EAI Endorsed Transactions
On Scalable Information Systems
Research Article EAI.EU

Data Mining Algorithm Based on Fusion Computer Artificial Intelligence Technology

Yingqian Bai[1,*], Kepeng Bao[2], Tao Xu[2]

[1] Department of Track Engineering, ShaanXi Railway Institute, Weinan 714000, Shaanxi, China
[2] School of Mechanical and Electrical Engineering, Xi'An Polytechnic University, Xi'an 710048, Shaanxi, China

Abstract

INTRODUCTION: The paper constructs a massive data mining model of distributed spatiotemporal databases for the Internet of Things. Then a homologous data fusion method based on information entropy is proposed. The storage space required by the tree structure is reduced by constructing the data schema tree of the merged data set. Secondly, the optimal dynamic support degree is obtained by using a neural network and genetic algorithm. Frequent items in the Internet of Things data are mined to achieve the normalization of the clustered feature data based on the threshold value. Experiments show that the F-measure of the data mining algorithm improves the efficiency by 15.64% and 18.25% compared with the kinds of other literatures respectively. RI increased by 21.17% and 26.07%, respectively.

Keywords: Artificial intelligence, Data mining, Information entropy, Data schema tree, Neural network

Received on 22 August 2023, accepted on 06 October 2023, published on 16 October 2023

doi: 10.4108/eetsis.3779

*Corresponding author. Email: 030203101@163.com

1. Introduction

With the rapid development of the Internet of Things industry, many smart devices are connected to the Internet of Things to monitor various objects in the natural environment. The central monitoring targets are traffic facilities, buildings, lakes, etc. [1]. Due to the various types and formats of data access, the scale of IoT data continues to increase, and the types and formats of data are increasingly complex. In addition, due to high dynamic heterogeneity, complex spatiotemporal characteristics and incompleteness of iot data, its practical application is minimal. These characteristics significantly increase the difficulty of mining practical knowledge from massive data. Existing data analysis techniques cannot be directly applied to IoT [2]. Literature [3] analyzes current network and computer technology development. The distributed privacy protection idea is introduced, and cluster analysis is carried out—research on data mining and privacy protection methods of personal information based on big data. The existing method can improve the accuracy of clustering results to ensure users' privacy, but its robustness is not high. The two-step discretization method is used to solve large-scale problems efficiently, and parallel computing technology is used to improve the running speed of the algorithm. An example analysis proves the effectiveness of this method in large-scale data processing [4]. Although this method dramatically improves the data processing speed, its scalability is not high. Literature [5] uses a clustering algorithm to screen the data that meets the conditions and obtains the data type distance calculation result based on data structure and similarity. Literature [6] revises the data model based on the CATS tree to make incremental mining easier. This makes the implementation efficiency of this method significantly improved. However, as the scale of big data increases exponentially, the existing computing time and storage cost based on association rules seriously restrict its application in extensive data analysis. Improving parallel computing performance to better deal with large-scale data is an urgent problem to be solved.

MapReduce is a new distributed parallel computing architecture developed by Google for large-scale data computation. It is characterized by simple use, low cost,

good system scalability and load-balancing ability. It has crucial research significance in such aspects as the analysis and processing of big data. Therefore, a parallel current item set Mining algorithm Inc Mining PFP is proposed in the literature [7] in a big data environment. The idea of MapReduce is introduced to improve the performance of parallel analysis effectively. However, because this method uses the tree to store all the data, thus accelerating the incremental mining later, the calculation amount of this method is enormous. Literature [8] adopted a method of constructing a data pattern tree according to the ordering of data volume to solve this problem. These data items are classified according to the frequency of each transaction recorded. By classifying different transactions, the same transaction item in different transactions shares a tree node as much as possible. This reduces the space required for the resulting tree-like structure. However, the application scope and feasibility of the algorithm are minimal, so the tree structure space of the data graph required by the algorithm needs to be further improved. Although the above methods have achieved some results, in the big data environment, how to Reduce the space occupation efficiently Can tree, obtain the corresponding support threshold and accelerate the data exchange between Map and Reduce still needs further research. This paper improves the existing method by combining information theory and artificial neural networks [9]. A homologous data fusion method based on information entropy is proposed. Constructing a data schema tree for the fused data set reduces the storage space required for the tree-like structure. The optimal dynamic support degree is obtained by combining neural networks and genetic algorithms. The threshold value is used to mine frequent items in the Internet of Things data to achieve the normalization of the clustered feature data. Experiments show that this method can improve the accuracy of data clustering.

2. Data definition model system based on a cloud platform

Cloud computing is an emerging technology. Integration with the Internet of Things is a trend in the progress of human society. Two different methods of mining and recommendation are efficiently parallelized and distributed. This pattern divides the whole system into three basic levels. This hierarchical design idea makes the overall information processing of the Internet of Things more efficient. At the same time, its computing speed is also greatly improved.

2.1 Cloud support platform level

In addition to providing data computing functions, it also provides file storage space. Integrate third-party data mining algorithm services into this platform. Users can conduct commercial operations on the cloud platform developed by themselves. It can also be implemented through a cloud platform provided by a third party.

2.2 Data mining level

This level provides basic functionality for overall data collection. This level requires algorithm service management, a scheduling engine, and data parallelization architecture. It must also provide the necessary support for the data collection business on the cloud computing platform. This will significantly impact the overall strength of the Internet of Things.

2.3 Data collection cloud computing service layer

The interface format of its functional packaging is various externally. Ways such as object access protocol, XML or local application programming interface can be used as the external interface of cloud computing. The Internet of Things' fundamental function is to make people more convenient through information technology. And cloud computing is used to strengthen its service functions. The service level of cloud computing is actually to integrate the information in the following two levels. In addition, the cloud service layer can also support access to the query language, which makes it easier to convert the language during data processing.

3. Internet of Things extensive data mining method based on parallel association rules incremental mining algorithm

3.1 Build a data model tree

Construct the dimension regulation and data modelling tree for IoT big data to ensure that it matches the dynamic nature of IoT big data. When designing the big data mining algorithm of the Internet of Things, an information pattern tree must be used to analyse the user's behaviour [10]. The remaining data nodes are processed by the classification method. The data model tree shown in Figure 1 is then formed.

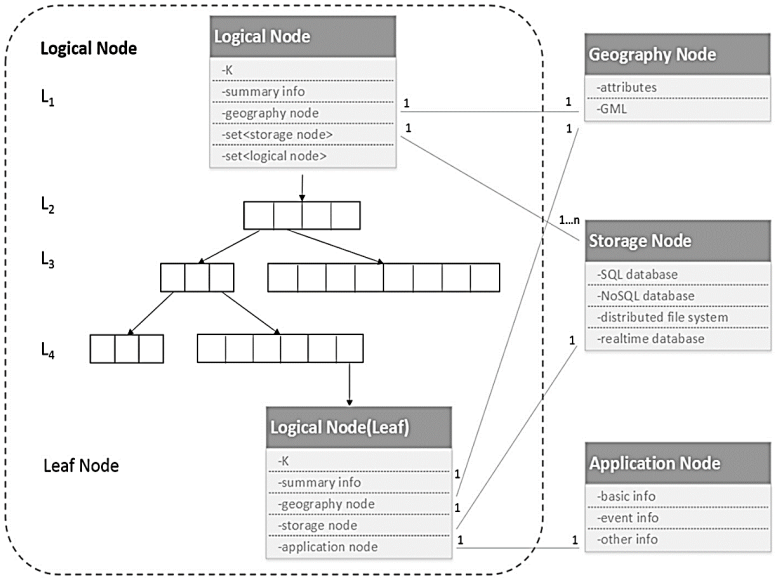


Figure 1. Data schema tree

3.2 Internet of Things characteristic information detection

A method of data graph based on multidimensional synthesis is proposed through the analysis of data graph. Given the existing problems in the Internet of Things big data, this project intends to use feature extraction technology to identify them [11]. Set the data set to be excavated to S. Then the paper sets the dimension of this data set to s. By analysing the properties of these data, a set E is obtained.

A subdomain L for data mining is established. The subspace is contained within the set of data attribute values and the data object in the subspace $r \in S$. From the outlier distribution of data, the neighbourhood (r, L) of the data object in its subspace is also non-uniform. The probability of an exception for an arbitrarily selected data object can be expressed in $Z_s(r, L)$. Use the following method to calculate the distance rate:

$$s_s = \frac{1}{Z_s(r, L)} \tag{1}$$

The distance is expressed as s. The standard distance γ from data s to data r can be calculated by the formula (2).

$$\gamma(r, s) = \sqrt{\frac{\sum_{s \in L} s(r, s)^2}{\sqrt{|L|^2}}} \tag{2}$$

The local discrete data has the characteristic of non-uniformity, so it must be described by the density of the discrete data and the approximation of the standard distance.

$$\mu = \frac{Z_s}{\gamma(r, s)} \tag{3}$$

The value of discrete characteristics μ is obtained by equation (3). The local discrete data assignment of the expected data in the local discrete data is obtained. This project proposes an information entropy detection algorithm for extensive Internet of Things data mining. Check the distribution u for an item in the data group U to be tested. According to the value probability function ξ, the information entropy $\psi(u)$ of data u is obtained:

$$\psi(u) = -\sum_{u \in U} \xi(u) \frac{\ln \xi(u)}{\ln \xi(u+1)} \tag{4}$$

All probed data contained in the data space is sorted in descending order. Then, according to the detection of the data set from the largest to the most minor order. Select several data with extensive information to detect the remaining data. The spacing between cluster centres is:

$$dist = \frac{|\alpha \cap \beta|}{|\alpha \cap \overline{\beta}| + |\overline{\alpha} \cap \beta| + |\alpha \cap \beta|} \tag{5}$$

α, β is the centre of the two clusters and is a randomly selected cluster. If the interval B between cluster centres is less than the specified threshold, the cluster centre must be selected again. The formula (5) calculation is repeated until all the calculation results exceed the specified threshold.

Calculate the distance from the data to the group centre in the detection data set. If there is no cluster, it is called a non-local exception. The existing data information in the classification cluster is all the feature data detected by the

information entropy algorithm. The detection data is normalized to increase the accuracy of data mining.

3.3 Standardized processing of feature data

Because the extracted feature quantity is highly correlated, some noise data will inevitably be mixed with the extracted feature quantity [12]. The data detected in the above operations are standardized to ensure the smooth progress of data analysis and processing in the future.

Because the detected data dimensions are different, it will adversely affect the results of extensive data mining in the Internet of Things, so the detected data must be processed according to the standard format. Follow the steps indicated in formula (6).

$$\eta' = \frac{\eta - \overline{\eta}}{V_\eta} \tag{6}$$

In the detected data, the result of data normalization η' must be calculated based on the characteristic mean η of the data and the characteristic standard deviation V_η of the data. The variance method is used to enhance the characteristics of data to ensure the accuracy of data mining. In addition, the result of data normalization can be obtained by the average deviation of data attributes from H_η.

$$\eta_i' = \frac{\eta - \overline{\eta}}{H_\eta} \tag{7}$$

The above calculation can improve the anti-interference performance of the method. The formula for calculating the mean η of the data attribute, the standard deviation V_η of the data attribute and the mean difference H_η of the data attribute is as follows:

$$\begin{cases} \overline{\eta} = \sum_n \eta \dfrac{1}{n} \\ H_\eta = \sum_n \dfrac{|\eta - \overline{\eta}|}{n} \\ V_\eta = \sqrt{\sum_n \dfrac{(\eta - \overline{\eta})^2}{n-1}} \end{cases} \tag{8}$$

The number of iterations is n. After normalizing the data, artificial intelligence technology is used to analyse the data to get the final data mining results.

3.4 Implementation of artificial intelligence data mining

This topic intends to use neural network technology to study big data in the Internet of Things. The primary model is the BP neural network with a three-layer transmission structure. Input normalized data into the neural network [13]. The weighting formula is as follows:

$$\varphi = \frac{1 - F_i}{\sum_{i=1}^{\psi} F_i} \tag{9}$$

The information entropy F_i of dimension i data is used to obtain the connection weight of the network (Figure 2 is quoted in Energies 2019, 12(21), 4124).

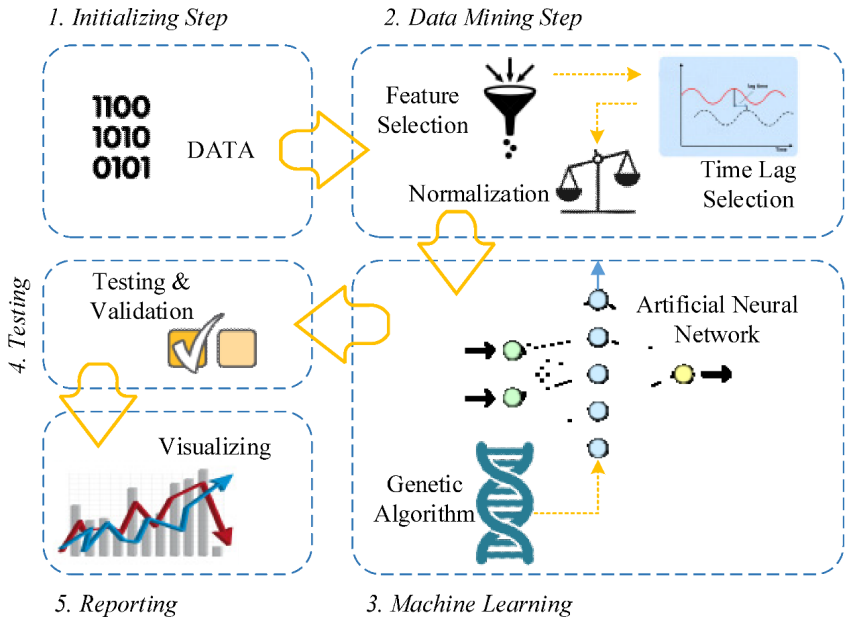


Figure 2. Data mining process based on neural network and genetic algorithm

A classifier based on network nonlinear classification ability and network structure is designed. A genetic algorithm solves the optimal solution. This improved approach to artificial intelligence technology can be connected to previous processing methods while ensuring nonlinear capabilities. This ensures the accuracy of data mining. The genetic algorithm is introduced into the neural network, and the genetic algorithm's hybridization operator and mutation operator are improved.

$$\begin{cases} \varepsilon_1 = \kappa\varepsilon_1 + (1-\kappa)\varepsilon_2 \\ \varepsilon_2 = \kappa\varepsilon_2 + (1-\kappa)\varepsilon_1 \end{cases} \tag{10}$$

$\varepsilon_1, \varepsilon_2$ is the linear combination of two data. The constant κ has a value between 0 and 1. If the constants do not change with time, the mixed operator is incompatible with solving. The parameters are changed under different iterations to improve the algorithm's performance. In this way, the gradual integration of big data in the Internet of Things is achieved. The change operator is modified in the process of data mining. Every random data j_k is likely to change. The value J_k of this data after a change is expressed randomly as follows:

$$J_k = \begin{cases} j_k = \Delta(t, UB - j_k),\ k \in (0, 0.5) \\ j_k = \Delta(t, j_k - LB),\ k \in (0.5, 1) \end{cases} \tag{11}$$

The change in the data is generated from the left and right adjacency LB, UB of the variable k and the return point of the function Δ. As the algebra t increases, the change value often approaches 0 indefinitely. The whole data set is retrieved through the above operation.

4. Simulation experiment

4.1 Acquisition of experimental data sets

The Internet of Things information containing 4000 2D vectors was selected as the experimental sample. The data points are broken down into sections, as shown in Figure 3.

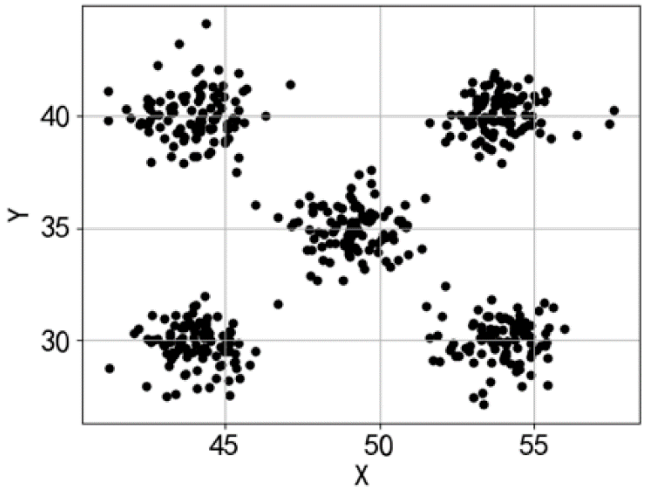


Figure 3. Data distribution map

As you can see from Figure 3, the data is divided into 20 categories. The ambiguity coefficient is 1.65 when the data points are clustered. A minimum is removed for each vector dimension to ensure that the vector dimensions are between 0 and 1. The algorithm divides the obtained result with the maximum value of each dimension so that the data of each dimension can be normalized [14]. The experimental data were sampled at random without repetition. This ensures that each segmented data block is the same, reducing the computational effort of the experiment. In the simulation test, the size of each subblock is set to 10,20,30,40.

4.2 Performance Specifications

The algorithms in references [15] and [16] are used in data mining experiments. F-measure and RI evaluate the method.

$$P_{ij} = \frac{n_{ij}}{n_i} \tag{12}$$

$$R_{ij} = \frac{n_{ij}}{n_j} \tag{13}$$

The RI index can also be used to evaluate data mining results accurately. By classifying the group data X, the actual classification results of the group C_{IN} and group C_T are obtained. Four groups, TP, TN, FP, and FN can be identified for a chaotic set of data points (x_i, x_j). These four groups are used to calculate the RI value:

$$\begin{cases} TP = \{(x_i, x_j) \mid x_i, x_j \in X,\ x_i, x_j \in C_{IN} \ \ and \ \ x_i, x_j \in C_T\}; \\ TN = \{(x_i, x_j) \mid x_i, x_j \in X,\ x_i, x_j \notin C_{IN} \ \ and \ \ x_i, x_j \notin C_T\}; \\ FP = \{(x_i, x_j) \mid x_i, x_j \in X,\ x_i, x_j \in C_{IN} \ \ and \ \ x_i, x_j \notin C_T\}; \\ FN = \{(x_i, x_j) \mid x_i, x_j \in X,\ x_i, x_j \notin C_{IN} \ \ and \ \ x_i, x_j \in C_T\}; \end{cases} \tag{14}$$

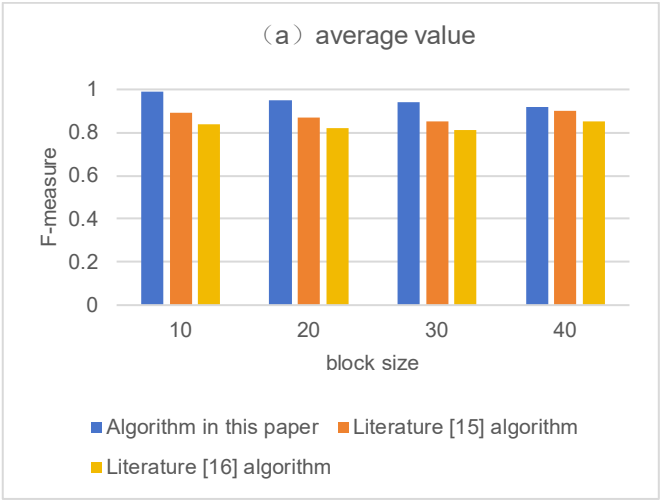
The RI calculation formula is obtained based on the above set:

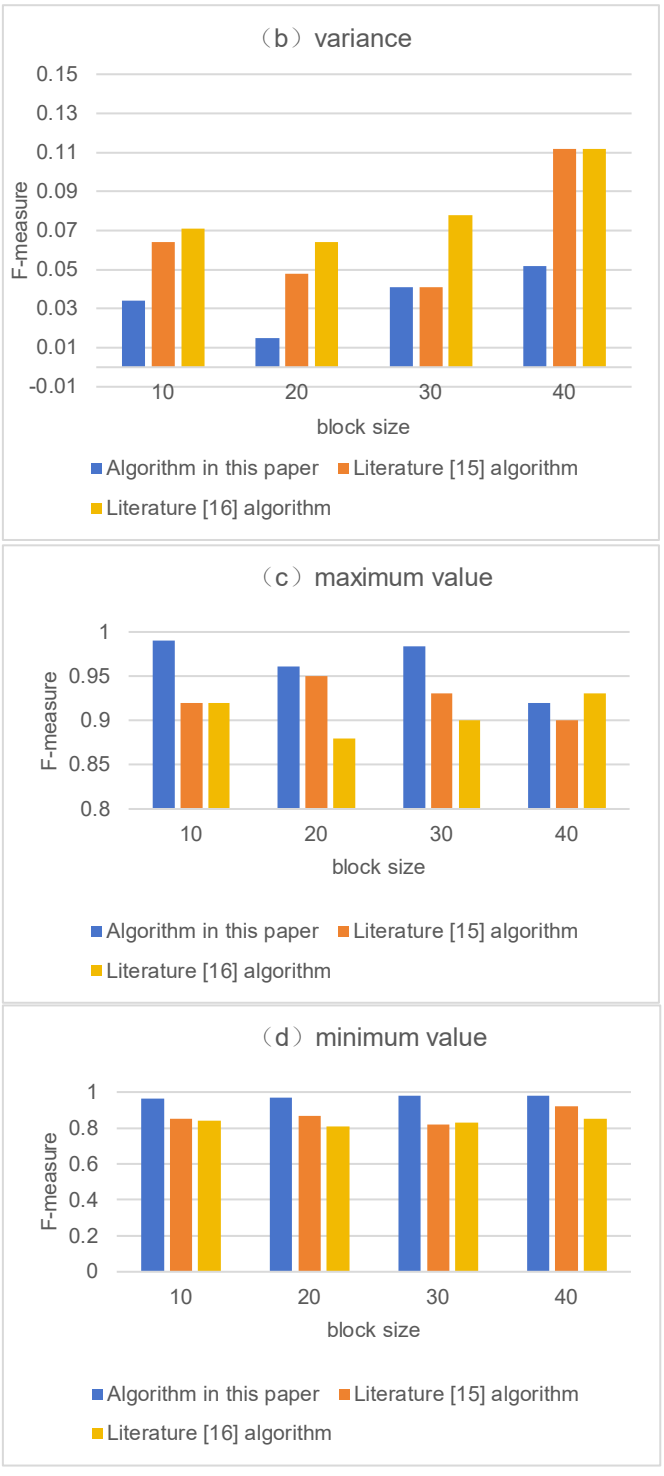
$$RI = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |RN|} \tag{15}$$

The similarity of the results of the two population sets is calculated by equation (15). Because the value of RI is in the range of 0 to 1, when the value of RI is closer to 1, it indicates that the data mining results are more similar to the actual results, and it also indicates that the clustering accuracy of the data mining algorithm is higher.

4.3 Algorithm performance

The mining algorithm designed in the paper is compared with the algorithms proposed in pieces of literature [15] and [16] on the experimental data set. F-metric and RI tests were carried out on the three methods, respectively. For better comparison, the paper sets the execution time of each algorithm to 20. The average value, variance, maximum, and minimum values of 20 experiments are analysed. The calculation of average value can measure the effect of data mining from the point of view of average value. The result pairs of the various algorithms for F-metrics are shown in Figure 4.





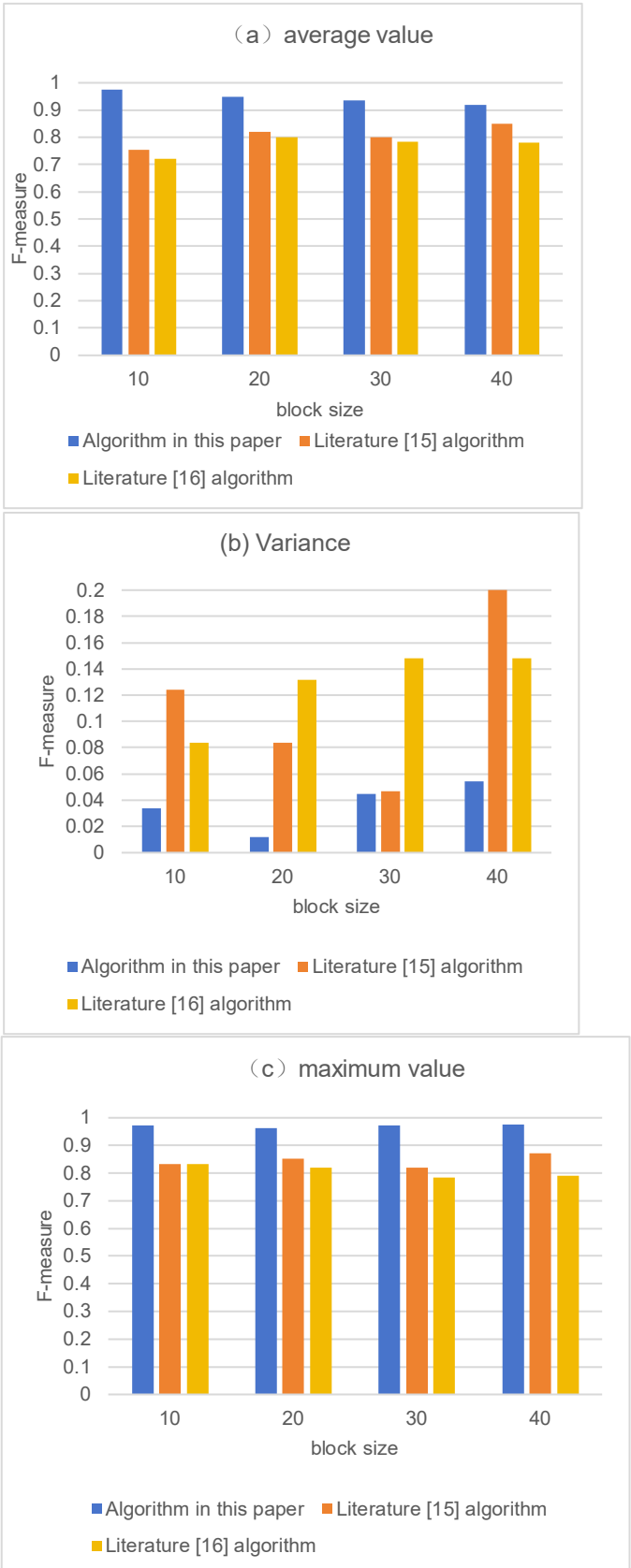


Figure 4. F-measure results

The data mining accuracy of the proposed algorithm increases by 15.64% at most compared with the literature [15]. Compared with the algorithm in the literature [16], the accuracy of data mining is increased by 18.25%. The algorithm presented in this paper shows promising results in the search process. Draw a comparison graph of the RI results of different algorithms, as shown in Figure 5.

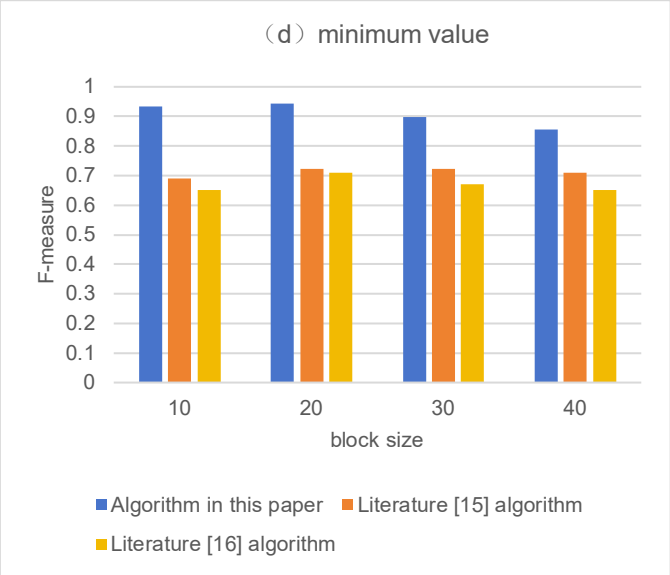


Figure 5. RI results

The analysis of the RI results of the three algorithms shows that the overall performance of the mining algorithm designed in this paper is better than that of the algorithms in the literature [15] and [16]. Especially in the case of 10%, the RI value of the proposed method increased by 21.17% and 26.07%.

Due to technical differences in extensive data analysis of the Internet of Things, the existing data mining methods are often brutal to process part of the data efficiently due to storage limitations. Artificial neural network technology combined with genetic algorithms solves the above problems. The imprecise transmission problem of clustering on big data sets is overcome through the special processing of data, and the efficiency of data mining is improved. Simulation shows that compared with the algorithms in literature [15] and [16], the algorithm designed in this paper has improved both F-results and RI results. It is proved that integrating artificial intelligence technology and the Internet of Things big data mining algorithm promotes the development of data mining technology.

5. Conclusion

The information entropy method is used to realize the fusion of the similarity in the data set. This reduces the complexity of the resulting data architecture and reduces the space footprint. Secondly, a dynamic threshold of support degree is obtained by combining an artificial neural network and genetic algorithm, and then related rules are explored to avoid redundant frequency pattern discovery. Finally, frequent itemset mining is migrated to MapReduce distributed computing platform to process large data sets in parallel, to realize parallel processing of massive data. Practice shows that the combination of AI technology and big data in the Internet of Things can effectively promote the development of data mining technology.

Acknowledgments.

This work was supported by Shaanxi Provincial Department of Education Service Local Special Research Program Project, Grant/Award. Research on Development Technology of Intelligent Drainage Big Data Platform Based on Microservice Architecture. Number: 22JC026.

References

[1] Mao, Y., Deng, Q., & Chen, Z. Parallel association rules incremental mining algorithm based on information entropy and genetic algorithm. Journal on Communications, 2021;42(5): 122-136.

[2] Heraguemi, K., Kadri, H., & Zabi, A. Whale optimization algorithm for solving association rule mining issue. International Journal of Computing and Digital Systems, 2021; 10(1): 333-342.

[3] Xu, W., Yuan, K., Li, W., & Ding, W. An emerging fuzzy feature selection method using composite entropy-based uncertainty measure and data distribution. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022; 7(1): 76-88.

[4] Chen, Q., Huang, M., Wang, H., & Xu, G. A feature discretization method based on fuzzy rough sets for high-resolution remote sensing big data under linear spectral model. IEEE Transactions on Fuzzy Systems, 2021;30(5): 1328-1342.

[5] Ghane, M., Ang, M. C., Nilashi, M., & Sorooshian, S. Enhanced decision tree induction using evolutionary techniques for Parkinson's disease classification. Biocybernetics and Biomedical Engineering, 2022; 42(3): 902-920.

[6] James, C. D., & Mondal, S. Optimization of decoupling point position using metaheuristic evolutionary algorithms for smart mass customization manufacturing. Neural Computing and Applications, 2021; 33(17): 11125-11155.

[7] Deng, G., & Fu, Y. Fuzzy rule based classification method of surrounding rock stability of coal roadway using artificial intelligence algorithm. Journal of Intelligent & Fuzzy Systems, 2021;40(4): 8163-8171.

[8] Hua, Y., Liu, Q., Hao, K., & Jin, Y. (2021). A survey of evolutionary algorithms for multi-objective optimization problems with irregular Pareto fronts. IEEE/CAA Journal of Automatica Sinica, 2021;8(2): 303-318.

[9] ZHAO, F., DONG, B., PAN, H., & SHI, A. A Mining Algorithm to Improve LSTM for Predicting Customer Churn in Railway Freight Traffic. Studies in Informatics and Control, 2023;32(2): 25-38.

[10] Qin, X., Zhan, P., Yu, C., Zhang, Q., & Sun, Y. Health monitoring sensor placement optimization based on initial sensor layout using improved partheno-genetic algorithm. Advances in Structural Engineering, 2021; 24(2): 252-265.

[11] Ferhat Taleb, S., Benalia, N. E. H., & Sadoun, R. Evolutionary algorithm applications for IoTs dedicated to precise irrigation systems: state of the art. Evolutionary Intelligence, 2023; 16(2): 383-400.

[12] Liu, W., Wang, J., Su, X., & Mao, Y. MR-DBIFOA: a parallel Density-based Clustering Algorithm by Using Improve Fruit Fly Optimization. Journal of Computers, 2022; 33(1): 101-114.

[13] Singh, L. K., Pooja, Garg, H., & Khanna, M. An IoT based predictive modeling for Glaucoma detection in optical coherence tomography images using hybrid genetic algorithm. Multimedia Tools and Applications, 2022; 81(26): 37203-37242.
[14] Fang, N., Fang, X., & Lu, K. Online incremental updating for model enhancement based on multi-perspective trusted intervals. Connection Science, 2022;34(1): 1956-1980.
[15] Ke, L., Li, M., Wang, L., Deng, S., Ye, J., & Yu, X. Improved swarm-optimization-based filter-wrapper gene selection from microarray data for gene expression tumor classification. Pattern Analysis and Applications, 2023; 26(2): 455-472.
[16] Thakkar, A., & Lohiya, R. A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions. Artificial Intelligence Review, 2022;55(1): 453-563.