Eng Int Syst (2021) 6: 371–378


**Engineering**
***Intelligent Systems***

# Simulation of Intelligent Internet of Things System Based on Machine Learning and Clustering Algorithm

**Liang Guo***

*School of Computer Science and Information Engineering, Anyang Institute of Technology, Anyang, Henan, 455000, China*

With the evolution of cloud computing technology, researchers are constantly examining ways to use network technology to simulate intelligent system operations, increase the number of practical application functions, and assist with the design of cloud computing platforms. The latter requires a sound understanding of computer science and engineering, to which research scholars have contributed several theories. The development of cloud computing network platforms has given users access to a great number of resources, which in turn has seen an increase in the number of users who use cloud computing network platforms for data processing. However, in the long term, the current structure of cloud computing network platforms will not be able to provide the quality that users expect from a network platform. In this paper, the particle swarm algorithm is analysed and its current shortcomings are addressed in order to improve its performance. The bad particles and parameters in the particle swarm are adjusted so that the algorithm can better meet the construction requirements of a cloud computing platform.

Keywords: machine learning; clustering algorithm; intelligent Internet of Things (IoT); system simulation

## 1. INTRODUCTION

This era of information explosion has seen the generation of an abundance of data, often leading to information overload that needs to be managed. Because people often need to quickly extract the information they need from a massive amount of data, data mining technology emerged and continues to be developed [1]. With the continuous advancement of science and technology, data mining technology is being applied in an increasing number of fields, two of which are medicine and e-commerce to the point where the latter has become an inextricable part of many people's lives [2–3]. The application of data mining technology provides people with more comprehensive data analysis and evaluation, making it more convenient for users to choose safe services in universities. By means of statistics pertaining to education data and analysis using related models, the relationship between different variables in these activities and the intensity of mutual influence can be determined [4].

When constructing a logistics network, data mining is an important part of the development of technology for the Internet of Things (IoT). The constant evolution of IoT technology has meant that traditional data mining technologies can no longer meet IoT requirements [5]. The continuous development of cloud computing informs the development of the IoT [6–7]. The establishment and development of a logistics network play an important role in promoting the development of product circulation. In the process of establishing the network, it needs to be applied to radio frequency identification technology, infrared sensing technology, positioning technology and various types of sensor equipment [8]. These technologies enable the logistics network to identify product performance, determine the logistics and transportation status, track logistics and transportation in real time, monitor logistics and transportation, and manage product circulation [9].

*Address for correspondence: Liang Guo, School of Computer Science and Information Engineering, Anyang Institute of Technology, Anyang, Henan, 455000, China, Email: djzxgl@126.com

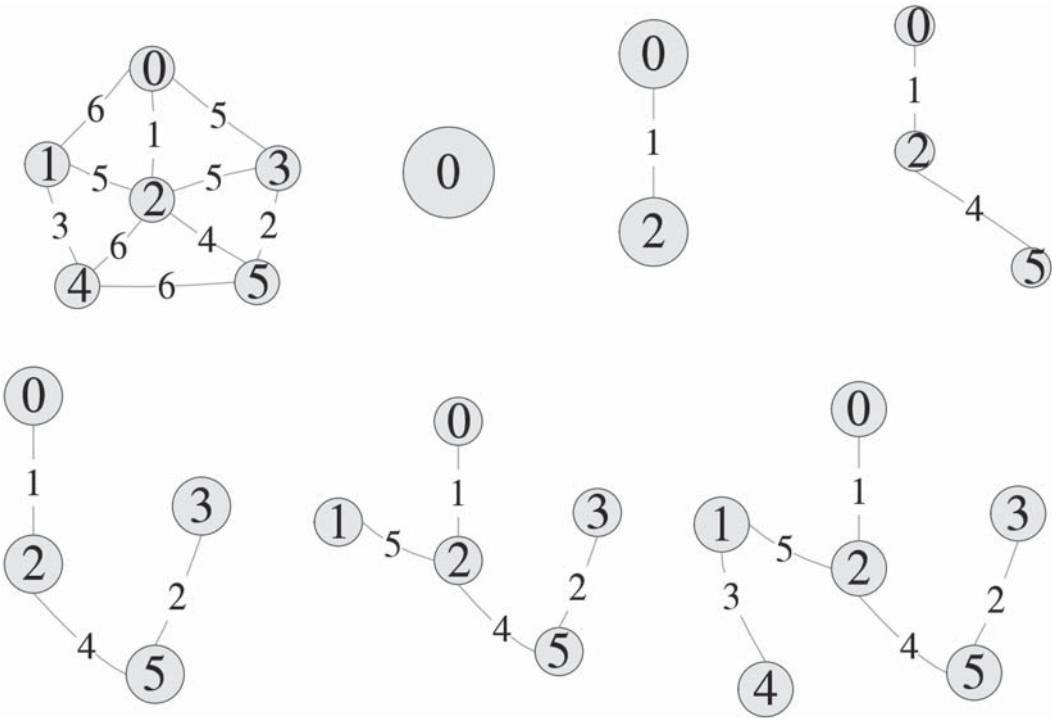

**Figure 1** The Plim algorithm constructs a minimum cost spanning tree.

## 2. THE THEORETICAL BASIS OF MACHINE LEARNING AND CLUSTERING ALGORITHMS

### 2.1 Improved K-Means Clustering Algorithm Based on Prim

Often, when a logistics network is being constructed, a large amount of data is generated, and these data can be complex. When a great amount of data has been accumulated over time, it is impossible to use traditional data statistics and data analysis methods to conduct pair evaluation. Hence, many scholars have researched and have begun using data mining technology [10]. In particular, due to the diversity and complexity of logistics data, many researchers have applied data mining technology to the data analysis of logistics management platforms.

The continuous development of data mining and data processing technology has made it necessary to classify data. The $k$-means clustering algorithm has become the most commonly-used algorithm for classifying data and determining related data. The final classification standard is based on the difference between the data. In the classification process, the data with less dissimilarity are classified under one category. After repeated calculations, data with smaller and smaller differences can be obtained. The smaller the difference between these data, the smaller the distance.

When using the $k$-means clustering algorithm, people generally apply the first spanning tree method to explain the classification of data. The construction process of this method is shown in the following figure:

When data is divided, the Prim algorithm and $k$-clustering algorithm are applied. By combining the advantages of these two algorithms, the clustering algorithm can better analyse the data. When selecting the initial center of the system, these two algorithms can provide researchers with a basis for making judgments [11–13]. The variables involved in these two algorithms and their definitions are given below.

Definition 3.1 The distance between the objects $x$ and $y$ of data analysis can be expressed as:

$$d(x, y) = \sqrt{(x^1 - y^1)^2 + (x^2 - y^2)^2 + \ldots (x^p - y^p)^2} \tag{1}$$

Definition 3.2 Data set cluster judgment [ center mi

$$m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \tag{2}$$

The definition of the error sum-of-squares function $E$ of data analysis can be expressed as:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \| X_{ij} - m_i \|^2 \tag{3}$$

### 2.2 ALDCK-Means Algorithm

When using different algorithms to calculate and analyze data, the algorithm is easily affected by the noise in the system and some abnormal values. If the noise and abnormal data are not handled well, the results are likely to have errors and influence [14–16] the final analysis result. Analysis and experience have led researchers to propose a method known as the LDC-means algorithm that combines the LDC algorithm and the $k$-means clustering algorithm.

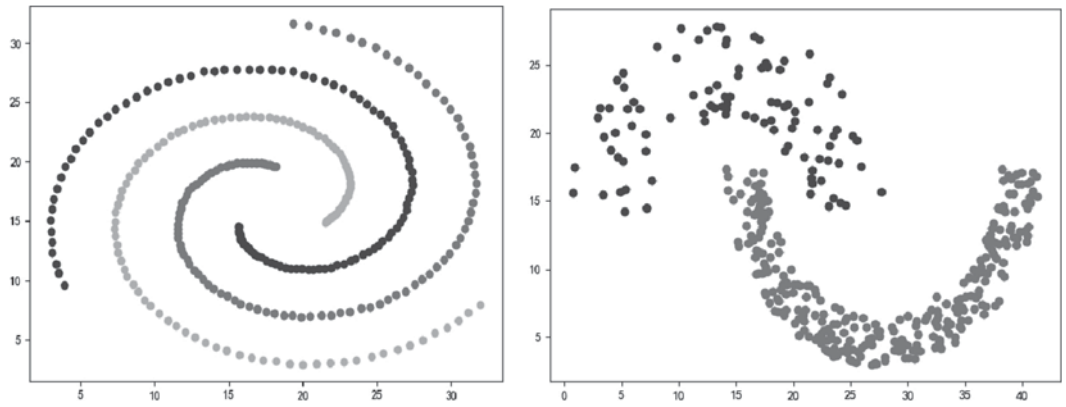

**Figure 2** LDCK-means clustering effect.

(1) When the $k$-clustering algorithm is used to select the clustering centers of data, it is likely to cause various changes in the final clustering results and the accuracy of the data calculation because this algorithm is very sensitive to the initial clustering centers of the system operation. When the cluster center changes, the corresponding clustering situation will also change, which will affect the results of the data analysis. The LDC algorithm is much better than the $k$-clustering algorithm when selecting the clustering center. This algorithm makes it easy to clearly distinguish the data with a larger distance, and it is easy to find the location from the center. In the process of clustering, the LDC algorithm can be used first to determine the location of the cluster center point of the data group, and then the $k$-clustering algorithm can be applied to cluster the data, which can indicate the cluster center to a certain extent When the point changes, the feasibility of the local optimal problem-solving strategy is improved.

When the LDC algorithm is used to determine the location of the center point of the cluster, people need to use a specific method to determine the specific situation of multiple cluster center points, and reflect the location of the center point in the decision diagram. People need to judge the location of the final cluster center with the naked eye when selecting the cluster center, which makes this algorithm very inefficient when applied. When selecting the cluster center, it can be found that the data group in the same category will contain extreme points showing the local density. The greater the distance between other data and the cluster center point, the smaller the extreme value of the local density. Clustered data can be either densely or sparsely clustered. If the clustered data is either dense or sparse, then local density cannot be used to describe the characteristics of the cluster center point. When selecting the position of the cluster center point, certain measures must be taken to remove the noise in the data group.

(2) When removing noise points, it is necessary to use filtering methods to exclude some data with relatively large density extreme points. When the $k$-clustering algorithm is used to select the clustering center, the noise points and abnormal data in the data group will cause certain errors in the result produced by the algorithm to determine the clustering center. When classifying data, it is very important that noise points be removed. The LDC algorithm can be used to filter the edge information in the data group and reduce the influence of noise on data classification.

(3) The process of $k$-means initial cluster center optimization based on LDC is as follows:

a) First, calculate the distance between the two data.

b) According to specific requirements, determine the data with higher density and the data with the longer distance between the data points; classify the data that does not meet the requirements as noise points, and delete the noise points from the data set.

c) Sort the data according to local density, and select the initial cluster center point of the system operation from the arranged data.

(4) In order to verify the feasibility of the combined clustering algorithm, this study uses experimental simulation to test the performance of the algorithm. The screenshots of the verification process are shown below in Figure 2.

When using the $k$-clustering algorithm to cluster data, the choice of $k$ value is directly related to the effect of data clustering. In order to ensure the effect of clustering, it is necessary to use a function to determine the value of $k$ so that the selected value of $k$ has the value of mathematical analysis. The parameters and definitions pertaining to the algorithm are as follows:

Definition 1: Dispersion between clusters

$$Disp = \frac{\sum_{i=1}^{k} Disp_i}{k} \tag{4}$$

Definition 2: Intra-cluster aggregation

$$Aggr = \frac{\sum_{i=1}^{k} Aggr_i}{k} \tag{5}$$

Definition 3: Clustering evaluation value

$$E = \frac{Aggr_k - Aggr_{k-1}}{Disp_k - Disp_{k-1}} \tag{6}$$

**Table 1** Iris data set Disp, Aggr and E list.

| | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Disp | 5.63 | 6.56 | 8.26 | 9.8 | 11.24 | 13.14 | 15.03 | 21.54 | 33.52 | 73.81 |
| Aggr | 0.92 | 0.983 | 1.05 | 1.12 | 1.24 | 1.33 | 1.48 | 1.85 | 2.40 | 2.99 |
| E | | 14.76 | 25.37 | 21.99 | 12.00 | 21.11 | 12.6 | 17.59 | 21.78 | 68.28 |
| | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | |
| Disp | 435.01 | 464.30 | 510.33 | 665.32 | 800.54 | 990.67 | 1295.46 | 1548.204 | 2089.654 | |
| Aggr | 67.38 | 77.32 | 92.81 | 123.93 | 138.69 | 152.66 | 190.64 | 216.98 | 235.62 | |
| E | | 2.947 | 2.972 | 4.980 | 9.161 | 13.610 | 8.025 | 9.595 | 29.048 | |

(1) The AK-means algorithm comprises these steps:

a) Randomly select several data objects in the data set as the initial clustering center of the system operation.

b) Using *k*-means clustering algorithm for the selected data, n clusters can be obtained through calculation.

c) Use formulas to calculate the degree of dispersion and aggregation of different data clusters, and use the letter E to indicate the evaluation value.

d) Assign the value of E to E0.

e) Calculate the distance between multiple cluster centers, merge the cluster centers that are closer together, determine a new cluster center point, and perform clustering again on the data group.

f) Use the formula to calculate the degree of dispersion and aggregation of the new data cluster, and use the letter E to indicate the evaluation value.

g) Compare the evaluation value in the previous calculation process with the new evaluation value. If certain conditions are met, you can go to the fourth step to iterate. If the conditions are not met, the calculation process of the algorithm ends here, and the result is the output.

(2) Effectiveness of AK-means algorithm

In order to test the effectiveness of the algorithm, experimental data was downloaded from the database for the analysis. The main analysis is shown in the table 1.

# 3. INTELLIGENT IOT SYSTEM DESIGN AND SIMULATION RESULT ANALYSIS

## 3.1 Physical Model Design of Intelligent IoT System

In order to better analyze the circulation of the logistics network, this study chose the logistics development data for a particular province as the basis for establishing the logistics model. The specific reasons are as follows:

(1) The economic development level of this province is relatively high, the population of each city is quite large, and the city's transportation network is relatively complete. In regard to road traffic, there are a great number of vehicles in many cities in this province, often producing traffic congestion.

(2) The area of this province is relatively small, there is no great distance between cities, and the cities are quite densely populated and well developed.

(3) This is a typical coastal province. The temperate monsoon climate affects the development of cities across the province. The characteristics of each season are predictable: the temperate monsoon climatic conditions produce high temperatures in summer and low temperatures in winter accompanied by high rainfalls.

The logistics development issues examined in this paper, and some new logistics development models have certain expressions in the development process of a certain province, so the data of this province is chosen as the basis for the model.

For the analysis, we assume that a certain goods supplier will transport goods from a certain area in Dalian to a certain area in Tieling City. These two areas are represented by the letters A and I respectively. Figure 3 below depicts a cargo transportation scenario using undirected graphs and where the speed is 50 km per hour.

The time transportation table and delivery distance between different cities are shown in Table 2 below.

## 3.2 Choice of Simulation System

### 3.2.1 Arena Basic Information

In order to verify the effectiveness of the logistics network operation, for this study, Arena software was chosen for the simulation and analysis of the specific operating conditions. Arena was developed by American researchers and has been in use for some time. The software offers researchers various tools for simulation analysis, enabling them to conduct simulation experiments successfully. The software can perform calculations and analyse the data input to the system, and shows users an animation video of the software communicating with other systems, providing a better understanding of the way the software works.

### 3.2.2 Arena's Processing and Analysis of Data

The software will also provide users with reliable data processors according to their different needs, and users can process and analyze data using the appropriate software modules. When analyzing data, the software's built-in function will

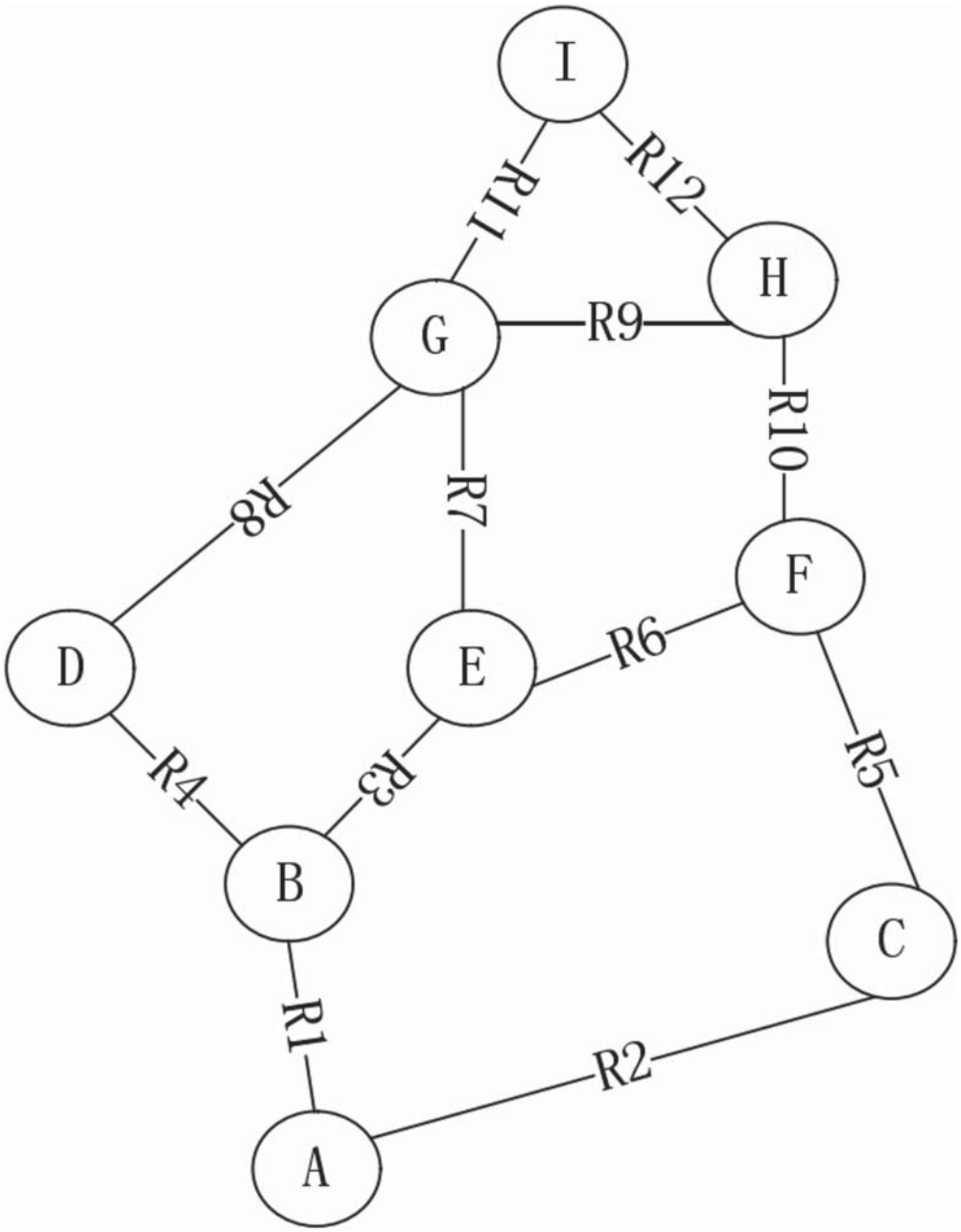


**Figure 3** Undirected map of a province.

**Table 2** Distance and time between distribution nodes.

| Between distribution nodes | Distance (km) | Time (hour) |
|---|---|---|
| A—B | 262 | 5.24 |
| A—C | 595 | 11.9 |
| B—D | 131 | 2.62 |
| B—E | 90 | 1.8 |
| C—F | 193 | 3.86 |
| E—F | 94 | 1.88 |
| D—G | 202 | 4.04 |
| E—G | 89 | 1.78 |
| F—H | 136 | 2.72 |
| G—I | 70 | 1.4 |
| G—H | 61 | 1.22 |
| H—I | 111 | 1.22 |

**Table 3** Experimental parameter settings.

| Parameter name | Parameter value |
|---|---|
| Sensor node communication range/m | 50 |
| MAC protocol | IEEE 802.15. 4 |
| Number of 6LoWPAN nodes | 40–320 |
| Network topology | Star network |
| Simulation area/m x m | 100 x 100 |

perform distribution fitting analysis and parameter estimation on the data. As long as users import the data to be processed into the software, the software will automatically analyze the fitting quality level of the distribution data, and the software will also select the most suitable distribution function for the analysis of the data.

#### 3.2.3 Lab Environment

The simulation software used in this experiment is Arena 14.00 version, which needs to be downloaded and installed locally on the computer to start the formal simulation experiment.

#### 3.2.4 Experimental Design

In this study, the specific conditions of the transport sector were not taken into consideration in the logistics plan, and excluded the impact of some natural disasters. The main aim of this experiment and the analysis focused on the routes for the distribution and delivery of goods. The experiment consisted of two phases: the static logistics situation which focuses on the impact, on the operation of the logistics network, of the distance involved in the delivery of goods In the second phase, the circulation of items is examined on the premise of analyzing the IoT technology.

### 3.3 Experimental Parameter Design

The parameters of the simulation experiment involved in this research are shown in Table 3 below.

### 3.4 Analysis of Results

The results of the experimental analysis are shown in Figure 4 below.

Compared to other countries, China has been relatively slow in developing data mining technology and its offshoots. Although China has lagged behind foreign countries in terms of technological development, research studies show that many professionals in this country have for quite some time been applying data mining technology in the domains of logistics and transportation. Because many foreign countries have done research on technological and scientific innovations and developments very early, many industries have realized the importance of data mining technology and the early establishment of a database. Numerous companies and several well-known tertiary education institutions have established their own research departments where they are deploying data mining technology and are engaged in ongoing research on the application of data mining. In recent years, the continuous progress and development of science and technology in China, has seen the widespread application and improvement of data mining technology. Many industries have realized the importance of data mining technology and are using it to promote and develop their companies. The fifth document points out the impact of the arrival of the information explosion era on people's lives and the development of various industries. The main point raised is that the excessive growth of information and the large amount of storage have made the analysis and evaluation of information difficult. Therefore, Effective analysis of information has become an important problem for people to develop.

An in-depth study of cluster analysis and related classification rules provides a foundation for research on data mining technology, and the data analysis for goods circulation has also been developed. After the data is stored in the system, the data will be preliminarily screened and processed according to the system's program, and finally the data from different data sources will be stored in the system according to the different standards established by the system's program regulations, forming a basis for the analysis of a database. When the data reaches the processing layer, this layer will extract the data from the database and divide the modules according to the behaviors represented by the data. After the different behavior modules have been divided, the cluster analysis technology is applied to analyze the students from the perspective of different behaviors. Clustering makes it easy for users to find the information they seek.

## 4. CONCLUSION

When constructing a logistics network, data mining is an important aspect of the development of IoT technology [17]. Because this technology is constantly evolving, traditional data mining technologies can no longer meet the requirements of the IoT [18], the development of which is directed by the continuous development of cloud computing. The establishment and development of the logistics network play an important role in promoting the development of product circulation. In the process of establishing the network, it needs to be applied to radio frequency identification technology, infrared sensing technology, positioning technology and various other types of sensor equipment [19]. These technologies enable the logistics network to identify product performance, determine the status of logistics and transportation and track them in real time, monitor logistics and transportation, and

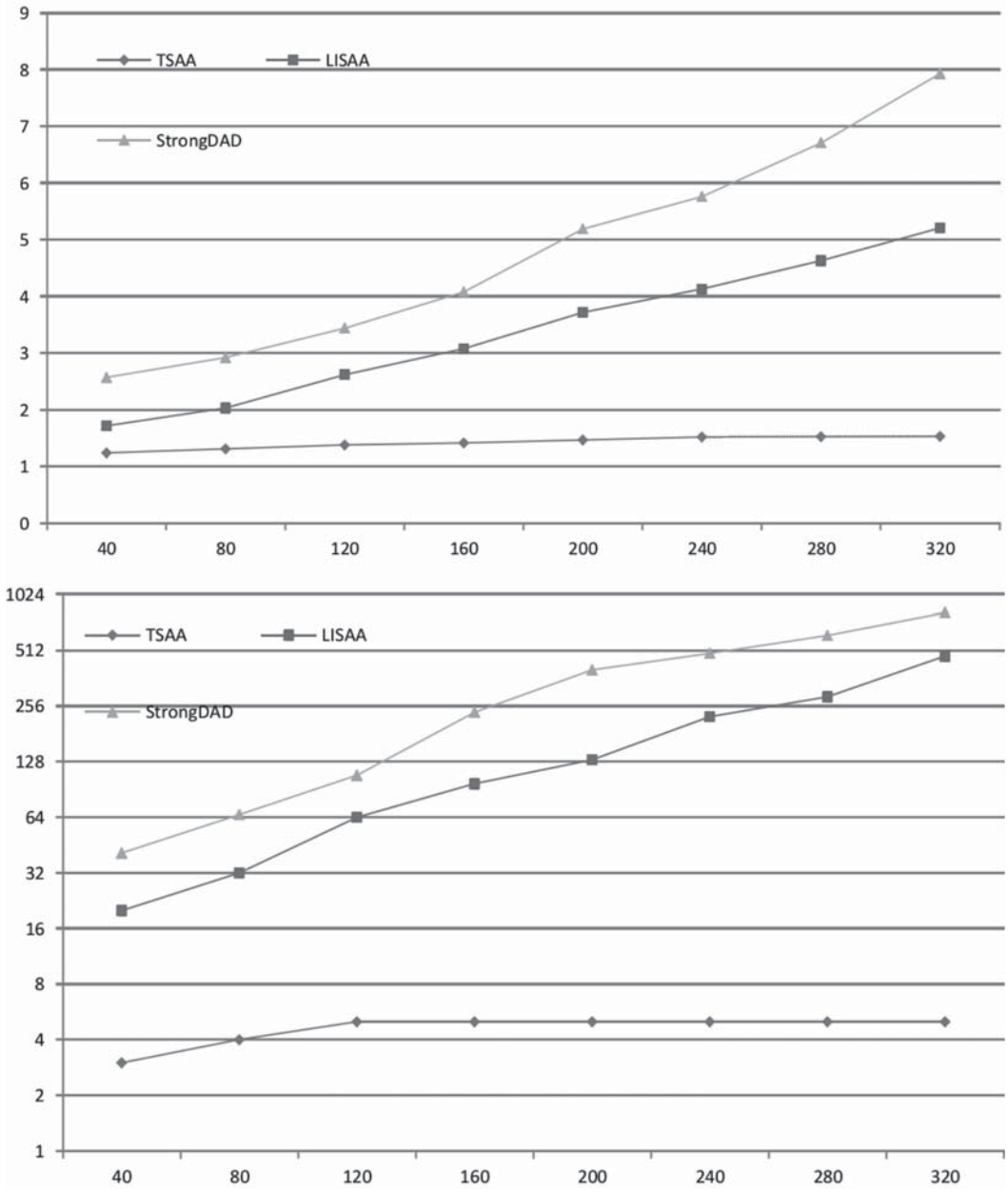


**Figure 4** Comparison of duplicate address detection overhead.

manage product distribution. When developing logistics networks, radio frequency technology can assist with the development of Internet technology. During the distribution of goods, radio frequency technology can help to track their delivery [20].

## REFERENCES


1. M. Minsky and S. Papert. Perceptrons: an introduction to computational geometry. *MIT Press*, Cambridge 78(3) (1969), 780–782.
2. K. Mistry, L. Zhang, SC. Neoh, CP. Lim, B. Fielding. A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition. *IEEE Trans Cybern* 47(5) (2016), 1496–1509.
3. S. Mitra and T. Acharya. Gesture recognition: a survey. *IEEE Trans Syst Man Cybern Part C* 37(3) (2007), 311–324.
4. A. Mollahosseini, D. Chan, MH. Mahoor. Going deeper in facial expression recognition using deep neural networks. In: *IEEE winter conference on applications of computer vision (WACV)*, Lake Placid, NY 11(3) (2016), 1–10.
5. XH. Nie, W. Wang, HY. Nie. Chaos quantum-behaved Cat Swarm Optimization Algorithm and its application in the PV MPPT. *Comput. Intell. Neurosci.* 31(4) (2017), 89–102.
6. X. Peng, Z. Xia, L. Li, X. Feng. Towards facial expression recognition in the wild: a new database and deep recognition system. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) workshops* 26(6) (2016), 27–30.
7. J. Peters, D. Janzing, B. Schlkopf. Identifying cause and effect on discrete data using additive noise models. In: *Proceedings 13th international conference artificial intelligence and statistics* 10(9) (2010), 597–604.

8. R. Ranjan, VM. Patel, R. Chellappa. HyperFace: A deep multitask learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(1) (2019), 121–135.
9. A.L. Rehab, M. Abd Elaziz, S. Lu. Chaotic opposition-based grey-wolf optimization algorithm based on differential evolution and disruption operator for global optimization. *Expert Syst. Appl.* 108(1) (2018), 1–27.
10. A. Samara, L. Galway, R. Bond, H. Wang. Affective state detection via facial expression analysis within a human–computer interaction context. *J Ambient Intell. Human Comput.* 10(8) (2019), 2175–2184.
11. C. Shan, S. Gong, P.W. McOwan. Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis. Comput.* 227(6) (2009), 803–816.
12. C.C. Shu, FW. Tsai. Computational intelligence based on the behavior of cats. *Int. J Innov. Comput. Inf. Control.* 3(1) (2007), 163–173.
13. I. Song, H. Kim, P.B. Jeon. Deep learning for real-time robust facial expression recognition on a smartphone. In: *IEEE international conference on consumer electronics (ICCE)*, Las Vegas, NV 53(2) (2014), 564–567.
14. A.N. Sreevatsan, K.G. Sathish Kumar, S. Rakeshsharma, M.M. Roomi. Emotion recognition from facial expressions: a targetoriented approach using neural network. In: *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing* 28(4) (2004), 1–6.
15. S.C. Tai and K.C. Chung. Automatic facial expression recognition using neural network. In: *IEEE Region 10 Conference—TENCON* 25(3) 2007, 689–699.
16. R.R. Walecki, O. Deep structured learning for facial expression intensity estimation. *Image Vis. Comput.* 25(9) (2017), 143–154.
17. D. Wang, C. Otto, A. Jain. Face search at scale: 80 million gallery. *Comput. Res. Repos.* 15(7) (2015), 1358–1366.
18. M. Wozniak and D. Połap. Adaptive neuro-heuristic hybrid model for fruit peel defects detection. *Neural Netw.* 98(4) (2018), 16–33.
19. M. Wozniak and D. Połap. Bio-inspired methods modeled for respiratory disease detection from medical images. *Swarm Evol. Comput.* 16(2) (2018), 1016–1025.
20. X.S. Yang. Engineering optimization: an introduction with metaheuristic applications. Wiley, Hoboken 25(3) (2010), 356–366.