Device Clustering Algorithm Based on Multimodal Data Correlation in Cognitive Internet of Things

Kai Lin*, Di Wang, Fuzhen Xia, Hongwei Ge

***Abstract*—With the development of information network, the popularity of Internet of Things (IoT) is an irreversible trend, and the intelligent demands for IoT is becoming more and more urgent. How to improve the cognitive ability of IoT is a new challenge and therefore has given rise to the emergence of Cognitive Internet of Things (CIoT). In this paper, a device level multimodal data correlation mining (DMDC) model is firstly designed based on the CCA to transform the data feature into a subspace and analyze the data correlation. The correlation of the device is obtained based on the comprehensive of data correlation and the location information of the device. Then a heterogeneous clustering model (HDC) is proposed by using the result of the correlation analysis to classify the device. Finally, we propose a device clustering algorithm based on multimodal data correlation (DCMDC) for CIoT, which combines the functions of multimodal data correlation analyze with device clustering. Extensive simulations are carried out and our results show that the proposed algorithm can effectively improve the quality of data transmission and the intelligent service.**

***Index Terms*—Multimodal, data colleration, Cognitive Internet of Things (CIoT), device clustering.**

I. Introduction

THE concept of Internet of Things (IoT) is proposed since 1999 [1], which is a technological revolution that brings us into an era of ubiquitous computing and communication. Meanwhile, cognitive IoT (CIoT) emerges to meet the current application requirements and becomes the development trend of IoT. And the center of IoT is shifted from connective to cognitive. The main idea of CIoT enables the traditional IoT to possess the features of self-sensing, decision-making, self-learning and self-adjusting intelligent. [2] proposed a view that CIoT has the ability to combine the physical world (such as goals, sources, etc.) with the social world (social behavior, user needs, etc.) to enhance the relation among intelligent resources allocation, automatic network operation and intelligent service provision. The research on CIoT is still in the development stage compared with IoT. [3] [4] proposed a cognitive management framework which could support the development of the sustainable intelligent city better than before. CIoT is regarded as an advance direction which is able to improve the performance and realize intellectualization to the current IoT [5] [6].

The data of CIoT is collected from multiple heterogeneous devices and different domains, such as numerical observations, the measurements from different devices or text from social media stream [7]. In order to meet the social enterprise needs and extract more valuable data information by mining the data correlation, some algorithms about data correlation and data clustering are studied to solve a practical application problem. [8] proposed a novel fusion learning framework which pays attention to cross-retrieval. The aim of the framework is retrieving the similar data from other types data by regarding a type of data as a query. For example, user retrieves relevant text and video by using a single picture. [9] described three clustering algorithms to analyze the data correlation of the user online behavior, which could solve the problem among clustering, person query, and social network prediction. [10] designed a novel CCA framework which combines CCA algorithm and norm-one regularization technology [11] [12], the CCA framework can extract relevant sensing data and cluster them into different clusters. [13] proposed a mobility prediction-based clustering scheme to solve the high mobility of nodes in ad hoc networks, which consists of two parts: the initial clustering stage and the cluster maintenance stage. [14] proposed an incremental clustering algorithm (ICFSKM) based on K-medoids, which can quickly find and discover the nodes with the density peak. [15] proposed a new heuristic clustering algorithm for numerical data, which aims to maximize DI (Dunn Index) [16] [17] or CHI (Calinshi Harabasz Index) [18]. [7] proposed an adaptive clustering method to design dynamic IoT data stream, the method is suitable for the underlying data drift of the data stream and can determine the number of clusters based on the data distribution, then an online clustering mechanism is used to cluster the input data stream. However, the above researches exist the defect that the processed data do not contain cognitive components and can not handle the data generated with high mobility.

In this paper, we focus on how to cluster the heterogeneous devices according to the data correlation and device distribution in CIoT. Firstly, a DMDC model is proposed based on the CCA method to analyze the multimodal data blocks by exploring the correlation among them, which aims to obtain accurate data packet results and provide the basis for detecting the correlation among devices which are explored by

Kai Lin is the corresponding author and with School of Computer Science and Technology, Dalian University of Technology, Dalian, China, 116024.
e-mail: link@dlut.edu.cn

Di Wang is with School of Computer Science and Technology, Dalian University of Technology, Dalian, China, 116024.
e-mail:dlut_ wd@mail.dlut.edu.cn.

Fuzhen Xia is with School of Computer Science and Technology, Dalian University of Technology, Dalian, China, 116024.
e-mail:7370559@mail.dlut.edu.cn.

Hongwei Ge is with School of Computer Science and Technology, Dalian University of Technology, Dalian, China, 116024.
e-mail: hwge@dlut.edu.cn.

Manuscript received XXX XX, 2017; revised XXX XX, 2017.

considering device distribution information and data packet correlation. Then a HDC model is designed to classify the heterogeneous devices according to their correlation. In the HDC model, the devices are divided into clusters and the clustering result is changed with time. Finally, a new DCMDC algorithm is proposed for multimodal data mining and device clustering in CIoT. To the best of our knowledge, it is the first work to study the device correlation problem with clustering. This paper offers the following contributions addressing the issues mentioned above:

- DMDC model is designed based on the CCA to analyze the multimodal data correlation according to their data modal. By using this model, the data blocks are mapped to the subspace to obtain the correlation of data blocks and generate the correlation among packets.
- HDC model is proposed to form the correlation among devices in CIoT, which combines the distribution of device and the correlation of the data packet. Then the device is classified into different clusters according to their device correlation and the clustering results vary with the time and the correlation of the device.
- DCMDC is designed by adopting the DMDC and the HDC model. The extensive simulations are performed to evaluate DCMDC. Simulation results demonstrate that DCMDC achieves high performance for clustering in CIoT.

The rest of this paper is organized as follows. In section II, the system model of CIoT and problem statement are introduced. DMDC and HDC model are proposed in section III. Section IV describes DCMDC in detail. Simulations and evaluations are given in Section V. In Section VI, the conclusion and the future work are discussed.

II. System Model and Problem Statement

A. Basic Topology and Architectures of CIoT

In this paper, the data cognitive and the intelligent management is integrated into IoT. Compared with traditional IoT, more human awareness are added into the interaction equipment and environment, which improves the accuracy and the efficiency of the sensor-driven complex system. The design of CIoT structure needs to meet the independent and intelligent requirements from users, which introduces the heterogeneous cognitive device to generate sensing information by adopting the interactive and convergent sharing mechanism. The above aspects provide the infrastructure of data collection and support the cognitive decision-making function and learning optimization mechanism in CIoT.

As shown in Fig. 1, the adopted CIoT architecture in this paper includes four layers. The first layer is the Data Sensing, which consists of heterogeneous intelligent devices equipped with various of sensors. These heterogeneous intelligent devices are responsible for generating and perceiving data and are able to communicate with each other under the control of the cognitive process.

The second layer named the Network Access supports different network protocols, which guarantees the compatibility of heterogeneous device communications. By adopting the cognitive function, the devices can sense the change of the local network environment. Under this premise, the devices in CIoT are able to choose the most appropriate access method according to the requirement and flexibly switch the communication mode, which is more convenient to build heterogeneous fusion network and make it possible to provide the seamless connection services to achieve the network integration and switching among different applications.

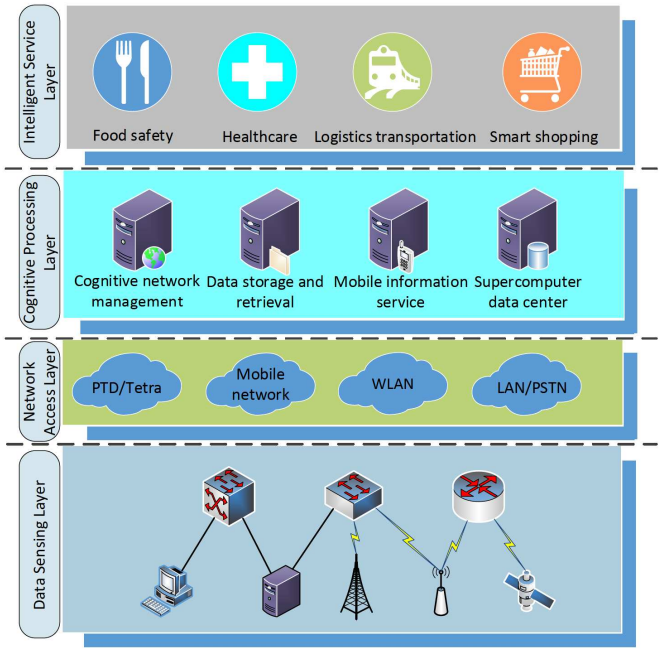


Fig. 1: The architecture of CIoT.

Cognitive Processing is the third layer, which deals with the data processing of CIoT and generates running process flows corresponding to the different applications. Massive heterogeneous sensing information is obtained by exploring the internal structure, operation mechanism and cooperative relationship among the devices, meanwhile, the network ecological environment observation and information perception are adopted to optimize network performance. The interconnection mechanism is employed by the devices to distribute and share the perceived information and the data fusion method is introduced to analyze and integrate the corresponding information. The collaborative intelligent decision is made according to the result of information fusion and execute the optimal processing. In this layer, the multimodal data correlation exploring and the device clustering need to be fulfilled.

The top layer of CIoT is the Intelligent Service, which is responsibility for transmitting the cognitive results to servers [19]–[22]. The resource demands of the users are analyzed in the Intelligent Service Layer to form the corresponding system requirements that are expressed as cognitive processes to the cognitive processing layer.

B. Problem Statement

The data modality of CIoT is increasing with the diversity of the devices which leads to the complexity of data processing.

For example, as shown in Fig. 2, CIoT collects a variety of different sensing data generated by heterogeneous devices in the Data Sensing Layer. These heterogeneous devices are divided into movable and stationary which can affect the network topology. Specifically, the correlation between multi-modal data generated by different devices is also varied with the topology and time. In this case, we assume that the sensing data generated by the devices in CIoT are heterogeneous, multidimensional and unstructured.

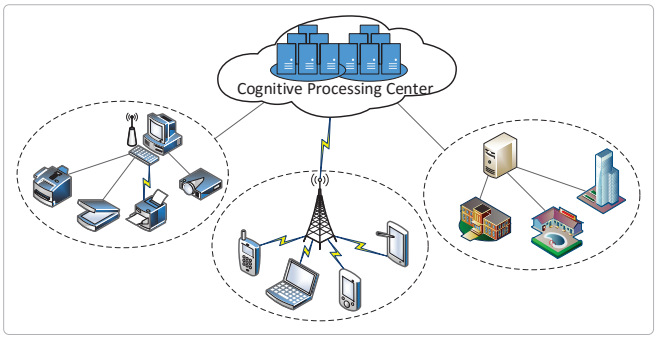


Fig. 2: The transmission method of CIoT.

In this paper, we consider that there are M heterogeneous devices in CIoT and the number of sensing data modal is N. The set of data modal is represented by $Type = \{Type_1, Type_2, ..., Type_N\}$. Considering an arbitrary time period t_κ in the time set $t = \{t_\kappa | \kappa = 1, ..., T\}$, the data set for all heterogeneous devices in CIoT is expressed as $Data = \{Data_1^{t_\kappa}, Data_2^{t_\kappa}, ..., Data_\tau^{t_\kappa}\}$. Each data packet consists of multiple different modal data blocks $Data_i^{t_\kappa} = \{Data_{i1}^{t_\kappa}, Data_{i2}^{t_\kappa}, ..., Data_{i\xi}^{t_\kappa}\}$. These different modality data blocks employ advisable approaches to extract their features. The data features are mapped into the corresponding binary code by using the hash method. The eigenvector of the data blocks is represented as $D = \{D_1, D_2, ..., D_n\} \in Data_{i\xi}^{t_\kappa}$. In the given n eigenvectors, p vectors are selected randomly ($p \ll n$), which constructs the kernel matrix K and zero-centered $D'_p = \frac{D_p - \bar{D}}{\sigma}$. $\bar{D}$ is the mean of the p vectors and σ is the variance. We define an all-zero vector e with length p for each hash equation. The vector e chooses q points in $[1, ..., p]$ to construct a row construct instruction vector e_s and assigns the corresponding value to 1. Then the $\varpi = K^{-1/2} e_s$ and the binary equation(1) are calculated.

$$h(\phi(x)) = sign(\sum_{i=1}^{p} \varpi(i)(\phi(x_i)^T \phi(x))) \tag{1}$$

The equation(1) is a binary feature coding matrix and $\phi(x)$ is the kernel function. Compared with the CCA algorithm can only map the different modality data into the same subspace E after encoding, we map the two different modality data into the subspace α and β after constructing the mapping relation. By this way, the features of different modalities exist corresponding relationships.

$$M_I : X \to \alpha \quad M_T : Y \to \beta \tag{2}$$

Two different modality data are mapped to two subspaces α and β and the two subspaces are reversible. The similarity search is performed in this subspace and a correlation measure of different data blocks is returned. In this paper, the CCA algorithm is used to train the different relationship space α and β.

To improve and utilize the cognitive function in CIoT, our objective is designing a multimodal data correlation exploring method to detect the correlation among different devices in CIoT, and proposing a device clustering algorithm based on the correlation results to enhance the cognitive ability of multimodal data in CIoT and provide more intelligent services to users.

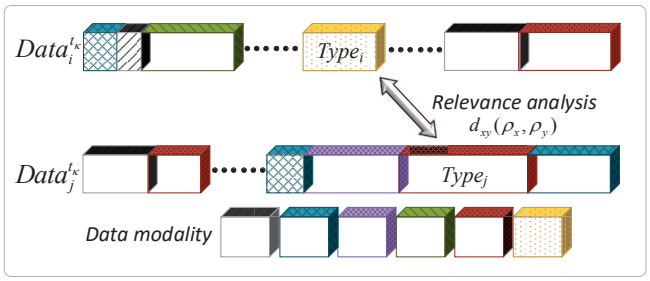


Fig. 3: The modal of data correlation.

III. Device-level Multimodal Data Correlation Mining and Clustering Model

In this section, a device-level multimodal data correlation mining (DMDC) model is designed, which utilizes the canonical correlation analysis (CCA) [23] to train the relationship space and analyze the correlation among multimodal data, and integrate the data generated by heterogeneous devices with high correlation. Based on the correlation analysis, the heterogeneous clustering model in CIoT is executed to classify the heterogeneous devices for further multimodal data fusion.

A. Multimodal Data Correlation Analysis in CIoT

Suppose there are two different modal data blocks $Data_{ix}^{t_\kappa}$ and $Data_{jy}^{t_\kappa}$ from two heterogeneous device data packets. X is the sample matrix of $m \times n_1$ and Y is $m \times n_2$. n_1, n_2 are the feature dimensions of X and Y. X and Y are represented as $\{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$. Each (x_i, y_i) is associated with a weight ζ_i which can show its importance. The weighted mean of X and Y are

$$\bar{x} = \frac{\sum_{i=1}^{m} \zeta_i x_i}{\sum_{i=1}^{m} \zeta_i} \quad \bar{y} = \frac{\sum_{i=1}^{m} \zeta_i y_i}{\sum_{i=1}^{m} \zeta_i} \tag{3}$$

The weighted sample variance and covariance are expressed as:

$$D(X) = \frac{\sum_{i=1}^{m} \zeta_i (x_i - \bar{x})^2}{\sum_{i=1}^{m} \zeta_i} \quad D(Y) = \frac{\sum_{i=1}^{m} \zeta_i (y_i - \bar{y})^2}{\sum_{i=1}^{m} \zeta_i} \tag{4}$$

$$\omega(X,Y)=\frac{\sum_{i=1}^{m}\zeta_i}{(\sum_{i=1}^{m}\zeta_i)^2-\sum_{i=1}^{m}\zeta_i^2}\sum_{i=1}^{m}\zeta_i(x_i-\bar{x})(y_i-\bar{y}) \quad (5)$$

Suppose that $\omega(X,X)$ is the weighted covariance matrix of the matrix X with itself and $\omega(Y,Y)$ is the covariance matrix of the matrix Y with itself. $\omega(X,Y)$ is the weighted covariance matrix of the matrices X and Y. The covariance matrix of the entire data set can be expressed as $\omega = \begin{bmatrix} \omega(X,X) & \omega(X,Y) \\ \omega(Y,X) & \omega(Y,Y) \end{bmatrix}$. For the linear representation of the matrix X, the linear coefficient vector is a, and the linear coefficient vector of Y is b. The linear representations are $\alpha = a^T x$ and $\beta = b^T y$, respectively.

As described in Fig.3, the optimization goal of clustering model is to maximize $corr(\alpha,\beta)$ for getting the corresponding projection vector a,b to measure the relationship between α and β, which means to find a set of optimal solutions to maximize $corr(\alpha,\beta)$. The resulting a and b are the weights that can maximize the relevance between α and β.

$$\underbrace{\arg\max_{a,b}\mathrm{corr}(\alpha,\beta)}_{a,b}=\frac{\mathrm{cov}(\alpha,\beta)}{\sqrt{D(\alpha)}\sqrt{D(\beta)}} \quad (6)$$

It can be concluded that the variance and covariance of α and β are

$$\begin{aligned} D(\alpha)=D(a^Tx)&=\frac{\sum_{i=1}^{m}(a^Tx_i-a^T\bar{x})^2}{\sum_{i=1}^{m}\zeta_i} \\ &=\frac{a^T\sum_{i=1}^{m}(x_i-\bar{x})^2a}{\sum_{i=1}^{m}\zeta_i}=a^T\omega(X,X)a \end{aligned} \quad (7)$$

$$\begin{aligned} D(\beta)=D(b^Ty)&=\frac{\sum_{i=1}^{m}(b^Ty_i-b^T\bar{y})^2}{\sum_{i=1}^{m}\zeta_i} \\ &=\frac{b^T\sum_{i=1}^{m}(y_i-\bar{y})^2b}{\sum_{i=1}^{m}\zeta_i}=b^T\omega(Y,Y)b \end{aligned} \quad (8)$$

$$\mathrm{cov}(\alpha,\beta)=a^T\omega(X,Y)b \quad (9)$$

The correlation between α and β is calculated as:

$$corr(\alpha,\beta)=\mathrm{corr}(a^Tx,b^Ty)=\frac{a^T\omega(X,Y)b}{\sqrt{a^T\omega(X,X)a}\sqrt{b^T\omega(Y,Y)b}} \quad (10)$$

By adjusting the values of the coefficients a and b of α and β to maximize $corr(\alpha,\beta)$, which becomes a convex optimization problem. The denominator in the above formula is fixed as a constant 1. Then this optimization problem is expressed as a mathematical formula as follows:

$$\begin{aligned} &Maximize: a^T\omega(X,Y)b \\ &s.t: a^T\omega(X,X)a=1, b^T\omega(Y,Y)b=1 \end{aligned} \quad (11)$$

The optimization target is finally transformed into a convex optimization process. The results of maximum value are the previously mentioned multi-dimensional X and Y correlation measure and the corresponding a,b are linear coefficients.

There are two general approaches to optimize this problem. The first is the Singular Value Decomposition(SVD) and the second is the Lagrangian Feature Decomposition. In this paper, we use the latter to solve the problem.

$$L=a^T\omega(X,Y)b-\frac{\mu}{2}(a^T\omega(X,X)a-1)-\frac{\upsilon}{2}(b^T\omega(Y,Y)b-1) \quad (12)$$

Derive the Equation(12):

$$\frac{\partial L}{\partial a}=\omega(X,Y)b-\mu\omega(X,X)a \quad (13)$$

$$\frac{\partial L}{\partial b}=\omega(Y,X)a-\upsilon\omega(Y,Y)b \quad (14)$$

Set the derivative equals to zero:

$$\omega(X,Y)b-\mu\omega(Y,Y)a=0 \quad (15)$$

$$\omega(Y,X)a-\upsilon\omega(Y,Y)b=0 \quad (16)$$

Set Equation (15) is multiplied by a^T and Equation (16) is multiplied by b^T. Under the condition of restriction $a^T\omega(X,X)a=1$ and $b^T\omega(Y,Y)b=1$, μ=υ=$a^T\omega(X,Y)b$. The equation (15) (16) are solved as

$$\begin{aligned} \omega(X,X)^{-1}\omega(X,Y)b&=\mu a \\ \omega(Y,Y)^{-1}\omega(Y,X)a&=\mu b \end{aligned} \quad (17)$$

Merging two equations (17):

$$\omega(X,X)^{-1}\omega(X,Y)\omega(Y,Y)^{-1}\omega(Y,X)a=\mu^2a \quad (18)$$

It can be seen that the result is the feature decomposition of $\omega(X,X)^{-1}\omega(X,Y)\omega(Y,Y)^{-1}\omega(Y,X)$ to find the largest generalized eigenvalue μ. In this case, the eigenvector corresponding to the largest eigenvalue is the linear coefficient a of X. Similarly, we can get the eigenvector of the linear coefficient b of Y.

For the designed device-level multimodal data correlation mining model, the mapping relation is reversible between the two spaces. As shown in Fig. 4, it helps to form a compact and efficient representation. In this representation, the vector ρ_x is employed as a spatial coordinate and the space X maps into the largest subspace α while the vector ρ_y is a spatial coordinate and the space Y maps into the largest subspace β. The set is expressed as $(\alpha,\beta)=\{(\rho_{x1},\rho_{y1}),(\rho_{x2},\rho_{y2}),...,(\rho_{xm},\rho_{ym})\}$.The distance function

$d_{xy}(\rho_x, \rho_y)$ is employed to determine the maximize correlation between the data block $Data_{ix}^{t_\kappa}$ and $Data_{iy}^{t_\kappa}$.

$$d_{xy}(\rho_x, \rho_y) = \sqrt{\sum_{i=1}^{m} (\rho_{xi} - \rho_{yi})^2} \tag{19}$$

The correlation is inversely proportional to the distance. The correlation between two heterogeneous device data packets can be speculated from the correlation between their data blocks set. The measure of different data blocks is determined by their bit length and the importance in the packet which are expressed as ℓ and Θ. For the data packets, the set is expressed as $\ell_i^{t_\kappa} = \{\ell_{i1}^{t_\kappa}, \ell_{i2}^{t_\kappa}, ..., \ell_{i\xi}^{t_\kappa}\}$ and $\Theta_i^{t_\kappa} = \{\Theta_{i1}^{t_\kappa}, \Theta_{i2}^{t_\kappa}, ..., \Theta_{i\xi}^{t_\kappa}\}$. The correlation among packets is depended on the weighted sum of the data blocks as shown in Equation (20).

$$\begin{aligned} r_{ij} &= corr(Data_i^{t_\kappa}, Data_j^{t_\kappa}) \\ &= \sum_{x \in Data_i^{t_\kappa}} \sum_{y \in Data_j^{t_\kappa}} \frac{\ell_x^{t_\kappa} \cdot \ell_y^{t_\kappa} + \Theta_x^{t_\kappa} \cdot \Theta_y^{t_\kappa}}{d_{xy}(\rho_x, \rho_y)} \end{aligned} \tag{20}$$

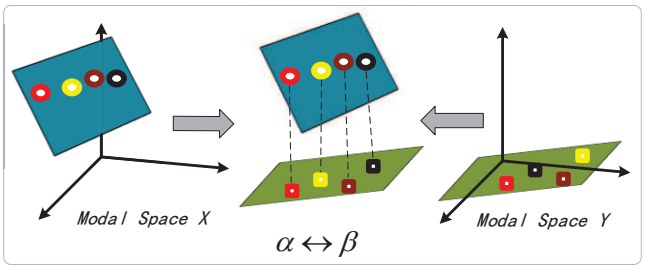


Fig. 4: The public subspace.

B. The Heterogeneous Device Clustering Model for the CIoT

As mentioned above, the data packets generated by the heterogeneous devices have different degrees of correlation. Next, we consider how to classify the heterogeneous devices in CIoT to make full use of such correlation among multimodal data. A heterogeneous device clustering Model (HDC) is designed for CIoT to improve the performance of the following processing, such as devices co-operation and data sharing, which classifies the heterogeneous devices according to the relativity between the data packets and their distribution. By utilizing the DMDC model, the data packet correlation can be extracted to get the correlation efficiency r_{ij}. We assume that each heterogeneous device transmits to one data packet at any time t_κ. In addition to the degree of correlation among the data, the distribution of the heterogeneous devices is also considered in the HDC model, which avoids the long-distance devices are classified into the same cluster to improve the network communication performance.

Therefore, the distance factor is taken into account during the clustering of the heterogeneous devices. For two heterogeneous devices $R_i, R_j \in \{R_M\}$, the distance between them is:

$$\Omega(R_i, R_j) = \sqrt{(R_i^x - R_j^x)^2 - (R_i^y - R_j^y)^2} \tag{21}$$

(R_i^x, R_i^y) and (R_j^x, R_j^y) are the location for the device R_i and R_j, respectively.

The data correlation and distance among heterogeneous devices are jointly adopted to decided the correlation of two devices:

$$\Phi(R_i, R_j) = \frac{\lambda r_{ij}}{\Omega(R_i, R_j)} \tag{22}$$

In the equation (22), λ is the equilibrium factor used to coordinate the relationship between the device distance and the data correlation. The results of $\Phi(R_i, R_j)$ is standardized to make $\Phi(R_i, R_j) \in [0, 1]$. If the two devices are not relevant, the $\Phi(R_i, R_j) = 0$. And we default the device has the correlation of 1 to itself, $\Phi(R_i, R_i) = 1$.

The clustering set $C = \{C_1, C_2, ...C_i, ...C_s\}$ represents the clustering results is firstly defined in the process of clustering. In the initial stage of clustering, each device in CIoT is classified as a cluster to complete the initialization of the clustering set, which means $C = \{R_1, R_2, ..., R_i, ..., R_M\} = \{C_1, C_2, ..., C_i, ..., C_s\}$ and $R_i = C_i$. In order to describe the correlation between devices and clusters clearly, a adjacency matrix $\Sigma_{s\times s}$ is defined to describe the correlation between them. Due to each device is a cluster at the initial stage of clustering, the value of each device in the adjacency matrix is the correlation between the two devices when the adjacency matrix is initialized. Where $\Phi(R_i, R_j) = \Phi(C_i, C_j)$. The adjacency matrix $\Sigma_{s\times s}$

$$\Sigma_{s\times s} = \begin{bmatrix} \Phi(C_1, C_1) & \bullet & \bullet & \bullet & \Phi(C_1, C_s) \\ \bullet & & \bullet & & \bullet \\ \bullet & & & \bullet & \bullet \\ \bullet & & & & \bullet \\ \Phi(C_s, C_1) & \bullet & \bullet & \bullet & \Phi(C_s, C_s) \end{bmatrix}$$

The elements on the main diagonal of the adjacency matrix represent the correlation of the cluster itself and its default value is equal to 1. Since the value of the correlation between the different clusters or the different devices is from 0 to 1, and the maximum value in the adjacency matrix $\Sigma_{s\times s}$ is required in the subsequent process, the elements on the main diagonal are set to 0. In order to reduce the time and space complexity of the clustering process and unnecessary operations, the value of the element in matrix $\Sigma_{s\times s}$ which less than the threshold value $\vartheta(\Phi(C_i, C_j) < \vartheta)$ is also assigned to zero $\Phi(C_i, C_j) = 0 (0 < i, j < s)$.

For any time period t^κ, the clustering result set C and the adjacency matrix $\Sigma_{s\times s}$ are initialized firstly. Then maximum value of the correlation $\Phi(C_i, C_j)$ is found from the matrix $\Sigma_{s\times s}$ and the corresponding two clusters C_i and C_j are extracted in the set C. The clusters C_i and C_j are combined into a new cluster C_η, which is $C_\eta = C_i \cup C_j$, and the clustering C_η is added into the clustering result set to form the new set C'. The Equation (23) is used to calculate the modularity Q' of the results in the clustered result set. The initial value of $Q = -1$.

$$Q = \frac{1}{2s}\sum_{ij}(\Sigma_{ij} - \frac{\Phi(C_i, C_j)}{2s})\delta(C_i, C_j) \tag{23}$$

δ is a discriminant function for determining whether the devices i and j belong to the same cluster. The value of function δ is 1 if the device i and j in the same cluster, otherwise it is equal to 0.

If $Q' < Q$, the two clusters C_i and C_j in clustering C_η are separated and reintroduced into the clustered result set C. The correlation between C_i and C_j of the clustering is set to 0 in the matrix $\Sigma_{s\times s}$. The algorithm is terminated if $Q' \geq Q$ while the difference between the current iteration module degree Q' and the previous iteration module degree Q is relatively small, $|Q' - Q| \leq \varepsilon$. The ε is a very small value. If $|Q' - Q| > \varepsilon$, the result set C' of current iteration is the initial result set of the next iteration, which means $C = C'$. In addition, the adjacency matrix $\Sigma_{|\mathrm{C}|\times|\mathrm{C}|}$ is reconstructed by using the elements in the clustering result set C, where $|C|$ is the number of elements in the set C. In the clustering result set C, some clusters are composed of only one device and others have multiple devices. For constructing the adjacency matrix $\Sigma_{|\mathrm{C}|\times|\mathrm{C}|}$, a detailed description of the correlation between clustering of a node is given from the Equation (22).

The Equation (24) is used to calculate the relationship between different clusters, and a cluster can contain one or multiple devices.

$$\Phi(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|}\sum_{l=1}^{|C_i|}\sum_{k=1}^{|C_j|}\Phi(R_l, R_k) \tag{24}$$

$|C_i|$ and $|C_j|$ represent the number of devices in clusters C_i and C_j respectively, $R_l \in C_i$ and $R_k \in C_j$. All the elements on the main diagonal are set to zero in the matrix $\Sigma_{|\mathrm{C}|\times|\mathrm{C}|}$. By adopting the reconstructed adjacency matrix $\Sigma_{|\mathrm{C}|\times|\mathrm{C}|}$ and the clustering result set C for iterative processing, the HDC model is executed according to the above steps until the module is calculated by the clustering result set C which satisfies the condition $|Q' - Q| \leq \varepsilon$. The heterogeneous devices in CIoT are classified into different clusters based on the correlation between the devices in different time periods through the above steps.

Algorithm 1 Device Clustering Based on Multimodal Data Correlation Algorithm.

Require:
At time t^κ, $Data = \{Data_1^{t_\kappa}, Data_2^{t_\kappa}, ..., Data_\tau^{t_\kappa}\} \in \{R_M\}$
Ensure:
The clustering set C;
for $i = 1$ to τ **do**
  for $j = 1$ to τ **do**
    Obtain two heterogeneous devices' data packets $Data_i^{t_\kappa} = \{Data_{i1}^{t_\kappa}, Data_{i2}^{t_\kappa}, ..., Data_{i\xi}^{t_\kappa}\}$ and $Data_j^{t_\kappa} = \{Data_{j1}^{t_\kappa}, Data_{j2}^{t_\kappa}, ..., Data_{j\xi}^{t_\kappa}\}$
  end for
  for $x = 1$ to ξ **do**
    for $y = 1$ to ξ **do**
      Extracting $Data_{ix}^{t_\kappa}$ $Data_{jy}^{t_\kappa}$ from two packets
      Extracting the feature $D_x = \{D_1, D_2, ..., D_m\} \in Data_{ix}^{t_\kappa}$ and $D_y = \{D_1, D_2, ..., D_m\} \in Data_{iy}^{t_\kappa}$
      Hashing $h(D_x, D_y) = \{X, Y\}$
      Adopting CCA to get $corr(\alpha, \beta)$ from Equation(10)
      Calculating $(\alpha, \beta) = (a^T x, b^T y)$
      Calculating $d_{xy}(\rho_x, \rho_y)$ between $Data_{ix}^{t_\kappa}$ and $Data_{jy}^{t_\kappa}$
    end for
  end for
  for $i = 1$ to τ **do**
    for $j = 1$ to τ **do**
      Calculating $corr(Data_i^{t_\kappa}, Data_j^{t_\kappa})$
      Calculating $\Phi(R_i, R_j)$ from the Equation(21)
    end for
  end for
  Initializing $C = \{C_1, C_2, ..., C_i, ..., C_s\}$ and $\Sigma_{s\times s}$
  Getting modularity Q'
  if $|Q' - Q| \leq \varepsilon$ **then**
    return New Cluster C' and $\Sigma_{|C|\times|C|}$;
  end if
  if $|Q' - Q| \leq \varepsilon$ **then**
    return The Cluster C and set $\Phi(R_i, R_j) = 0$
  end if
end for

IV. Device clustering algorithm based on multimodal data correlation

In the previous section, DMDC and HDC model are adopted to analyze data correlation and classify the devices. The clustering result is able to be applied to the corresponding network routing distribution or other application services in CIoT. By utilizing DMDC and HDC model, the device clustering algorithm based on multimodal data correlation (DCMDC) is designed to improve the performance of CIoT.

As mentioned above, the device is heterogeneous and the data is multimodal in CIoT. In order to explore the correlation among multimodal data, the DMDC model is used to detect the correlation among the multimodal data from heterogeneous devices. The two packets $Data_i^{t_\kappa}$ and $Data_j^{t_\kappa}$ from heterogeneous devices which contain multiple different modal data blocks. The correlation of each data block and the distribution of device information are both analyzed to obtain the two devices correlation. The data block feature is extracted and the corresponding eigenvector $D_x = \{D_1, D_2, ..., D_m\} \in Data_{ix}^{t_\kappa}$ and $D_y = \{D_1, D_2, ..., D_m\} \in Data_{iy}^{t_\kappa}$ from two different modal data blocks $Data_{ix}^{t_\kappa}$ and $Data_{jy}^{t_\kappa}$. Next, the eigenvector converts into a hash representation $h(D_x, D_y)$ and is mapped into the same subspace. The CCA algorithm is adopted to transform linearly the vector $\{X, Y\} = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ into the subspace α and β with the maximize correlation. The function $d_{xy}(\rho_x, \rho_y)$ is used to determine the correlation of two data blocks α and β. The correlation $corr(Data_i^{t_\kappa}, Data_j^{t_\kappa})$ from the heterogeneous device data packets is obtained after synthesizing the

weight information of the data blocks.

The correlation between the two heterogeneous devices $\Phi(R_i, R_j)$ is obtained by taking into account the combination of the device distribution $\Omega(R_i, R_j)$ after obtaining the data correlation. The HDC model is established to cluster the heterogeneous devices in CIoT to integrate the various requirements of the device clustering according to the correlation among devices. In the clustering process, each device in CIoT is divided as a cluster and composed a clustering result $C = \{C_1, C_2, ..., C_i, ..., C_s\}$. The adjacency matrix $\Sigma_{s\times s}$ is adopted to represent the correlation from the devices. The new cluster C' is formed through the correlation between the devices and the modularity Q is used to measure the new clustering results. If the new module Q' satisfy the condition $|Q' - Q| \leq \varepsilon$, the cluster is converted to a new cluster C' and update the adjacency matrix to $\Sigma_{|C|\times|C|}$. The result set C' of current iteration is the initial result set of the next iteration. The clustering results are formed on this basis which is served as the initial value for the next iteration. If $|Q' - Q| \geq \varepsilon$, the clustering result is split and update $\Sigma_{s\times s}$ to set the correlation of the two heterogeneous devices to 0. The device provides the necessary assistance for network routing and other application services for CIoT based on their correlation clustering. The algorithm of building and working process of the DCMDC is described in Algorithm1.

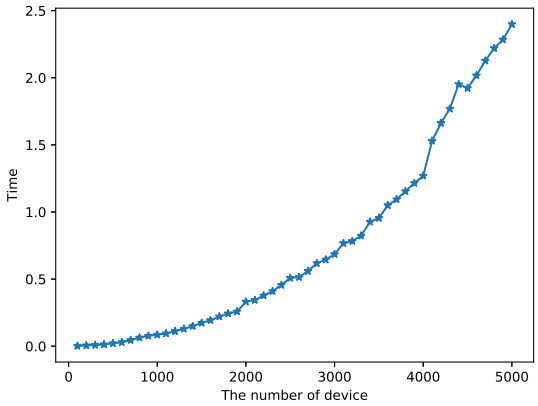


Fig. 5: The clustering time of devices

Now, we discuss the complexity of clustering algorithms. The correlation among the devices is calculated and sorted in the initial clustering. For example, if there are n devices in the network, the $n(n-1)/2$ devices correlation is obtained in the initial clustering. In the clustering process, the number of inter-cluster correlations is less than $n(n-1)/2$. So DCMDC calculates the correlation of $n(n-1)/2$ devices at most, the time complexity is $O(n^2)$. As shown in Figure 5, when there are only 1000 devices in the network, the clustering time has small growth with the number of devices increases, and the fluctuation does not exceed 0.1 seconds. When the number of devices increases to 5000, the device cluster time exponentially increase and has a greater impact on the clustering. At this time, the clustering effect of DCMDC is not better than before. Therefore, the algorithm takes into account the common relationship between data correlation and device distance in running process. For the environment with a large number of network devices, the clustering process is running in a small environment. In Equation (22), the common role of device location information and data correlation is defined.

V. Simulation and Result

This section discusses the results of simulations to evaluate the performance of DCMDC. In this section, we use an LFR artificial network [24] to evaluate the clustering result of DCMDC. The LFR is a simulation network that can contain overlapping clusters. The artificial network is constructed by controlling the power distribution of device degree and clustering size. At the same time, the network can also initialize the degree of community overlap.

In the simulation, four different sizes and characteristics networks are built. As shown in Table I, three parameters are considered in the network construction, the number of devices (ND), device degree distribution index (D3I) and clustering size distribution index (CSDI). In the artificial network experiment, we gradually increase the average number k of devices, $k = 10, k = 20$ and $k = 30$ respectively, observe the experimental results from three levels.

TABLE I: Parameters for eight different types networks

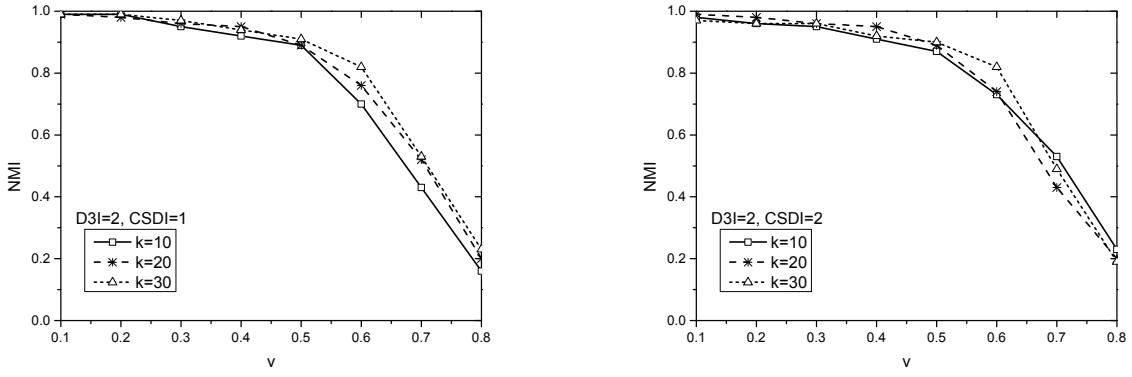
Name	G1	G1	G3	G4	G5	G6	G7	G8
ND	1000	1000	1000	1000	5000	5000	5000	5000
D3I	2	2	3	3	2	2	3	3
CSDI	1	2	1	2	1	2	1	2

We use the appropriate evaluation criteria of Normalized Mutual Information (NMI) [25] to evaluate the effect of clustering. The NMI calculates the information between two clusters to obtain the similarity of two clustering. NMI is used to evaluate the matching degree of the clustering structure detected by the algorithm with the real clustering structure. NMIs formula is defined as follows:

$$NMI(\mathrm{E}|\mathrm{Z}) = 1 - [H(\mathrm{E}|\mathrm{Z}) + H(Z|E)]/2 \qquad (25)$$

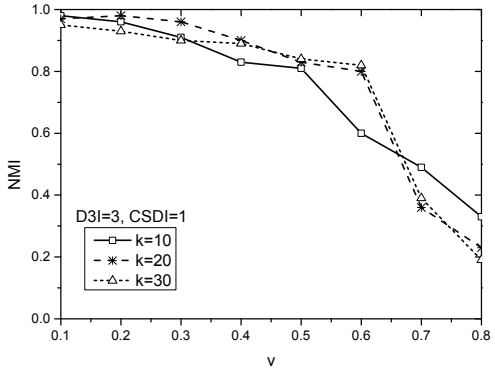
Where E and Z are given two clustering structures. The larger of the value NMI, the more similar to the two clustering structures. The clustering result calculated by the algorithm is similar to the clustering structure of the artificial network, where $NMI \in [0, 1]$.

As shown in the Figures 6 and 7, the NMI value between the clustering structure generated by the algorithm and the artificial network clustering structure is gradually reduced with the increase of the network topology mixed parameter v. The similarity between the two results reduced and this is a normal phenomenon. Because with the increase of v, the structure of the network becomes more complex and it is difficult to detect the real clustering structure. When the value of v is relatively small, the clustering structure generated by the algorithm is almost the same as the artificial network structure, which shows that the clustering effect of the algorithm is quite good. Generally speaking, when $v \leq 0.5$, NMI value can reach more than 0.8. The clustering structure calculated by

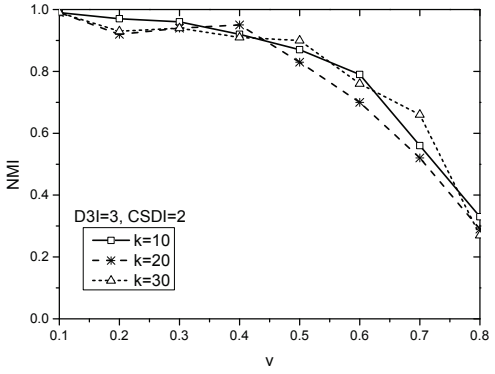


(a) D3I=2,CSDI=1

(b) D3I=2,CSDI=2



(c) D3I=3,CSDI=1



(d) D3I=3,CSDI=2

Fig. 6: Simulation results for DCMDC in LFR networks, network size ND=1000

the algorithm has very good results. In addition, the results are compared with the corresponding experimental in Figures 6 and 7. When the number of devices reach to 5000 and the size of the network increases, the corresponding NMI value will be slightly reduced. But the algorithm proposed in this paper can get better results.

VI. Conclusion

To increase the data cognitive ability of CIoT, this paper introduces a device clustering algorithm based on multimodal data correlation which including the function of data correlation analyze and device clustering. A device-level multimodal data correlation mining model is firstly proposed based on the CCA algorithm to analyze the multimodal data and device correlation, which is capable of classifying the device according to the data correlation and device distribution. The DCMDC clusters the heterogeneous devices in CIoT according to their correlation by using the result of the data correlation mining model. Extensive simulations are performed to evaluate the proposed algorithm. The results show that the designed algorithm can achieve a satisfying quality of device clustering and has the potential to transform into a practical technique in CIoT.

Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities under grant No. DUT16QY18.

References

[1] Hwang K. Chen M. *Big Data Analytics for Cloud/IoT and Cognitive Computing*, Wiley, U.K., ISBN: 9781119247029, 2017.

[2] Wu Q. Ding G. Xu Y. Feng S. Du Z. Wang J. Long K. *Cognitive internet of things: a new paradigm beyond connection*. IEEE Internet of Things Journal, vol. 1, no. 2, pp. 129-143, 2014.

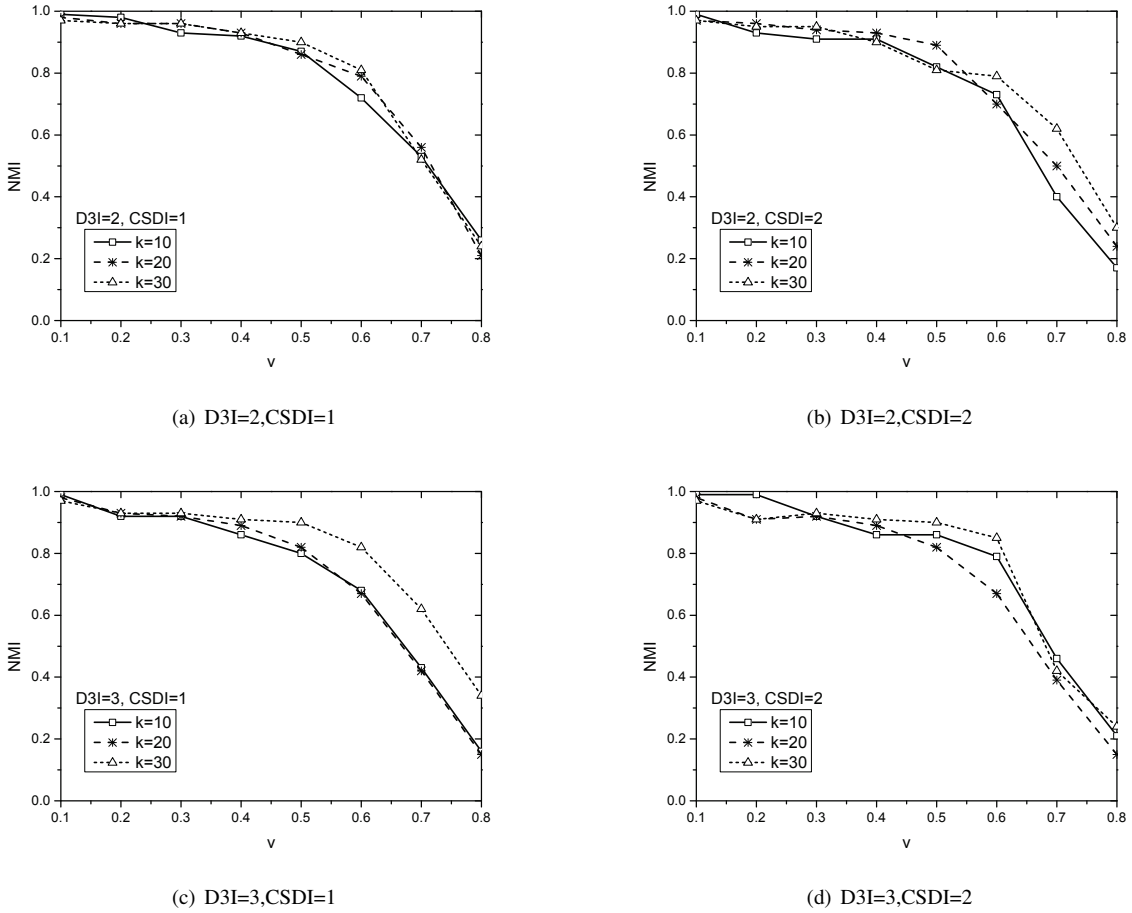
[3] Vlacheas P. Giaffreda R. Stavroulaki V. Kelaidonis D. Foteinos V. Poulios G. Demestichas P. Somov A. Biswas AR. Moessner K. *Enabling smart cities through a cognitive management framework for the internet of things*. IEEE Communications Magazine, vol. 51, no. 6, pp. 102-111, 2013.

[4] Chen M. Yang J. Zhu X. Wang X. Liu M. Song J. *"Smart Home 2.0: Innovative Smart Home System Powered by Botanical IoT and Emotion Detection"*, Mobile Networks and Applications, DOI 10.1007/s11036-017-0866-1, 2017.

[5] Zhang M. Zhao H. Zheng R. Wu Q. Wei W. *Cognitive Internet of Things: Concepts and application example,* International Journal of Computer Science Issues, vol. 9, no. 3, pp. 151-158, 2012.

[6] Chen M. Yang J. Hao Y. Mao S. Hwang K. *"A 5G Cognitive System for Healthcare", Big Data and Cognitive Computing*, vol. 1, no. 1, DOI:10.3390/bdcc1010002, 2017.

[7] Puschmann D. Barnaghi P. Tafazolli R. *Adaptive Clustering for Dynamic IoT Data Streams*. IEEE Internet of Things Journal, vol. 4, no. 1, pp. 64-74, 2016.



(a) D3I=2,CSDI=1

(b) D3I=2,CSDI=2

(c) D3I=3,CSDI=1

(d) D3I=3,CSDI=2

Fig. 7: Simulation results for DCMDC in LFR networks, network size ND=5000

[8] Wang K. He R. Wang L. Wang W. Tan T. *Joint Feature Selection and Subspace Learning for Cross-modal Retrieval.*. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 38, no. 10, pp. 2010-2023, 2016.

[9] Yang Y. Wang C. *A novel method of data correlation analysis of the big data based on network clustering algorithm*. Communication Software and Networks (ICCSN), IEEE International Conference on, pp. 360-366, 2015.

[10] Chen J. Schizas I.D. *Distributed sparse canonical correlation analysis in clustering sensor data*, Signals, Systems and Computers, 2013 Asilomar Conference on. IEEE, pp. 639-643, 2013.

[11] Tibshirani R. *Regression Shrinkage and subset Selection via the Lasso,* Journal of the Royal Statistical Society, Series B, vol. 58, no. 1, pp. 267-288, 1996.

[12] Zou H. Hastie T. Tibshirani R. *Sparse Principal Component Analysis,* Journal of Computational and Graphical Statistics, vol. 15, no. 2, pp. 265-286, 2006.

[13] Ni M. Zhong Z. Zhao D. *MPBC: A Mobility Prediction-Based Clustering Scheme for Ad Hoc Networks.* IEEE Transactions on Vehicular Technology, vol. 60, no. 9, pp. 4549-4559, 2011.

[14] Zhang Q. Zhu C. Yang L. Chen Z. Zhao L. Li P. *An Incremental CFS Algorithm for Clustering Large Data in Industrial Internet of Things.* IEEE Transactions on Industrial Informatics, vol. 13, no. 3, pp. 1193-1201, 2017.

[15] Siddiqi U.F. Sait S.M. *A New Heuristic for the Data Clustering Problem.* IEEE Access, vol. 5, pp. 6801-6812, DOI: 10.1109/ACCESS.2017.2691412, 2017.

[16] Bezdek C. Pal N. *Cluster validation with generalized dunns indices,* in Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, New Zealand, 1995.

[17] Dunn J.C. *Well-separated clusters and optimal fuzzy partitions,* Journal of Cybernetics, vol. 4, no. 1, pp. 95-104, 1974.

[18] Calinski T. Harabasz J. *A dendrite method for cluster analysis,* Communications in Statistics, vol. 3, no. 1, pp. 1-27, 1974.

[19] Chen M. Ma Y. Li Y. Wu D. Zhang Y. Youn C. *"Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System"*, IEEE Communications, vol. 55, no. 1, pp. 54-61, 2017.

[20] Chen M. Zhou P. Fortino G. *"Emotion Communication System"*, IEEE Access, vol. 5, pp. 326-337, 2017.

[21] Lin K. Luo J. Hu L. Shamim Hossain M. Ghoneim A. *Localization based on Social Big Data Analysis in the Vehicular Networks*. IEEE Transactions on Industrial Informatics, DOI 10.1109/TII.2016.2641467, 2016.

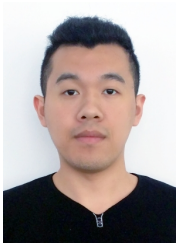
[22] Lin K. Chen M. Deng J. Hassan M.M. Fortino G. *Enhanced fingerprinting and trajectory prediction for IoT localization in smart buildings*. IEEE Transactions on Automation Science and Engineering, vol. 13, no. 3, pp. 1294-1307, 2016.

[23] Hardoon D.R. Szedmak S. Shawe-Taylor J. *Canonical correlation analysis: An overview with application to learning methods*. Neural computation, vol. 16, no. 12, pp. 2639-2664, 2004.

[24] Lancichinetti A. Fortunato S. Kertsz J. *Detecting the overlapping and hierarchical community structure of complex networks*. New Journal of Physics, vol. 11, no. 3, pp. 19-44, 2008.

[25] McDaid A.F. Greene D. Hurley N. *Normalized mutual information to evaluate overlapping community finding algorithms*. arXiv preprint arXiv:1110.2515, 2011.

Kai Lin is an associate professor at the School of Computer Science and Technology, Dalian University of Technology. He received his M.S. and Ph.D. degrees in communication engineering from Northeastern University, China. His research interests include wireless communication, data mining and data fusion, big data analysis, mobile ad hoc networks, cyber physical systems, and sensor networks. He is an Associate Editor of Recent Patents on Telecommunications, and has also served several journals and Special Issues as Editor or Guest Editor. He has served as General Chair, Technical Program Committee Chair, and Publicity Chair for many international conferences, including IEEE I-SPAN, CSE, SCALCOM, EmbeddedCom, MWNS, IWSMN, and MSN. He has also participated in more than 40 international TPCs. He has authored or co-authored over 50 papers in international journals and conferences.



Di Wang is a student at the School of Computer Science and Technology, Dalian University of Technology. He received her B.S. degree in sensor network from Dalian Maritime University, China, in 2015. He is currently pursuing his M.S. degree, in computer science and technology from Dalian University of Technology. His current research interests is big data analysis, 5G, IoT, network spectrum allocation.

Fuzhen Xia is a student at the School of Computer Science and Technology, Dalian University of Technology. She received her B.S. degree in computer science and technology professional from Zhengzhou University, China, in 2016. She is currently pursuing her M.S. degree, in computer software and theory from Dalian University of Technology. Her current research interests is big data analysis.

Hongwei Ge received Ph.D degree in computer application technology from Jilin University, in 2006. He is currently an associate professor with the College of Computer and Science, Dalian University of Technology, Dalian, China. His current research interests include machine learning, computational intelligence, optimization and modeling, system control. He has published more than 70 papers in these areas. His research was featured in the IEEE Transactions on Cybernetics, the Pattern Recognition, the Information Science, the Computers and Structures and so on.