou, I. Khalil, A. Khreishah, and A. Al-Fuqaha, "Robust insider attacks countermeasure for Hadoop: Design and implementation," *IEEE Syst. J.*, vol. 12, no. 2, pp. 1874–1885, Jun. 2018.
[16] G. Cattaneo, U. F. Petrillo, R. Giancarlo, and G. Roscigno, "An effective extension of the applicability of alignment-free biological sequence comparison algorithms with Hadoop," *J. Supercomput.*, vol. 73, no. 4, pp. 1467–1483, 2017.
[17] B. Kong, S. Liu, J. Yin, S. Li, and Z. Zhu, "Demonstration of application-driven network slicing and orchestration in optical/packet domains: On-demand vDC expansion for Hadoop MapReduce optimization," *Opt. Express*, vol. 26, no. 11, pp. 14066–14085, 2018.
[18] C. T. Yang, J. C. Liu, and S. T. Chen, "Implementation of a big data accessing and processing platform for medical records in cloud," *J. Med. Syst.*, vol. 41, no. 10, p. 149, 2017.
[19] M. M. Rathore, H. Son, A. Ahmad, A. Paul, and G. Jeon, "Real-time big data stream processing using GPU with spark over Hadoop ecosystem," *Int. J. Parallel Program.*, vol. 46, no. 3, pp. 630–646, 2017.
[20] R. Chaudhary, G. S. Aujla, N. Kumar, and J. J. P. C. Rodrigues, "Optimized big data management across multi-cloud data centers: Software-defined-network-based analysis," *IEEE Commun. Mag.*, vol. 56, no. 2, pp. 118–126, Feb. 2018.
[21] X. Ma, X. Fan, J. Liu, H. Jiang, and K. Peng, "vLocality: Revisiting data locality for MapReduce in virtualized clouds," *IEEE Netw.*, vol. 31, no. 1, pp. 28–35, Jan./Feb. 2017.
[22] J. M. Luna, F. Padillo, M. Pechenizkiy, and S. Ventura, "Apriori versions based on MapReduce for mining frequent patterns on big data," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2851–2865, Oct. 2017.
[23] M. Khan, S. Din, S. Jabbar, M. Gohar, H. Ghayvat, and S. C. Mukhopadhyay, "Context-aware low power intelligent SmartHome based on the Internet of Things," *Comput. Electr. Eng.*, vol. 52, no. C, pp. 208–222, 2016.
[24] M. C. Ruiz, D. Cazorla, D. Pérez, and J. Conejero, "Formal performance evaluation of the Map/Reduce framework within cloud computing," *J. Supercomput.*, vol. 72, no. 8, pp. 3136–3155, 2016.
[25] A. Paul, A. Ahmad, M. M. Rathore, and S. Jabbar, "SmartBuddy: Defining human behaviors using big data analytics in social Internet of Things," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 68–74, May 2016.
[26] J. H. Um, S. Lee, T. H. Kim, C. H. Jeong, S. K. Song, and H. Jung, "Distributed RDF store for efficient searching billions of triples based on Hadoop," *J. Supercomput.*, vol. 72, no. 5, pp. 1825–1840, 2016.
[27] W. Zhang, L. Xu, Z. Li, Q. Lu, and Y. Liu, "A deep-intelligence framework for online video processing," *IEEE Softw.*, vol. 33, no. 2, pp. 44–51, Mar. 2016.
[28] K. Matsuzaki and R. Miyazaki, "Parallel Tree Accumulations on MapReduce," *Int. J. Parallel Program.*, vol. 44, no. 3, pp. 466–485, 2015.
[29] D. Chen, Y. Chen, B. N. Brownlow, P. P. Kanjamala, C. A. G. Arredondo, B. L. Radspinner, and M. A. Raveling, "Real-time or near real-time persisting daily healthcare data into HDFS and elasticsearch index inside a big data platform," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 595–606, Apr. 2017.
[30] D. Park, J. Wang, and Y.-S. Kee, "In-storage computing for Hadoop MapReduce framework: Challenges and possibilities," *IEEE Trans. Comput.*, to be published, doi: 10.1109/TC.2016.2595566.
[31] Y. Sun, H. Qiang, J. Xu, and G. Lin, "IoT-based online condition monitor and improved adaptive fuzzy control for a medium-low-speed maglev train system," *IEEE Trans. Ind. Informat.*, to be published, doi: 10.1109/TII.2019.2938145.
[32] M. Shi, Y. Tang, and J. Liu, "Functional and contextual attention-based LSTM for service recommendation in Mashup creation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 5, pp. 1077–1090, May 2019.
[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

YUANPAN ZHENG was born in Henan, China, in 1983. He received the Ph.D. degree from the Beijing Institute of Technology, in 2010. He has been a Teacher with the Zhengzhou University of Light Industry (ZZULI), since 2010, where he has also been an Associate Professor with the School of Computer and Communication Engineering, since 2012. In 2015, he became a Master Tutor of ZZULI. He has published more than 50 articles and holds two patents and three Software copyrights. His research interests include but not limited to numerical computation, algorithm design, big data, and cloud computing.

GUANGYU CHEN was born in Henan, China, in 1996. She received the bachelor's degree from the Zhengzhou University of Light Industry, in 2018, where she is currently pursuing the master's degree. During her master's degree, she holds one software copyright. Her research interests include data mining and machine learning.