Semantic Information Classification of IoT Perception Data Based on Density Peak Fast Search Clustering Algorithm

Lin Chen[1]*, Jinli Hu[2], Weisheng Wang[3]
The Internet of Things and Artificial Intelligence College,
Fujian Polytechnic of Information Technology, Fuzhou, 350001, China[1, 3]
Industrial Teaching and Research Cooperation Division,
Fujian Polytechnic of Information Technology, Fuzhou, 350001, China[2]

***Abstract*—In the rapidly developing field of the Internet of Things today, effective processing and analysis of perceptual data has become crucial. The perception data of the Internet of Things is usually large, diverse, and presents high-dimensional characteristics, which poses new challenges to data clustering algorithms. This study utilizes the K-center point algorithm to optimize the density peak fast search clustering algorithm, proposes a new clustering algorithm, and applies it to the research of semantic classification of perception data in the Internet of Things. Firstly, the K-center algorithm was used to optimize the clustering center optimization process of the density peak fast search clustering algorithm. Then, the optimized algorithm was applied to the automatic semantic classification model. Thus, a new automatic semantic annotation model for IoT aware data has been established. The research results showed that the classification accuracy of the proposed optimization algorithm was as high as 0.98, and the running stability of the automatic semantic annotation model optimized using this algorithm was as high as 0.99, with a running time as low as 1s. In summary, the automatic semantic annotation model built in this study can effectively improve the efficiency and accuracy of semantic classification, thereby providing more accurate and efficient data support for intelligent services.**

***Keywords*—Clustering algorithm; Internet of Things; perceived data; classification; peak density; semantic information**

I. Introduction

With the rapid development of the Internet of Things (IoT) technology, more and more devices are connected to the Internet, generating a large amount of sensory data. These data are not only large and diverse in volume, but also exhibit high-dimensional characteristics, bringing unprecedented challenges to effective information processing and analysis [1-2]. Especially in intelligent service domains such as smart city, smart home, health monitoring, etc., how to accurately and efficiently extract valuable semantic information from massive perceptual data has become an urgent problem to be solved [3]. Clustering by Fast Search and Find of Density Peaks (CFSFDP) algorithm has gained wide attention in the field of data science due to its superior performance, especially in identifying the cluster centers, which shows significant advantages. However, CFSDP algorithms still face problems such as inconsistent sample density and sensitivity to noisy data when dealing with IoT sensory data [4]. In addition, the K-center algorithm has better robustness in data classification, but its performance is limited by the choice of centroids [5]. However, existing research mainly focuses on data acquisition and transmission optimization, with insufficient exploration of efficient data processing and accurate semantic classification, failing to give full play to the potential of IoT data in intelligent applications. This study intends to fill this research gap by proposing a Fusion Clustering Algorithm Based on K-Centroids and Fast Search of Density Peaks (FCA-KCFSDP) based on the optimization of K-centroid algorithm, which aims to improve the accuracy and efficiency of semantic information classification of IoT sensory data. This algorithm not only improves the stability and operational efficiency of the classification model by optimizing the clustering center searching process of the clustering algorithm, but also provides a strong technical support for achieving more accurate and personalized intelligent services. The contribution of this study is to clearly point out the shortcomings of the existing research and to achieve excellent performance in semantic information classification of IoT sensory data by proposing and validating a new clustering algorithm. The algorithm outperforms the existing clustering algorithms in terms of classification accuracy, operation stability, and operation time, which provides a new solution for the processing of IoT sensory data, as well as new ideas and methods for subsequent related research.

II. Related Work

CFSFDP is a density based clustering method aimed at addressing some of the limitations of traditional clustering algorithms when dealing with complex datasets. Many experts have conducted a series of studies using this clustering algorithm. In industrial applications, ensuring the reliability of rolling bearing rotating machinery is crucial. Wu J et al. proposed a new bearing fault diagnosis method that extracts bearing features through improved complete set empirical mode decomposition and uses CFSFDP for fault identification. This method was superior to traditional methods in fault diagnosis [6]. Chunhao Z et al. proposed an improved RNN-CFSFDP algorithm to address the limitations of the CFSFDP algorithm. This new algorithm redefined the sample density metric by introducing inverse nearest neighbors,

enhancing the robustness of the allocation process, effectively reducing the domino effect, and avoiding incorrect selection of density peaks as clustering centers. The clustering performance of RNN-CFSFDP on manifold and non-uniform density datasets was superior to or equivalent to traditional methods [7]. To ensure vehicle driving safety, Wang H et al. studied vehicle stability identification and coordinated control. Firstly, a vehicle dynamics model was established using the vehicle simulation software Carsim, and an attribute dataset representing the lateral stability of the vehicle was obtained. Subsequently, the CFSFDP algorithm was applied to classify lateral stability. The final simulation results validated the advantages of the proposed method and coordinated control strategy [8]. Ren W et al. proposed an improved algorithm to address the limitations of the original CFSFDP algorithm in anomaly detection. This algorithm effectively reduced storage and computing costs by using a small number of key data points and reducing redundant data, while maintaining the arbitrary shape clustering characteristics of CFSFDP. Compared with traditional clustering algorithms, the improved CFSFDP algorithm performed better in generating anomaly detection files in terms of speed and accuracy, achieving a balance between detection accuracy and real-time performance [9].

The Semantic Information Classification (SIC) refers to the process of classifying text or data based on the semantic information it contains. Currently, many experts have used various algorithms to build various SIC models. Borges J B et al. proposed an IoT time series classification strategy and named it TSCLAS. TSCLAS is a time series classification strategy for IoT data, which mainly distinguishes different categories by transforming the original data into the ordinal pattern domain. At the same time, this strategy also enhanced the dynamic class separability of time series by selecting the optimal parameters. TSCLAS performed well in processing large-scale and incomplete IoT data, and had advantages in classification accuracy and computational time compared to other classification algorithms [10]. Wang Z et al. proposed a novel network structure for multi label image classification, which is a semantic supplementary network with prior information. This network first generated prior information through prior information networks with different convolutional layers, and then used semantic supplementation modules to generate semantic information of potential labels highly related to the current information based on the prior information. The proposed architecture achieved better classification performance in predicting certain semantic related labels [11]. Chen L et al. proposed a method for inferring regional level metadata from building automation system data to address the issue of inconsistent and incomplete metadata in existing building automation systems. Even in the absence of intuitive labels, the proposed information classification method could accurately classify and associate regional level building automation system points. The average accuracy of its classification and association stages was 90% and 85%, respectively [12]. Liu Z et al. proposed a global semantic memory network for aspect level emotion classification tasks. Traditional attention neural networks usually only consider the interaction between aspects in a single sentence and its context when solving this task, ignoring the rich semantic information available in other sentences. This network innovatively treated contexts with similar meanings as global semantic information and incorporated them as domain knowledge into the model to generate domain specific labels, proving its effectiveness [13].

In summary, many current studies have covered the application of CFSFDP and its variants in multiple fields. These studies indicate that the CFSFDP algorithm and its improved versions have advantages in processing high-dimensional, complex, and non-uniform density datasets, effectively identifying key information, and achieving efficient SIC in multiple application fields. However, these methods still have certain limitations in the processing of the IoT sensing data (IoT-SD), especially in terms of accuracy and efficiency of SIC. To further improve the classification performance of the current model for IoT-SD, this study aims to propose a new clustering method for optimizing the SIC of IoT-SD by combining the K-center point (K-CP) algorithm and the CFSDP algorithm.

III. SIC OF IOT-SD BASED ON IMPROVED CLUSTERING ALGORITHM

To efficiently process these IoT-SDs, this study first fused K-CP and CFSFDP, designed a new clustering algorithm and named it Fusion Clustering Algorithm Based on K-Centroids and Fast Search of Density Peaks (FCA-KCFSDP). On this basis, an Automatic Semantic Annotation (ASA) model for IoT-SD was further designed, aiming to achieve automatic annotation of semantic information and improve the efficiency of information classification.

A. Design of IoT-SD Clustering Algorithm Integrating CFSDP and K-CP

In the CFSDP algorithm, local density and the minimum distance from data points to higher density points are two key basic concepts. Local density is usually measured by calculating the number of points around each point [14]. For each point, CFSDP calculates the distance from it to the closest point with higher density, which helps determine the cluster center. CFSDP typically uses a large number of samples to achieve efficient identification of cluster centers. In order to obtain more clustering centers, it is necessary to use the objective function in Eq. (1) for data selection [15].

$$Of = \sum_{i=1}^{n} p\left(obj_i, e_i\right) \tag{1}$$

In Eq. (1), Of represents the objective function. e_i represents an example of distance between multiple measurement objects. obj_i represents the measurement object. p represents the correlation between the measured object and the distance example. i represents the number of objects, and n represents the upper limit of their values. Assuming ρ_i represents local density, its calculation formulas are Eq. (2) and Eq. (3).

$$\rho_i = \sum_j \chi\left(d_{ij} - d_c\right) \tag{2}$$

In Eq. (2), d_{ij} and d_c represent the distance from object

obj_i to obj_j and the truncation distance, respectively. The density value $\chi(x)$ is represented by the difference between d_{ij} and d_c, and its specific value is Eq. (3).

$$\chi(x)=\begin{cases}1, x<0\\ 0, x\geq 0\end{cases} \tag{3}$$

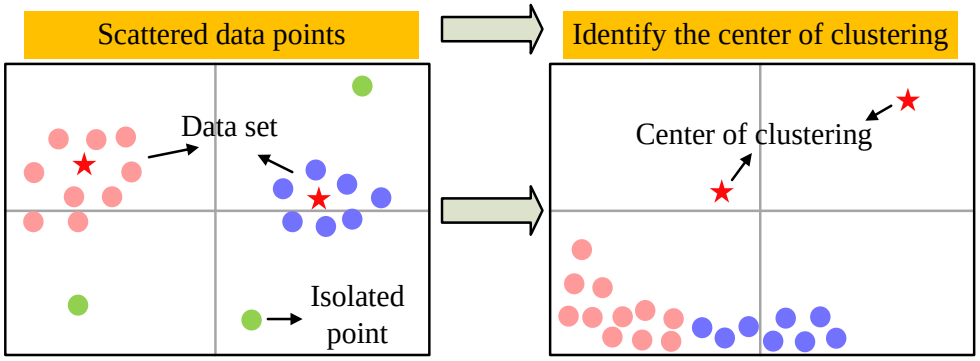
In Eq. (3), when the difference between d_{ij} and d_c is less than 0, i.e. $x<0$, the density value is $\chi(x)=1$. When the difference between d_{ij} and d_c is greater than or equal to 0, i.e. $x\geq 0$, the density value $\chi(x)=0$. Assuming that the minimum distance from a data point to a higher density point is δ_i, the calculation formulas are shown in Eq. (4) and Eq. (5).

$$\delta_i=\min\delta_i \quad \rho_j>\rho_i \tag{4}$$

In Eq. (4), when the local density ρ_j of obj_j is greater than the local density ρ_i of obj_i, the minimum distance δ_i can achieve a minimum value.

$$\delta_i=\max\delta_i \quad \rho_j\leq\rho_i \tag{5}$$

In Eq. (5), when ρ_j of obj_j is less than or equal to ρ_i of obj_i, the minimum distance δ_i can achieve a maximum value. Fig. 1 is the recognition decision diagram of the cluster center obtained by combining Eq. (1) to Eq. (5).



(a) Distribution of data before identification

(b) Distribution of data after identification

Fig. 1. Identification decision diagram of the clustering center.

Fig. 1(a) and (b) represent the distribution of surrounding data before and after cluster center recognition, respectively. Due to the high local density and minimum distance of the two pentagrams in Fig. 1(b) during the calculation process, these two points will be separately identified as outliers. And these two points happen to be the cluster centers of the two datasets in Fig. 1(a), so this method can be used to determine all the remaining cluster centers one by one. After determining all cluster centers, the remaining data will be automatically divided into nearby clusters based on the principle of nearest distance allocation. The identification formula for cluster centers is Eq. (6).

$$\gamma_i=\rho_i\times\delta_i \tag{6}$$

In Eq. (6), γ_i represents the product of local density and minimum distance. The larger the value, the more likely the data is to be the cluster center. Fig. 2 shows the running process of the CFSDP algorithm.

In Fig. 2, the execution of the CFSDP algorithm starts by calculating the local density of each data point. This step is usually achieved by quantifying the number of points within a certain radius around each point. Next, the distance from each point to the closest point with higher density and the local density value are calculated. Both local density and minimum distance serve as the axes of the decision graph to identify potential cluster centers. After determining the cluster center, the algorithm assigns the remaining points to the high-density points closest to them, forming independent clusters. Finally, to improve the accuracy of clustering, the algorithm will refine these preliminary clusters through a series of post-processing steps, and ultimately output the clustering results.

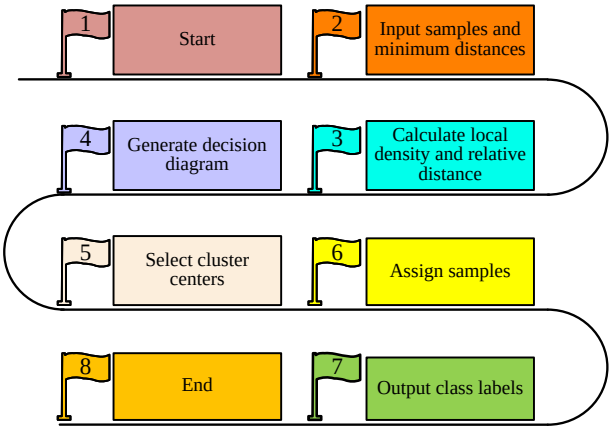


Fig. 2. Flowchart of running the CFSDP algorithm.

K-CP is similar to the K-means algorithm, but it differs in selecting cluster centers. In K-means, the cluster center is the mean of all points within the cluster, while in K-CP, the cluster center is the actual point that exists in the data, that is, the center point. Fig. 3 is the operational diagram of K-CP.

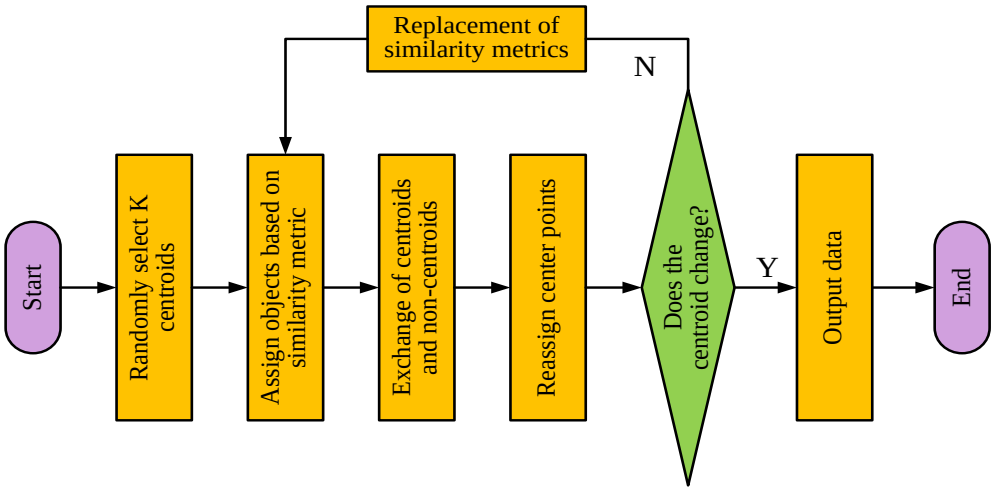


Fig. 3. Flow chart of the K-CP algorithm.

In Fig. 3, the calculation steps of K-CP are mainly divided into four parts: initializing the center point, data allocation, updating the center point, and multiple iteration algorithms. Assuming that all center point data (CPD) is denoted as m and non-CPD is denoted as o, the calculation formula for the exchange criterion function of CPD and non-CPD is Eq. (7).

$$E = \sum_{i=1}^{k} \sum_{p \in C_j} dist\left(p', o_i\right) \tag{7}$$

In Eq. (7), E represents the exchange criterion function. p' represents all objects. o_i represents an object in the C_j dataset. k represents the number of center points. To optimize the selection effect of the initial center point (ICP) of K-CP, this study adopts the dissimilarity measurement method to select the ICP, and its calculation is Eq. (8).

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{i=1}^{n} d_{ij}} \tag{8}$$

In Eq. (8), v_j represents the measure of dissimilarity of calculation object j. To sort the v_j values of each CPD and select the k objects with the top k minimum values as ICP.

Due to the fact that IoT-SD is usually high-dimensional, dynamically changing, and may contain noise and outliers, a single clustering algorithm is difficult to effectively handle such complex data. In order to improve the accuracy and robustness of IoT-SDSIC, this study combined CFSDP and K-CP to design FCA-KCFSDP, and its operating process is Fig. 4.

In Fig. 4, the calculation steps of the FCA-KCFSDP algorithm are mainly divided into three main steps: initialization clustering, initial cluster allocation, and cluster update. In order to optimize the FCA-KCFSDP, the study meticulously examined several key parameters, including the selection of the cluster radius, the density threshold, and the centroid selection criteria. The main rationale for selecting these parameter sets is based on the following considerations. When choosing the clustering radius, this study considered the distributional characteristics and density variations of the dataset. By comparing the effects of different radius values on the clustering results, it was ultimately found that the selected radius values could effectively differentiate between high-density regions and low-density regions, thus identifying the peak density points more accurately. In addition, the study also tried multiple radius values to evaluate their impact on classification accuracy and algorithm efficiency. The density thresholds were determined based on an in-depth analysis of the dataset features. By setting different density thresholds, it enables the final designed FCA-KCFSDP algorithm to control the tightness of clustering and thus optimize the clustering results. Different density thresholds were tried in this study, aiming to find a balance to ensure the quality of clustering while not over-dividing or merging real clusters. Finally, the choice of centroid directly affects the quality of clustering and the efficiency of the algorithm operation. The study develops a set of center point selection criteria based on the distribution characteristics and density information of the data. It is verified through pre-experiments that this set of criteria can effectively identify suitable clustering centers and improve the accuracy of clustering. In the initialization clustering stage, it is first necessary to calculate the local density and relative distance. Next, it is necessary to obtain the identification decision map of the clustering center based on the local density value, and calculate the initial clustering center based on the decision map. Then to change the center point of the cluster and calculate the distance from the object to the center point, obtaining the calculation formula for initial cluster allocation as shown in Eq. (9).

$$dist\left(a_i, a_j\right) = \sqrt{\sum_{t=1}^{n} \left(a_{it} - a_{jt}\right)^2} \tag{9}$$

In Eq. (9), a_i and a_j both represent objects. $a_{it} - a_{jt}$ represent the distance between two objects at time t. Introduce the variance of data objects as the weight factor for cluster center updates in cluster updates, and its expression is Eq. (10).

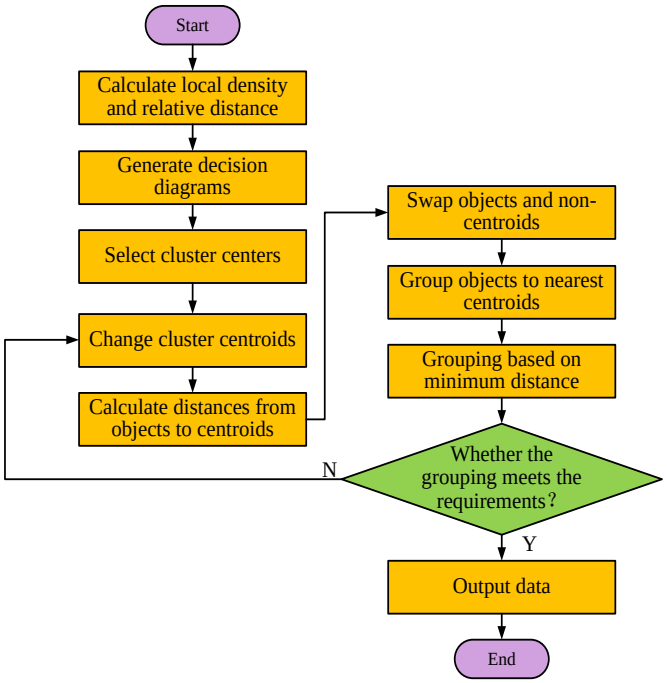


Fig. 4. Flowchart of running the FCA-KCFSDP algorithm.

$$\sigma_i = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n} dist\left(t_i - a_j^2\right)} \quad (10)$$

In Eq. (10), σ_i represents variance. t_i represents the closest data object. Based on Eq. (10), Eq. (11) is introduced to measure the total distance from all nearest data objects to object a_j.

$$D_i = \sum_{j=1}^{n}\left(dist_i - a_j\right) \quad (11)$$

In Eq. (11), D_i represents the total distance. According to Eq. (11), it is possible to update the cluster and achieve dynamic classification of data, ultimately completing the SIC of IoT data.

B. Construction of ASA Model for IoT-SD

To improve the classification efficiency of IoT-SD and further achieve automatic classification of IoT-SD, this study combined the FCA-KCFSDP clustering algorithm to build an ASA model for IoT-SD. The current ASA research mainly focuses on two methods, namely pattern based and machine learning based semantic annotation methods. For data documents with consistent formats and preprocessed data, pattern based semantic annotation methods are more effective [16-17]. This method relies on identifying patterns and implementing specific rules based on data characteristics to perform semantic annotation. On the other hand, machine learning based methods are more suitable for processing text or other unstructured data types. This type of method typically combines natural language processing technology and various machine learning algorithms for data analysis and feature extraction to reveal hidden information and knowledge in text or data. Fig. 5 shows a common ASA model structure.

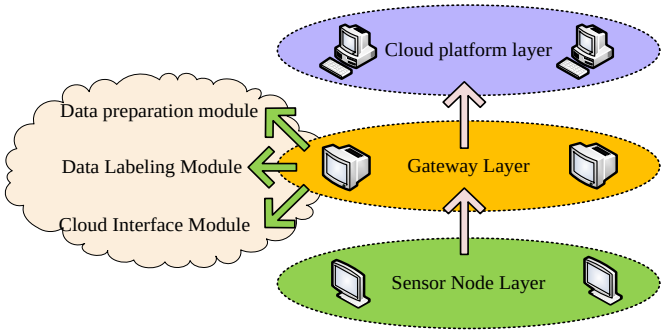


Fig. 5. Structure diagram of the traditional ASA model.

In Fig. 5, the entire ASA model consists of a cloud platform layer, a gateway layer, and a sensor node layer. The key architecture for implementing ASA is the gateway layer. In the gateway layer, it mainly includes three modules: data preparation module (DPM), data labeling module (DLM), and cloud interface module (CIM). DPM is mainly responsible for filtering and removing excess data, while converting the remaining data into XML format. This process analyzes and processes the initial data emitted by sensor nodes to minimize the computational resources required for the annotation process. Subsequently, DLM receives XML formatted data from DPM and annotates it, utilizing the concept of mapping to the application domain ontology to label the data. CIM is responsible for connecting cloud services and IoT data gateways, and implementing three main tasks. These include ensuring functional independence between the physical layer and the cloud service layer, transferring annotated IoT data to the cloud in RDF file format, processing sensor discovery content from upper layer applications, and querying requests for real-time data. The FCA-KCFSDP clustering algorithm was introduced as the core semantic classification component in the original ASA model, and the optimized structure of the IoT-SD oriented ASA model is Fig. 6.

Fig. 6 shows the ASA model with the addition of FCA-KCFSDP semantic classification component. The optimized ASA model consists of four parts, namely data pre-processing, semantic classification optimization (SCO), semantic annotation module (SAM), and CIM. Among them, the data pre-processing module aims to maintain the original functions of data aggregation, data filtering, and structured data representation. Structured data representation ensures that data is transformed in a way that is easy to process by the FCA-KCFSDP algorithm. The SCO module aims to replace the original semantic classification module with the FCA-KCFSDP clustering method. This module is responsible for assigning data to the correct semantic categories based on its characteristics and patterns. SAM will continue to receive the output of the semantic classification module and use domain ontology mapping and referencing concepts for semantic annotation of data. CIM refers to the transmission of semantically annotated data in RDF format to the cloud platform and processing of requests from the cloud platform and high-level applications. The optimized ASA model introduces the FCA-KCFSDP clustering method as the core of semantic classification, and all semantic classification work is carried out through this newly integrated algorithm. Compared

to traditional ASA models, ASA models that use FCA-KCFSDP clustering method as the semantic classification core have better classification performance and adaptability. It not only enables reasonable clustering and annotation of various types of information, further reducing the need for subsequent data processing and storage, but also reduces computational costs without sacrificing performance.

IV. Results

To test the effectiveness of the research method, the results analysis section first tested the performance of the FCA-KCFSDP clustering algorithm and proved that the algorithm performed better than other comparative algorithms in error performance and SIC. Subsequently, this study applied the FCA-KCFSDP clustering algorithm to the ASA model and tested the model's performance in actual IoT-SD classification.

A. Performance Testing of FCA-KCFSDP Clustering Algorithm

To evaluate the performance of the FCA-KCFSDP clustering algorithm in the IoT-SD semantic classification problem, this study constructed a comprehensive dataset containing multidimensional temporal data as the experimental dataset. The data set consists of readings from different sensors, including temperature, humidity, light intensity, and motion sensor data. In addition, the data was collected from three different indoor environments, covering a duration of four weeks to ensure inclusion of various environmental changes and possible anomalies. Table I shows the specific dataset data.

Table I provides the dataset information for this study. To ensure that experimental errors caused by equipment changes can be avoided in multiple repeated experiments, this study conducted experiments in the same simulation environment. The experimental operating system is Ubuntu 20.04 LTS, with an Intel Core i7-9700K CPU @ 3.60GHz and 32GB DDR4 RAM. The algorithm design was completed using Python 3.8 and TensorFlow 2.4.1. Firstly, the changes in Mean Square Error (MSE) and Mean Absolute Error (MAE) of FCA-KCFSDP, CFSDP, and K-Means Clustering algorithms in the training dataset were compared, as shown in Fig. 7.

TABLE I. Dataset Information Table

Data Indicators	Specific description
Data Source	Indoor environment sensor data (temperature, humidity, light, motion)
Sampling Period	1 min
Total Duration	4 weeks
Number of data points	10000
Pre-processing operation Labelling information	Missing value interpolation, outlier rejection, normalization Provided by domain experts, contains labels such as normal, abnormal operation, equipment failure, etc.
Data division	Training set (80%), validation set (10%), test set (10%)

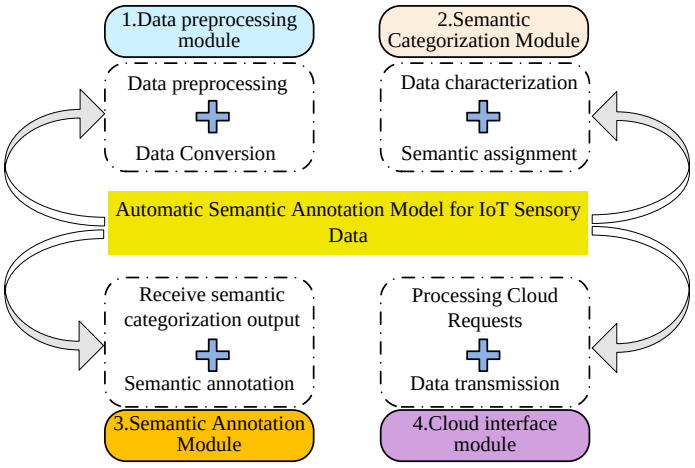
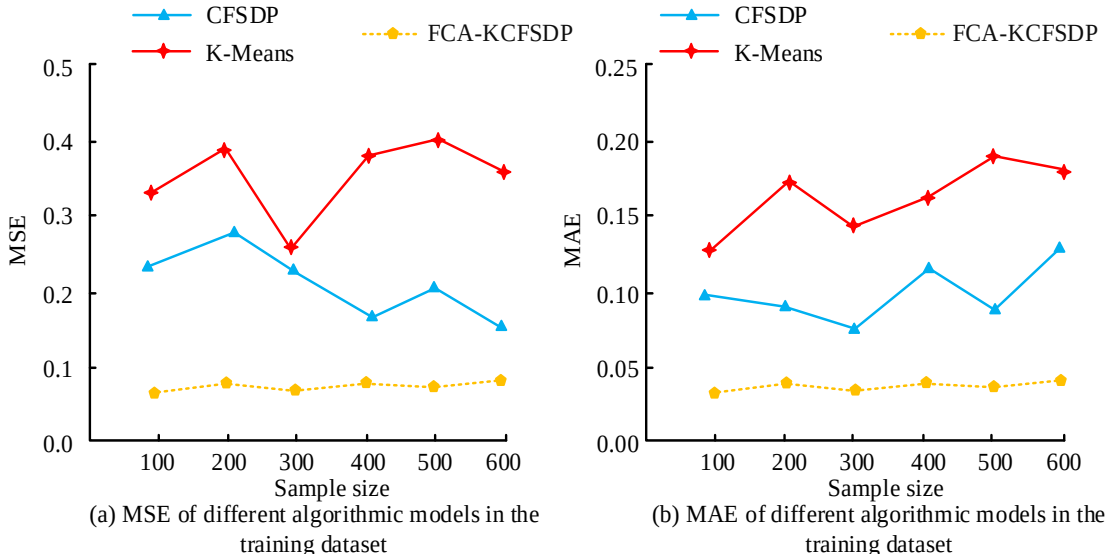


Fig. 6. Structural diagram of the ASA model for introducing FCA-KCFSDP.



(a) MSE of different algorithmic models in the training dataset

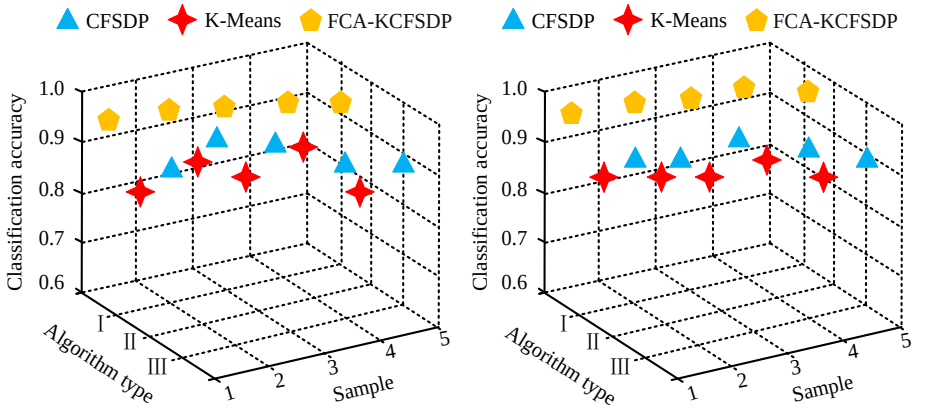
(b) MAE of different algorithmic models in the training dataset

Fig. 7. MSE and MAE for the different clustering algorithms.

Fig. 7(a) and Fig. 7(b) show the MSE and MAE values of the three clustering algorithms in the training dataset, respectively. In Fig. 7(a), as the number of training samples increases, the MSE value of the FCA-KCFSDP algorithm remains below 0.1, CFSDP is between 0.1 and 0.3, and K-Means is between 0.2 and 0.4. In Fig. 7(b), the increase in the number of training samples resulted in the MAE value of the FCA-KCFSDP algorithm always being below 0.05, CFSDP between 0.05 and 0.15, and K-Means between 0.10 and 0.20. In summary, the FCA-KCFSDP algorithm has better error performance.

Fig. 8(a) and 8(b) show the classification accuracy of the three algorithms in the training and validation sets, respectively. In Fig. 8(a), the FCA-KCFSDP, CFSDP, and K-Means algorithms have the highest classification accuracy in the training set of 0.96, 0.89, and 0.84, respectively. The highest classification accuracy of the three algorithms in Fig. 8(b) in the validation set is 0.98, 0.91, and 0.86, respectively. Therefore, the FCA-KCFSDP algorithm has higher classification accuracy in both training and validation processes, indicating that the algorithm can better perform SIC on experimental data.

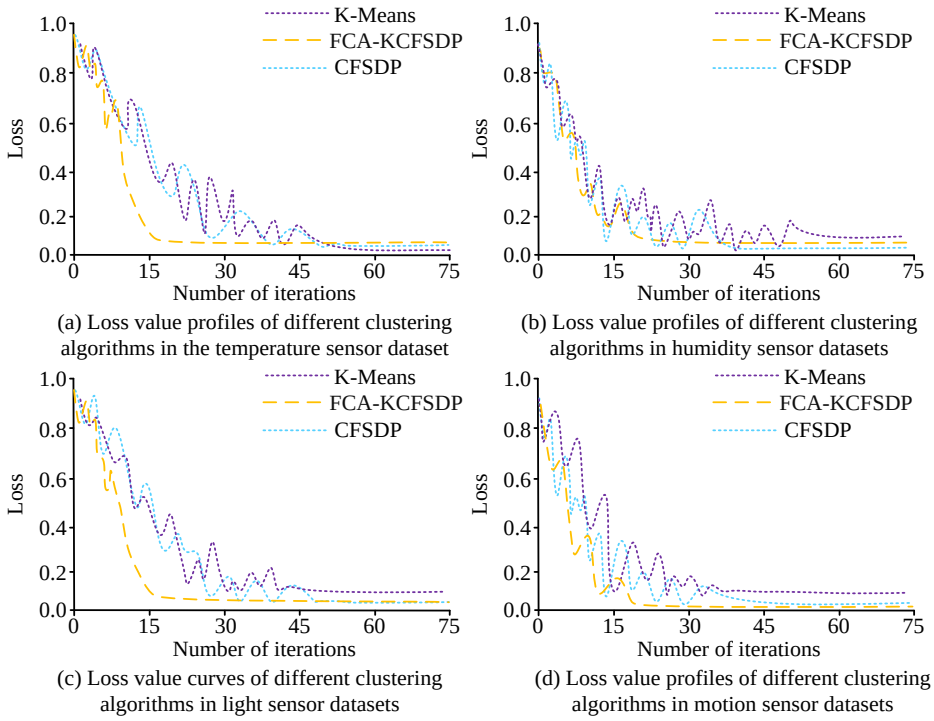
The temperature, humidity, lighting, and motion data of indoor environmental sensors are classified separately, and the loss curves of three clustering algorithms on the four classification datasets are obtained. Fig. 9(a) to Fig. 9(d) indicate that compared to CFSDP and K-Means, the FCA-KCFSDP algorithm always obtains a more stable loss curve. The FCA-KCFSDP algorithm can achieve stable classification on four datasets: illumination, temperature, motion, and humidity by iterating 15, 22, 16, and 20 times respectively.



(a) Classification accuracy of different clustering algorithms in the training set

(b) Classification accuracy of different clustering algorithms in the test set

Fig. 8. Classification accuracy of different clustering algorithms in the training and test sets.



(a) Loss value profiles of different clustering algorithms in the temperature sensor dataset

(b) Loss value profiles of different clustering algorithms in humidity sensor datasets

(c) Loss value curves of different clustering algorithms in light sensor datasets

(d) Loss value profiles of different clustering algorithms in motion sensor datasets

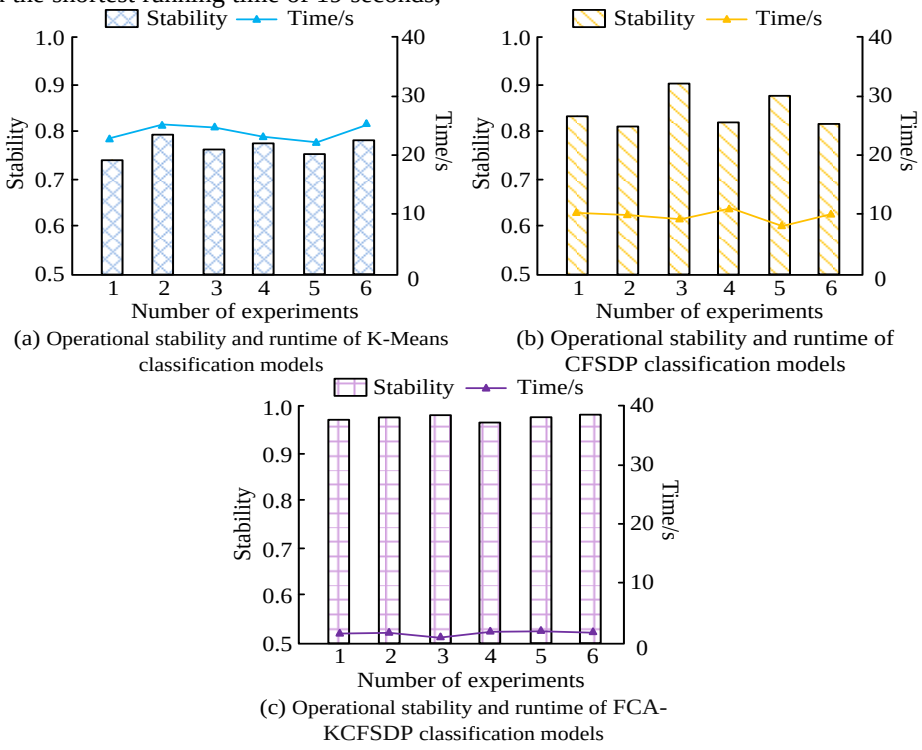
Fig. 9. Loss curves for the different clustering algorithms in the four datasets.

B. Test of the Classification Performance of ASA Models for IoT-SD

After testing the performance of the FCA-KCFSDP algorithm, this study applied it to the ASA model to further verify the classification performance of the IoT SDASA model optimized by the algorithm. Firstly, the operational stability and time variation of ASA models built by various clustering algorithms were compared, as shown in Fig. 10.

Fig. 10(a) to Fig. 10(c) shows the operational stability and time variation of three models under multiple actual tests. The above figure shows that the semantic annotation model combined with K-Means, CFSDP, and FCA-KCFSDP algorithms has the highest running stability of 0.79, 0.90, and 0.99, respectively, and the shortest running time of 19 seconds, 8 seconds, and 1 second, respectively. Therefore, applying the FCA-KCFSDP algorithm to the IoT-SDASA model can enable the model to have higher operational stability and shorter data classification time.

Fig. 11(a) and Fig. 11(b) show the satisfaction levels of IoT company users and experts with three different classification models, respectively. The satisfaction levels of users with the annotation models under K-Means, CFSDP, and FCA-KCFSDP algorithms are 0.76, 0.83, and 0.96, respectively. The satisfaction rates of experts with the models under K-Means, CFSDP, and FCA-KCFSDP algorithms are 0.72, 0.86, and 0.95, respectively.

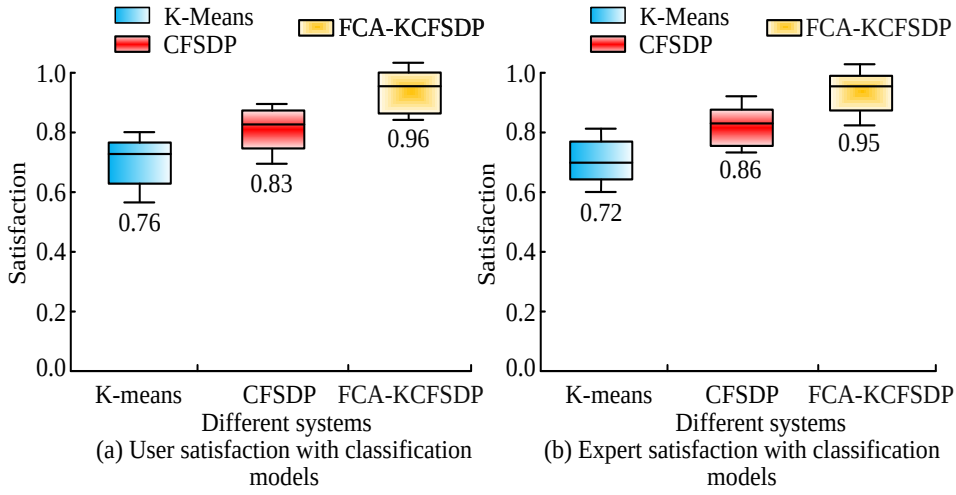


(a) Operational stability and runtime of K-Means classification models

(b) Operational stability and runtime of CFSDP classification models

(c) Operational stability and runtime of FCA-KCFSDP classification models

Fig. 10. Stability and running time of each classification model.



(a) User satisfaction with classification models

(b) Expert satisfaction with classification models

Fig. 11. Satisfaction of users and experts with the classification model.

The classification effects of four different clustering algorithms in practical applications are given in Table II. One week's IoT sensing data was collected from an enterprise, which was divided into noisy data and non-noisy data, and 1,000 of each was taken. In Table II, the classification accuracy and classification time of K-Means algorithm are 84.25% and 1.87s for noisy data, and 86.32% and 1.59s for non-noisy data, respectively. The classification accuracy and classification time of CFSDPs algorithm are 88.30% and 1.38s for noisy data, and 88.30% and 1.38s for non-noisy data, respectively. The classification accuracy and time were 89.94% and 1.26s for noisy data. k-means++ algorithm was 91.05% and 1.12s for noisy data, and 92.17% and 1.05s for non-noisy data. fca-KCFSDP algorithm was 98.49% and 1.59s for noisy data. Classification time is 98.49% and 0.25s for noisy data, and 98.85% and 0.21s for non-noisy data. Taken together, all the clustering algorithms are better at dealing with noisy data than non-noisy data, and the FCA-KCFSDP algorithm, compared to the other three comparative algorithms, has higher classification accuracy and classification efficiency.

TABLE II. ACTUAL CLASSIFICATION EFFECT OF DIFFERENT ALGORITHMS

Algorithmic models	Data Type	Classification accuracy	Classification execution time
K-Means	Noise data	84.25%	1.87s
	Non-noise data	86.32%	1.59s
CFSDP	Noise data	88.30%	1.38s
	Non-noise data	89.94%	1.26s
K-means++	Noise data	91.05%	1.12s
	Non-noise data	92.17%	1.05s
FCA-KCFSDP	Noise data	98.49%	0.25s
	Non-noise data	98.85%	0.21s

V. CONCLUSION

In response to the shortcomings of low classification accuracy and poor classification performance of the current IoT SDSIC tool, this study combined the K-CP and CFSDP algorithms to design an optimized FCA-KCFSDP, and used it to build an ASA model for IoT-SD. The research results indicate that compared to K-means and CFSDP algorithms, FCA-KCFSDP clustering algorithm had better error performance, with MAE and MSE below 0.05 and 0.1, respectively. In addition, the classification accuracy of FCA-KCFSDP, CFSDP, and K-means algorithms in the entire dataset could reach up to 0.98, 0.91, and 0.86, respectively. The sensor data in the dataset was subdivided into four types of data: motion, humidity, temperature, and lighting. It was found that the FCA-KCFSDP algorithm can reach a stable state by iterating 15, 22, 16, and 20 times respectively. Therefore, the performance of FCA-KCFSDP algorithm was superior to the other two comparative algorithms. Finally, this study also compared the stability and running time of classification models under K-means, CFSDP, and FCA-KCFSDP algorithms, and found that their highest running stability reached 0.79, 0.90, and 0.99, respectively, and their shortest running time was 19 seconds, 8 seconds, and 1 second. The FCA-KCFSDP classification model could also achieve higher classification satisfaction. In summary, the clustering algorithm and classification model designed in this study can achieve good semantic classification results. Subsequent research can further expand the types of semantic information in the dataset, thereby proving that the model has better generalization properties. Future research work includes the following points. Firstly, explore the possibility of integrating density peak based fast search clustering algorithms with deep learning models to improve the accuracy and efficiency of the model when dealing with complex datasets. Secondly, the research will be extended to more application areas, such as intelligent transportation systems, environmental monitoring, etc., to verify the universality and applicability of the algorithm. In addition, focus on security and privacy protection during data processing, and study how to ensure algorithm performance while ensuring data security and user privacy are not compromised.

ACKNOWLEDGEMENT

The research is supported by Fujian Provincial Science and Technology Plan Project, "Development and Application of Marine Ecological Environment Monitoring System for Marine Ranches" (No.2023H0029).

REFERENCES

[1] Chu Z, Cui X, Zhai X, Liu S, Qiu W, Waseem M. Anomaly detection and clustering-based identification method for consumer-transformer relationship and associated phase in low-voltage distribution systems. Energy Conversion and Economics, 2022, 3(6): 392-402.

[2] Wang S, Hua W, Liu H, Jiao L. Unsupervised classification for polarimetric SAR images based on the improved CFSFDP algorithm. International journal of remote sensing, 2019, 40(8): 3154-3178.

[3] Mohan A, Thalapala V S, Guravaiah K, Dhanyamol M V. FMGNR: fuzzy median graph for network routing applications. Wireless Networks, 2023, 29(2): 821-832.

[4] Dalski A, Kovács G, Ambrus G G. No semantic information is necessary to evoke general neural signatures of face familiarity: evidence from cross-experiment classification. Brain Structure and Function, 2023, 228(2): 449-462.

[5] Ma Z, Xia M, Lin H, Qian M, Zhang Y. FENet: Feature enhancement network for land cover classification. International Journal of Remote Sensing, 2023, 44(5): 1702-1725.

[6] Wu J, Lin M, Lv Y, Cheng Y. Intelligent fault diagnosis of rolling bearings based on clustering algorithm of fast search and find of density peaks. Quality Engineering, 2023, 35(3): 399-412.

[7] Chunhao Z, Bin X, Yiran Z. Reverse-Nearest-Neighbor-Based Clustering by Fast Search and Find of Density Peaks. Chinese Journal of Electronics, 2023, 32(6): 1341-1354.

[8] Wang H, Zhou J, Hu C, Chen W. Vehicle lateral stability control based on stability category recognition with improved brain emotional learning network. IEEE Transactions on Vehicular Technology, 2022, 71(6): 5930-5943.

[9] Ren W, Zhang J, Di X, Lu Y, Zhang B, Zhao J. Anomaly detection algorithm based on CFSFDP. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2020, 24(4): 453-460.

[10] Borges J B, Ramos H S, Loureiro A A F. A classification strategy for Internet of Things data based on the class separability analysis of time series dynamics. ACM Transactions on Internet of Things, 2022, 3(3): 1-30.

[11] Wang Z, Fang Z, Li D, Yang H, Du W. Semantic supplementary network with prior information for multi-label image classification. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(4): 1848-1859.

[12] Chen L, Gunay H B, Shi Z, Shen W, Li X. A Metadata inference method for building automation systems with limited semantic information. IEEE Transactions on Automation Science and Engineering, 2020, 17(4): 2107-2119.

[13] Liu Z, Wang J, Du X, Rao Y, Quan X. Gsmnet: global semantic memory network for aspect-level sentiment classification. IEEE Intelligent Systems, 2020, 36(5): 122-130.

[14] Li M. Learning behaviors and cognitive participation in online-offline hybrid learning environment. International Journal of Emerging Technologies in Learning (iJET), 2022, 17(1): 146-159.

[15] Tong W, Wang Y, Liu D, Guo X. A multi-center clustering algorithm based on mutual nearest neighbors for arbitrarily distributed data. Integrated Computer-Aided Engineering, 2022, 29(3): 259-275.

[16] Pellizzoni P, Pietracaprina A, Pucci G. Adaptive k-center and diameter estimation in sliding windows. International Journal of Data Science and Analytics, 2022, 14(2): 155-173.

[17] G Mehdi, H Hooman, Y Liu, S Peyman R Arif. Data Mining Techniques for Web Mining: A Survey. Artificial Intelligence and Applications, 2022, 1(1): 3-10.