doi:10.3772/j.issn.1006-6748.2021.01.007

Clustering residential electricity load curve via community detection in network①

Huang Yunyou (黄运有)[*,**], Wang Nana[**,***], Hao Tianshu[**,***], Guo Xiaoxu[**], Luo Chunjie[**,***], Wang Lei[**], Ren Rui[**], Zhan Jianfeng②[**,***]

([*] School of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, P. R. China)

([**] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, P. R. China)

([***] University of Chinese Academy of Sciences, Beijing 100049, P. R. China)

Abstract

Performing analytics on the load curve (LC) of customers is the foundation for demand response which requires a better understanding of customers' consumption pattern (CP) by analyzing the load curve. However, the performances of previous widely-used LC clustering methods are poor in two folds: larger number of clusters, huge variances within a cluster (a CP is extracted from a cluster), bringing huge difficulty to understand the electricity consumption pattern of customers. In this paper, to improve the performance of LC clustering, a clustering framework incorporated with community detection is proposed. The framework includes three parts: network construction, community detection, and CP extraction. According to the cluster validity index (CVI), the integrated approach outperforms the previous state-of-the-art method with the same amount of clusters. And the approach needs fewer clusters to achieve the same performance measured by CVI.

Key words: smart meter data, electricity load curve (LC), clustering methods, community detection, demand response (DR)

0 Introduction

With the rapid development of economy and society, electricity consumption increased to 5802.00×10^9 kWh in 2015 in China and aggravated energy crisis[1]. It is becoming much more pressing to increase energy efficiency and reduce emissions. Demand response (DR) has been proposed as an important tool to improve a utility's economic and energy efficiency, reduce emissions, and integrate renewables[2]. To achieve it, an essential step is to analysis the electricity consumption behavior. Analysis of electricity consumption behavior above is a task to discover unknown knowledge using load curves (LC) to support the development of DR. As the most common-used LC analysis technology, the clustering method is an unsupervised approach to discover the electricity consumption pattern (CP) in data set by grouping similar LC into the same sub-group.

Currently, various clustering methods have been applied to LC clustering[3], such as K-means, fuzzy clustering, hierarchical clustering, and self-organizing maps. However, the performance of previous LC clustering methods is poor in two folds: larger number of clusters, huge variances within a cluster. The reasons are as follows.

(1) As each LC in the data set is treated as an individual time series during the clustering process, the indirect relationship among LCs is hard to be directly considered, restraining the performance of the clustering methods[4]. Here, the indirect relationship refers to the relationship that requires the participation of one or more third parties.

(2) Due to the significant volatility and uncertainty of the residential electricity LC, most previous clustering methods easily result in a poor performance[2,5].

Although reason (1) is raised in the general domain, it has not been considered in LC clustering. Reason (2) is a common problem for LC clustering. Current state-of-the-art clustering method is dynamic time warping (DTW)[2]. However, when using the state-of-the-art clustering method, a new problem is introduced. CP, which is extracted by traditional averaging method, has a significant difference from the corresponding LCs, bringing huge difficulty to understand the real electricity consumption pattern of customers.

① Supported by the Major Program of National Natural Science Foundation of China (No. 61432006).

② To whom correspondence should be addressed. E-mail: zhanjianfeng@ ict.ac.cn

Received on Mar. 18, 2020

In this paper, an integrated framework clustering method incorporated with community detection (CICD) is proposed to improve LC clustering performance. CICD consists of network construction, community detection, and CP extraction. First, an LC data set is converted into an ε-nearest neighbor network (ε-NN) using the distance measurement DTW, and characterize both direct and indirect inherent relationships between any pair or any group of LC. Second, a modularity-based algorithm Louvain is employed to synchronously optimize the local and global modularity, and obtain optimal community partitioning, where a community represents a cluster. Third, the centers, each of which represents a CP, is extracted from clusters using the averaging method (DTW Barycenter averaging[6]) to obtain the typical electricity CP of the customers. Compared with the state-of-the-art method, the proposed method has a significant improvement in terms of the metric of common cluster validity indexes.

The contributions of this paper are as follows.

(1) The indirect relationship between LCs directly is made accessible and usable by converting the dataset into a network.

(2) LC clustering is performed by a high-efficient community detection algorithm named Louvain, which is robust to noise.

(3) The DTW Barycenter averaging is applied to extract the CPs of the customers. It makes the CP more precise.

(4) Combined with distance measurement DTW, community detection algorithm Louvain, and TLP extraction method DTW Barycenter averaging, a clustering framework CICD is proposed to cluster residential electricity load curve and extract the CP of customers.

The paper is structured as follows. Section 1 summarizes the previous LC clustering methods. Section 2 introduces the innovative approach. Section 3 presents the experiment setup. Section 4 presents and discusses the results. Section 5 draws a concluding remark.

1 Related work

Various methods are proposed to cluster LC, including K-means, self-organizing maps, and hierarchical clustering. According to the clustering criterion, these commonly-used methods are roughly grouped into four categories: partitioning, hierarchical, density-based, and model-based[3].

Due to the simplicity and low time complexity, the partitioning methods are the most commonly-used ones in LC clustering. The partitioning methods initially select K centroids, and iteratively update these centroids to optimize the cost function[7]. Typical examples include K-means[8-9], K-medoids[2,10], and fuzzy C-means[11-12].

The hierarchical methods include two types: agglomerative and division. The agglomerative one is the most commonly used method[13]. For the agglomerative methods, each LC is firstly initialized to be a cluster. And, the two closest clusters are iteratively combined into a new one, thus reducing the number of clusters[14]. In addition, in Ref. [8], hierarchical clustering is employed to merge the clusters generated by the K-means, and reduce the number of the clusters.

The density-based methods consider the LC in high-density regions of the space as clusters, and the ones in low-density regions as outliers or noise[7,15]. In Ref. [16], a density-based method is proposed to cluster the customers using the daily load curve[17]. In Ref. [18], the classic density-based method (density-based spatial clustering of applications with noises) is applied to cluster the daily residential LC.

The model-based method assumes a model for each cluster, and finds the best fit of the data for the given model[7]. In Ref. [5], finite mixture models (Gaussian mixture models) are applied to cluster the LC. In Ref. [19], self-organizing maps model is applied to cluster the LC.

In addition to the above methods, several new methods, such as spectral clustering[20], hierarchical K-means[21], support vector clustering[22], and the iterative self-organizing data-analysis technique algorithm[23] have been proposed for LC clustering.

For LC clustering methods above, each LC in the data set is treated as an individual time series, ignoring the inherent relationship between LCs, restraining the performance of the clustering methods. However, network is a powerful tool to represent the inherent relationship between the LC and is able to improve the performance of time series clustering[4]. Thus, this work converts the LC data set into a network and clustering LC via the community detection in the network, filling the research gap above.

2 Methodology

In this section, a detailed description of the proposed method is given, which consists of 4 steps: data preparation, network construction, community detection, and CP extraction.

2.1 Data preparation

The normalization is an indispensable step in LC clustering, as it avoids the amplitude interference and

easily identifies user consumption pattern contained in the LC. In our work, for a given daily LC s_t, a normalized daily LC l_t is obtained by

$$l_t = \frac{s_t}{a} \quad \text{and} \quad a = \sum_{t=1}^{n} s_t \tag{1}$$

here, a is the sum electricity consumption of a customer for one day, and n is the number of the point in LC.

2.2 Network construction

2.2.1 Dynamic time warping

The pattern of an LC, which is presented by the shape of LC, is the most important feature in the LC clustering task. Thus, the shape-based distance measurement is the most appropriate method of measuring the similarity between the LCs. In this work, the famous shape-based distance measure DTW, which is used to find the optimal alignment between two time series and calculate the distance between them[24], is used to measure the distance between the LCs. For two given LCs $X = \{x_0, x_1, \cdots, x_{n-1}\}$ and $Y = \{y_0, y_1, \cdots, y_{n-1}\}$, the warping path $P = \{p_0, p_1, \cdots, p_k\}$, $p_k = (i_k, j_k) \in [0:n-1] \times [0:n-1]$ is obtained according to the following restrictions[2].

(1) Boundary condition: $p_0 = (0,0)$ and $p_k = (n-1, n-1)$.

(2) Monotonicity condition: $i_{k-1} \leqq i_k$ and $j_{k-1} \leqq j_k$.

(3) Continuity condition: $i_k - i_{k-1} \leqq 1$ and $j_k - j_{k-1} \leqq 1$.

(4) Warping window: $|i_k - j_k| < \omega$.

Here, i_k is the index of X, and j_k is the index of Y. In addition, we assume that a typical consumer exhibits less than 1 h variation in time of electricity consumption. As a case, when the customer electricity consumption data are collected every 15 minutes, the warping window ω of DTW is set to 4.

The minimum total cost of the warping path $c_{P(X,Y)}$ is equal to the distance between the LC X and Y. And, c_P is obtained by Eq. (2).

$$\begin{aligned} c_{P(X,Y)} &= \eta(n-1, n-1) \\ \eta(i,j) &= \delta(i,j) \\ &+ \min[\eta(i-1,j), \eta(i-1,j-1), \eta(i,j-1)] \\ \delta(i,j) &= \begin{cases} d(x_i, y_j) & |i-j| < 4 \\ Max\_value & \text{else} \end{cases} \end{aligned} \tag{2}$$

where, $\eta(i,j)$ is the cost of the warping path $P = \{p_0(x_0, y_0), p_1(x_0, y_1), \cdots, p_k(x_i, y_j)\}$, $\delta(i,j)$ is the distance between point x_i and y_j.

2.2.2 Converting an LC data set into a network

A network is a powerful tool, and it is useful for representing the complex relationship between the LCs. A network $G(V, E)$ consists of n vertices $V = \{v_0, v_1, \cdots, v_{n-1}\}$ and m edges $E = \{e(v_i, v_j) \mid v_i, v_j \in V\}$, where $e = (v_i, v_j)$ is an edge that connects v_i and v_j. In this work, we convert the LC data set into an ε-nearest neighbor network (ε-NN) as following steps.

(1) Every LC_i is considered as a vertex v_i in a network G.

(2) For each v_i, respectively calculate the distance between the v_i and other vertices by DTW.

(3) For each v_i, obtain a distance threshold ε using Eq. (3).

(4) For each v_i, if the distance between the v_i and other v_j is less than the distance threshold ε, an edge e that connects v_i and v_j is added to the network G.

$$\begin{aligned} \varepsilon &= \lambda \times \mu_d \\ \mu_d &= \frac{1}{n} \sum_{j=0}^{n-1} d(v_i, v_j) \end{aligned} \tag{3}$$

where, μ_d is the mean distance between v_i and all of the other vertices, λ is a parameter to adjust the number of the edge of the network. The higher λ, the more edges a network contains and the more complex the network is.

In addition, the weight of the edge w_e is determined by

$$\begin{aligned} w_e &= 1 - \frac{d(v_i, v_j)}{d_{\max}}, \ e = (v_i, v_j) \in E \\ d_{\max} &= \text{Max}[d(v_k, v_l) \mid e = (v_k, v_l) \in E] \end{aligned} \tag{4}$$

where, $d_{\max}$ is the max distance between the vertices in the network.

2.3 Community detection

Community detection is a method to discover apriori unknown patterns in a network, and has been attracting a lot of attention[25], as most real-world systems can be modeled by networks[4]. In this work, the most commonly used modularity-based algorithm Louvain is used to extract communities from a network, as it has advantages on computation time and quality of the community detection in terms of the metric of the modularity of a partition Q in Eq. (5)[25]. As shown in Algorithm 1, Louvain iteratively extracts communities from a network to find the optimal partition.

$$\begin{aligned} Q &= \frac{1}{2m} \sum_{i,j \in V} \left[A_{i,j} - \frac{k_i k_j}{2m}\right] \delta(c_i, c_j) \\ m &= \frac{1}{2} \sum_{u,v \in V} A_{u,v} \\ \delta(c_i, c_j) &= \begin{cases} 1 & c_i = c_j \\ 0 & \text{else} \end{cases} \\ k_i &= \sum_{i,l \in V} A_{i,l} \end{aligned} \tag{5}$$

here, $A_{i,j}$ is the weight of the edge between vertices i and j, k_i is the sum of the weights attached to i, c_i is the community to which i is assigned, and m is the total weight of the network.

Algorithm 1 Louvain

Input: Network $G(V,E)$.
Output: Communities C_{final}
Set *Change* = *True*
while *Change* do
  Set each vertex $v_i \in V$ as a different community $c_i \in C$.
  Set *Change* = *False*
  Set *LocalChange* = *True*
  While *LocalChange* do
    Set *LocalChange* = *False*
    for all Vertex $v_j \in V$ do
      Set the vertices which directly connect to vertex v_j as V_{nei}.
      for all Vertex $v_l \in V_{nei}$ do
        Set the community of v_j as c_u and the community of v_l as c_v.
        if $c_u \neq c_v$ then
          Calculating the ΔQ for moving v_j into c_v by Eq. (6).
        end if
      end for
      if The max value of ΔQ large than 0 then
        Move v_j to the c_v which obtain the max ΔQ.
        Set *LocalChange* = *True*
      end if
    end for
    Set *Change* = *LocalChange* || *Change*
  end while
  if *Change* then
    for $c_i \in C$ do
      Merge the c_i.
    end for
    Construct the new Network G.
  end if
end while
return C

$$\Delta Q = \left[\frac{\sum_{in} + \gamma k_{j,in}}{2m} - \left(\frac{\sum_{tot} + k_j}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_j}{2m}\right)^2\right] \quad (6)$$

here, $\sum_{in}$ is the sum of the weights of the edges inside a community c_v, $\sum_{tot}$ is the sum of the weights of the edges incident to the vertices inside c_v, k_i is the sum of the weights of the edges connected to vertex i, $k_{i,in}$ is the sum of the weights of the edges from i to the vertices in c_v, and m is the sum of weights of a network[25], γ is a parameter to adjust the number of communities. Usually, the smaller γ is, the more communities are extracted from a network.

2.4 Typical electricity consumption pattern extraction

The center of an LC cluster is a CP that reflects an electricity consumption pattern. Usually, the center of an LC cluster is obtained by an averaging method. However, when DTW is used as a distance measurement, it is a difficult task to average an LC cluster, because it has to be consistent with the ability of DTW to realign sequences over time[6]. In this paper, as shown in Algorithm 2, an averaging technology, named DTW Barycenter averaging[6], is used to extract CPs from the clusters.

Algorithm 2 DTW Barycenter averaging

Input: All of the LC $lc \in C$ and the initial center *lcc*.
Output: *lcc*
Set a table with size n as CT, and n equals to the length of *lcc*.
While *lcc* is not stable **do**
  Set $CT = [\varphi, \varphi, \cdots, \varphi]$
  for $lc_i \in C$ do
    Obtain the warping path $P = DTW(lcc, lc_i)$
    for $p_k = (i_k, j_k) \in P$ do
      Put $lc_i[j_k]$ to $CT[i_k]$.
    end for
  end for
  for $0 \rightarrow n-1$ do
    $ct = CT[i]$
    $lcc[i] = \frac{\sum_{x \in ct} x}{|ct|}$
  end for
end while

3 Experiment setup

3.1 Data description

To evaluate the performance of the proposed method, the experiments on two most used data sets are conducted. The first one was provided by the Pecan Street Inc.[26], which is marked as Dataport. It contains 22 113 daily LCs from 351 households, and the LCs were collected from the houses for a period of 63 d from July 6, 2015 to September 6, 2015. The sampling interval of the smart meter is 15 min. The second one is

from smart metering electricity customer behaviour trials (CBTs) in Ireland[27], which is marked as CBT. It contains 58 527 LCs from 929 households and the LCs were collected from the houses for a period of 63 d from July 6, 2010 to September 6, 2010. The smart meter data is half-hourly sampled electricity consumption (kWh) data from each customer.

3.2 The cluster validity indexes

The LC clustering is an unsupervised process, and usually lacks of the 'gold standard'[28]. The internal cluster validity index (CVI), which does not need to reference any 'gold standard', is the common way to measure the quality of clustering results in recent LC clustering studies[2,29]. However, there is no single internal CVI outperforms the rest, and every internal CVI may be able to prefer any method just in the way how the indices are proposed[30]. Thus, it is necessary to select several internal CVIs, which have excellent performance in evaluating different clustering algorithms applied to different datasets, to diminish the limitation of the internal CVI. According to a previous study, which investigated 30 commonly used CVIs in many different environments with different clustering methods (3 clustering methods) and different datasets (20 real datasets and several synthetically generated datasets), Davies-Bouldin index (*DB*, the most commonly used internal CVI), VCN index (*VCN*, an improvement of Silhouette proposed in recent study) and COP index (*COP*) have better performance in most environments and are selected in this work[28,31-33]. In addition, because the demand-response is one of the most important applications in smart grid, the *entropy*, a common index to identify demand-response potential customers, is also considered to measure the clustering results in our work[2]. The smaller *DB* and *COP* are, the higher *VCN* is, the higher cluster quality is.

Davies-Bouldin index is a very commonly used index. It assesses the cohesion according to the distance between the points and the centroid of the cluster, and the separation according to the distance between centroids[28].

$$\begin{aligned} DB &= \frac{1}{k}\sum_{c_i\in C}\max_{c_j\in C,\, c_j\neq c_i}\left\{\frac{S_{c_i}+S_{c_j}}{d(lcc_i,\, lcc_j)}\right\} \\ S_{c_i} &= \frac{i}{|\,c_i\,|}\sum_{x_m\in c_i} d(x_m,\, lcc_i) \end{aligned} \tag{7}$$

here, k is the number of the cluster, c_i is a cluster, lcc_i is the center of the c_i, $d(\_,\_)$ is the dissimilarity function and S_{ci} is the average dissimilarity between the instance x_m, which belongs to c_i, and its center lcc_i.

VCN index is an improvement of Silhouette and recently proposed. It is used to evaluate the clustering quality according to the cluster center and the nearest neighboring cluster[32].

$$\begin{aligned} VCN &= \frac{1}{k}\sum_{i=1}^{k}\frac{bd(c_i)-wd(c_i)}{\max\{bd(c_i),\, wd(c_i)\}} \\ bd(c_i) &= \min_{1\leqslant j\leqslant k,\, j\neq i}\left(\frac{1}{|\,c_i\,|}\sum_{m=1,\, x_m\in c_i}^{|c_i|} d(x_m,\, lcc_j)\right) \\ wd(c_i) &= \frac{1}{|\,c_i\,|}\sum_{m=1,\, x_m\in c_i}^{|c_i|} d(x_m,\, lcc_i) \end{aligned} \tag{8}$$

where, k is the number of the cluster, c_i is a cluster, lcc_i is the center of the c_i, $d(\_,\_)$ is the dissimilarity function, x_m is an instance which belongs to c_i, $bd(c_i)$ is the minimum between-cluster dissimilarity, and $wd(c_i)$ is the within-cluster dissimilarity.

COP is a ratio-type index and it assesses the cohesion by the distance between the points in a cluster and their centroid, and separation by the distance from points to the furthest neighbor[33].

$$COP = \frac{1}{n}\sum_{c_i\in C}|\,c_i\,|\,\frac{\dfrac{1}{|\,c_i\,|}\displaystyle\sum_{x_m\in c_i} d(x_m,\, lcc_i)}{\min\limits_{x_l\notin c_i}\max\limits_{x_m\in c_i} d(x_m,\, x_l)} \tag{9}$$

here, n is the number of the instance in dataset, c_i is a cluster, lcc_i is the center of the c_i, $d(\_,\_)$ is the dissimilarity function and x_m is an instance which belongs to c_i.

The *entropy* is used to estimate the variability of a consumer according to the cluster which the LC is assigned to[2].

$$entropy = -\sum_{i=1}^{k} p_i \log(p_i) \tag{10}$$

where, k is the number of the cluster and p_i is the ratio between the number of load curve that falls into cluster c_i and the total number of load curves[2].

3.3 Clustering method for comparison

CICD is compared with the previous clustering methods which are applied to LC clustering: K-medoids[2], K-means, agglomerative clustering (AC), Gaussian mixture models (GMM), spectral clustering (SC), and clustering by fast search and find of density peaks (CFSFDP)[17] on the same data set. The details on the models above are as follows.

(1) K-medoids is applied to LC clustering and achieves the state-of-the-art performance, combined with the distance measurement DTW[2].

(2) K-means is a most used partitioning method to cluster LC.

(3) AC is a hierarchical method, and the most used hierarchical clustering algorithm in LC clustering.

(4) GMM is a model-based clustering algorithm which is based on the expectation-maximization (EM) algorithm.

(5) SC is a graph-based clustering algorithm. Similar to the CICD, SC also converts the dataset into network.

(6) CFSFDP is a density-based clustering algorithm that is proposed to cluster the customers using the daily load curve[16].

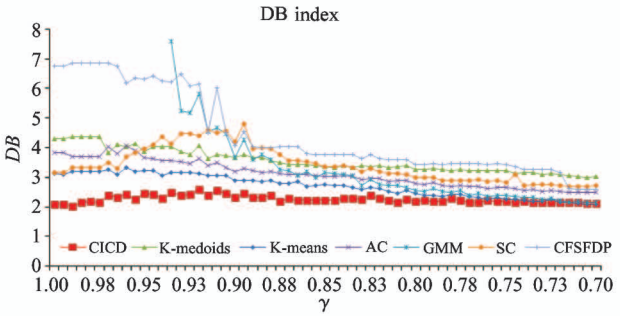
4 Results and discussion

To evaluate the efficiency of our method over state-of-the-art approaches, different clustering methods are validated on two datasets in the same task.

4.1 Cluster performance evaluation

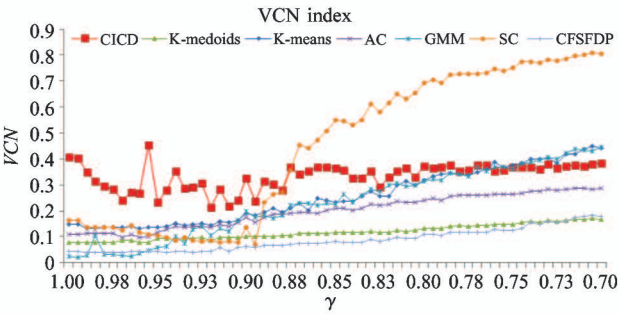
With the metrics of *DB*, *COP*, and *VCN*, we evaluate the CICD quality compared with other clustering methods.

For the Dataport dataset, Fig. 1, Fig. 2, and Fig. 3 show the value of *DB*, *VCN*, *COP* after our method is used to cluster LC with different γ. Similarly, the value of *DB*, *VCN*, *COP* of the results of other clustering methods are also respectively shown in Figs 1-3, where the cluster number of other clustering methods are set to be the same as the one in our method. In addition,



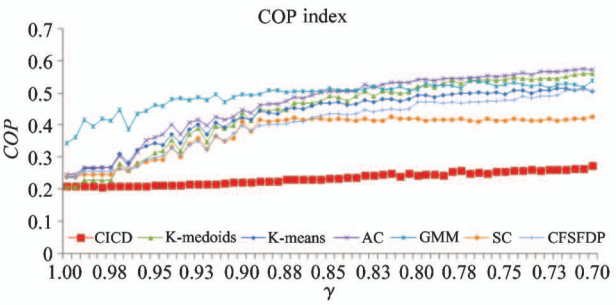
CICD is better than other algorithms in terms of *DB*, especially when γ is larger

Fig. 1 DB index on dataport



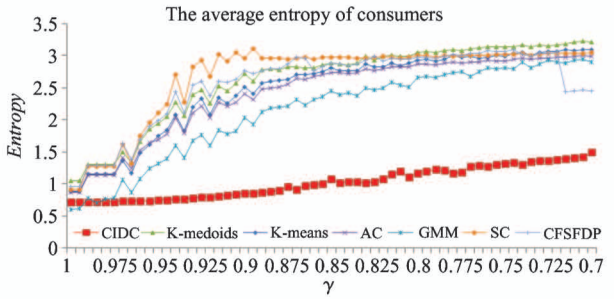
CICD is better than other algorithms (except SC) in terms of *VCN*

Fig. 2 VCN index on dataport



CICD is better than other algorithms in terms of *COP*, especially when γ is smaller

Fig. 3 COP index on dataport



The result of CICD has lower entropy than that of other algorithms

Fig. 4 The average entropy of consumers on dataport

the cluster number increases with the decrease of γ.

In general, the following phenomena are observed.

(1) When the cluster number is the same as the one of other clustering methods, the results of CICD have smaller *DB* and *COP*, and higher *VCN* respectively. It means that the CICD outperforms other clustering methods in general.

(2) With the change of γ (or cluster number), the *DB*, *COP*, and *VCN* of the CICD are more stable compared with the corresponding values of the other clustering methods, especially for the value of *DB*, *COP* and *VCN*. It means that the cluster number has less influence on the CICD.

(3) Compared with the other clustering methods, only the CICD obtains the best performance in *DB*, *COP* and *VCN* when the γ is relatively large (or cluster number is relatively small). It means that the CICD has excellent performance when the cluster number is small, and the CICD tends to cluster the LC into a small number of cluster.

Except for the phenomena above, it is observed that the value of *DB* of CICD is significantly smaller than the values of other clustering methods and the value of *VCN* of CICD is significantly higher than these values of other clustering methods when the γ is relatively large (or cluster number is relatively small). It means that compared with other clustering methods the

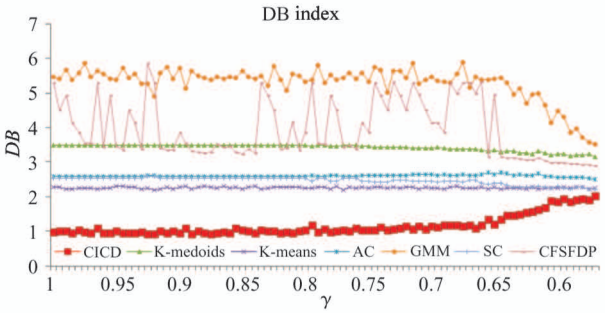
CICD has significant advantages when the γ is relatively large (or cluster number is relatively small). Besides, the *VCN* of the SC is significantly higher than the values of other clustering methods when $\gamma < 0.87$ (or cluster number is larger than 190) as shown in Fig. 2. However, researchers are more concerned about the performance of the clustering method when the γ is relatively large (or cluster number is relatively small). As when the $\gamma > 0.87$ (or cluster number is smaller than 190) the *VCN* of CICD is higher than the values of other clustering methods. It is concluded that the CICD has significant advantages compared with other clustering methods.

Furthermore, as shown in Table 1, compared with the other clustering methods, CICD has obtained the best mean value of cluster validity indexes except *VCN* (the mean value of CICD in second place). Compared with the K-medoids recently proposed, CICD has a huge improvement. The minimum improvement is 36.45% in *DB* and the maximum improvement is 184.13% in *VCN*. Compared with SC clustering method which works on graph (or network), though the value of VCN of CICD is smaller than the one of SC, the DB, and COP have significant improvement.

Table 1 The statistics of the mean cluster validity indexes of the clustering results on Dataport

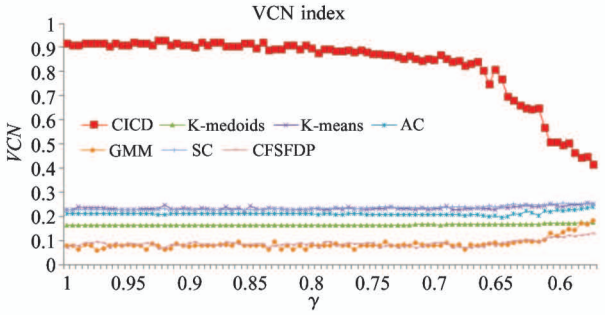
	DB	*VCN*	*COP*	*entropy*
CICD	**2.2569**	0.3347	**0.2350**	**1.0235**
K-medoids	3.5513	0.1178	0.4400	2.6156
K-means	2.7142	0.2564	0.4307	2.4703
AC	3.1158	0.1994	0.4648	2.4122
GMM	5.0560	0.2327	0.4919	2.1346
SC	3.4198	**0.4466**	0.3758	2.6729
CFSFDP	4.4449	0.0881	0.4026	2.5588
Improvement from K-medoids to CICD	36.45%	184.13%	46.59%	60.87%
Improvement from SC to CICD	34.00%	−25.06%	37.47%	38.29%

For the CBT dataset, Fig. 5 – 7 and Table 2 show the clustering results of the CICD and other clustering methods. Similar to the results on Dataport, it is evident to be observed that the performance of CICD is superior to the other clustering methods with the metrics of *DB*, *COP*, and *VCN*. What's more, it is also observed that the performance of the CICD on CBT is more excellent than the one on Dataport. For example, the improvement from K-medoids to CICD on CBT is 66.03%, 404.68%, and 44.86% with the metrics of *DB*, *VCN*, and *COP* respectively. The improvement from



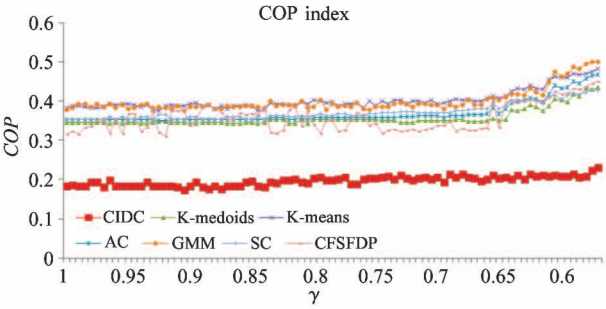
CICD is better than other algorithms in terms of *DB*, especially when γ is larger

Fig. 5 DB Index on CBT



CICD is better than other algorithms in terms of *VCN*

Fig. 6 VCN Index on CBT



CICD is better than other algorithms in terms of *COP*, especially when γ is smaller

Fig. 7 COP Index on CBT

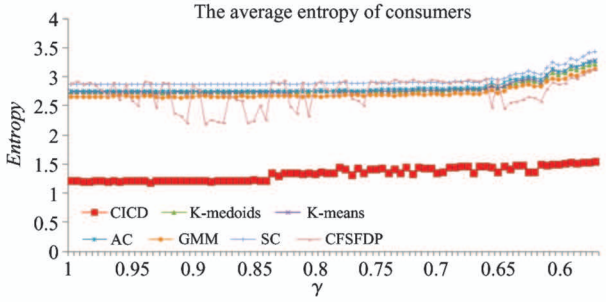
Table 2 The statistics of the mean cluster validity indexes of the clustering results on CBT

	DB	*VCN*	*COP*	*entropy*
CICD	**1.1641**	**0.8302**	**0.1970**	**1.3387**
K-medoids	3.4271	0.1645	0.3573	2.7975
K-means	2.2606	0.2337	0.4028	2.8055
AC	2.6051	0.2104	0.3690	2.8259
GMM	5.2864	0.0858	0.4000	2.7225
SC	2.4653	0.2347	0.3716	2.9403
CFSFDP	3.9664	0.0877	0.3579	2.7370
Improvement from K-medoids to CICD	66.03%	404.68%	44.86%	52.15%
Improvement from SC to CICD	52.78%	846.65%	46.99%	54.47%

K-medoids to CICD on Dataport is 36.45%, 184.13%, and 46.59% with the metrics of *DB*, *VCN*, and *COP* respectively.

In addition, classifying consumers into stable representative groups is helpful to understand and predict individual energy consumption patterns according to the recent study[2]. Thus, the stability of the consumer needs to be investigated when the LC of the consumer are assigned into different clusters by the clustering method.

The *entropy* of consumers decreases as the stability of consumers increases. It is observed that the *entropy* of the CICD is significantly lower than the *entropy* of the other clustering method as shown in Fig. 4 and Fig. 8. Besides, as shown in Tables 1 and 2, the mean value of *entropy* of CICD is less than 50% of the *entropy* of other clustering methods. It means that the CICD is able to classify the consumer into representative groups in a much more stable manner according to the assignment of consumer LC.



The result of CICD has lower *entropy* than that of other algorithms

Fig. 8 The average entropy of consumers on CBT

It is concluded that CICD outperforms other clustering methods according to CVIs (*DB*, *VCN*, and *COP*), though the SC outperforms CICD in *VCN* when cluster number is larger than 190 on Dataport. Besides, CICD is able to classify the consumers into more stable groups that is helpful to understand and predict individual energy consumption patterns.

4.2 The extraction of cluster center

The center of the LC cluster is usually considered as the CP of the customer, and researchers are able to understand the electricity consumption behavior of the customer according to the CP. As shown in Fig. 9, it is observed that every LC in the cluster has very violent fluctuation all the time. However, as shown in Fig. 10, the center extracted by the traditional averaging method is smoother than the one extracted by the DTW Barycenter averaging. It means that the fluctuation of the LC in the cluster is hardly preserved by the traditional averaging method, and inversely easily preserved by the DTW Barycenter averaging. In addition, the electricity consumption pattern preserved by DTW Barycenter averaging is more precise than the one preserved by the traditional averaging method. For example, it is observed that near the time 38 there is a short consumption peak period in cluster #1. However, the center extracted by the traditional averaging method contains 2 small consumption peak periods near the time 38 instead of a large peak period. The reason may be that the consumption peaks of different LCs are short and not simultaneous. Inversely, the center extracted by the DTW Barycenter averaging preserved the consumption peak period near the time 38. It is concluded that the DTW Barycenter averaging outperforms the traditional averaging method.

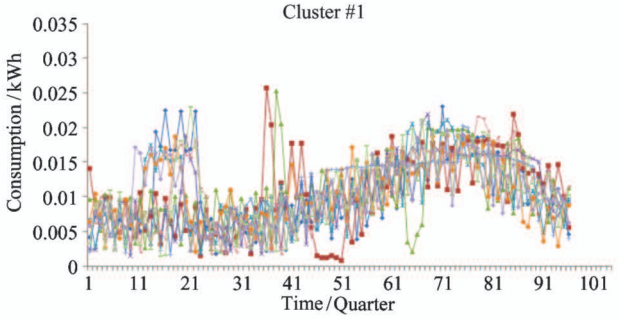


Fig. 9 A cluster of LC

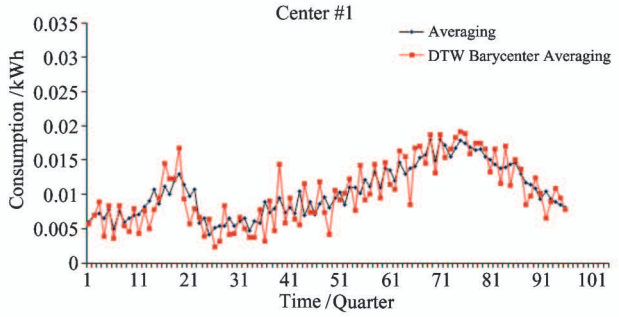


Fig. 10 The centers extracted from LC cluster

5 Conclusion

An integrated framework CICD is proposed to perform load curve clustering. The method is incorporated with community detection to improve the performance of LC clustering, which includes network construction, community detection, and CP extraction. With metrics of four cluster validity indexes (Davies-Bouldin index, VCN index, COP index and entropy), our method is validated to be effective, outperforming the other clustering methods. With the four metrics above, the cluster number has less influence on the CICD, and the CICD tends to classify the LC into fewer clusters. In addition, the center extracted by CICD significantly outperforms the traditional averaging method. Thus, CICD is able to help us obtain a better understanding of the consumption behaviors of customers.

References

[1] National Bureau of Statistics of China. Annual data[EB/OL]. http://data. stats. gov. cn/easyquery. htm? cn = C01: Stats, 2019

[2] Teeraratkul T, O'neill D, Lall S. Shape-based approach to household electric load curve clustering and prediction [J]. *IEEE Transactions on Smart Grid*, 2018, 9(5): 5196-5206

[3] Wang Y, Chen Q, Hong T, et al. Review of smart meter data analytics: applications, methodologies, and challenges[J]. *IEEE Transactions on Smart Grid*, 2018, 10(3): 3125-3148

[4] Ferreira L N, Zhao L. Time series clustering via community detection in networks [J]. *Information Sciences*, 2016, 326:227-242

[5] Li R, Li F, Smith N D. Multi-resolution load profile clustering for smart metering data[J]. *IEEE Transactions on Power Systems*, 2016,31(6):4473-4482

[6] Petitjean F, Ketterlin A, Gançarski P. A global averaging method for dynamic time warping, with applications to clustering[J]. *Pattern Recognition*, 2011, 44(3):678-693

[7] Ramos S, Duarte J M, Duarte F J, et al. A data-mining-based methodology to support MV electricity customers' characterization[J]. *Energy and Buildings*, 2015, 91:16-25

[8] Kwac J, Flora J, Rajagopal R. Household energy consumption segmentation using hourly data [J]. *IEEE Transactions on Smart Grid*, 2014, 5(1):420-430

[9] Quilumba F L, Lee W-J, Huang H, et al. Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities [J]. *IEEE Transactions on Smart Grid*, 2015, 6(2):911-918

[10] Laurinec P, Lucká M. Interpretable multiple data streams clustering with clipped streams representation for the improvement of electricity consumption forecasting [J]. *Data Mining and Knowledge Discovery*, 2019, 33(2): 413-445

[11] Yang H, Zhang L, He Q, et al. Study of power load classification based on adaptive fuzzy C means[J]. *Power System Protection and Control*, 2010,16(111-115):2238

[12] Nikolaou T G, Kolokotsa D S, Stavrakakis G S, et al. On the application of clustering techniques for office buildings' energy and thermal comfort classification[J]. *IEEE Transactions on Smart Grid*, 2012, 3(4):2196-2210

[13] Wei Y, Zhang X, Shi Y, et al. A review of data-driven approaches for prediction and classification of building energy consumption[J]. *Renewable and Sustainable Energy Reviews*, 2018, 82:1027-1047

[14] Jota P R, Silva V R, Jota F G. Building load management using cluster and statistical analyses[J]. *International Journal of Electrical Power and Energy Systems*, 2011, 33(8):1498-1505

[15] Haneen A, Noraziah A, Wahab M H A. A review on data stream classification[J]. *Journal of Physics: Conference Series*, 2018, 1018: 012019

[16] Wang Y, Chen Q, Kang C, et al. Clustering of electricity consumption behavior dynamics toward big data applications[J]. *IEEE Transactions on Smart Grid*, 2016, 7(5):2437-2447

[17] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. *Science*, 2014, 344(6191):1492-1496

[18] Jin L, Lee D, Sim A, et al. Comparison of clustering techniques for residential energy behavior using smart meter data[C] // AAAI-17 Workshop on Artificial Intelligence for Smart Grids and Buildings, San Francisco, USA, 2017: 260-266

[19] Mcloughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data[J]. *Applied Energy*, 2015, 141: 190-199

[20] Lin S, Li F, Tian E, et al. Clustering load profiles for demand response applications[J]. *IEEE Transactions on Smart Grid*, 2019,10(2): 1599-1607

[21] Xu T S, Chiang H D, Liu G Y, et al. Hierarchical K-means method for clustering large-scale advanced metering infrastructure data[J]. *IEEE Transactions on Power Delivery*, 2017, 32(2):609-616

[22] Gavrilas M, Gavrilas G, Sfintes C V. Application of honey bee mating optimization algorithm to load profile clustering[C] // 2010 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA), Taranto, Italy, 2010:113-118

[23] Mutanen A, Ruska M, Repo S, et al. Customer classification and load profiling method for distribution systems [J]. *IEEE Transactions on Power Delivery*, 2011, 26(3):1755-1763

[24] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series[C] // Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Seattle, USA, 1994: 359- 370

[25] Blondel V D, GUILLAUME J-L, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 2008(10):P10008

[26] Consortium P S R. Dataport2017[EB/OL]. https://www. pecanstreet. org/: Pecanstreet, 2017

[27] SMART C E. Commission for Energy Regulation. Metering customer behaviour trials (CBT) findings report [EB/OL]. https://www. cru. ie/wp-content/uploads/2011/07/cer11080ai. pdf,: CRU, 2021

[28] Arbelaitz O, Gurrutxaga I, Muguerza J, et al. An extensive comparative study of cluster validity indices [J]. *Pattern Recognition*, 2013, 46(1):243-256

[29] Wang Y, Chen Q, Kang C, et al. Clustering of electricity consumption behavior dynamics toward big data applications[J]. *IEEE Transactions on Smart Grid*, 2016, 7(5):2437-2447

[30] Maulik U, Bandyopadhyay S. Performance evaluation of some clustering algorithms and validity indices[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(12):1650-1654

[31] Davies D L, Bouldin D W. A cluster separation measure [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,1979(2):224-227

[32] Zhou S, Xu Z. A novel internal validity index based on the cluster centre and the nearest neighbour cluster[J]. *Applied Soft Computing*, 2018, 71: 78-88

[33] Gurrutxaga I, Albisua I, Arbelaitz O, et al. SEP/COP: an efficient method to find the best partition in hierarchical clustering based on a new cluster validity index[J]. *Pattern Recognition*, 2010, 43(10):3364-3373

Huang Yunyou, born in 1990. He is an assistant professor in computer science at the School of Computer Science and Information Technology, Guangxi Normal University. He received his Ph. D degree from University of Chinese Academy of Sciences in 2020, and received his M. S. and B. S. degrees from Guangxi Normal University in 2015 and 2012 respectively. His research interests include data analytics in smart grid and machine learning.