Electricity Consumption Behaviour Analysis Based on Time Sequence Clustering

Haibo Li¹,a, Zheng Zhang¹, Xinyi Wang², Mushi Zhou¹ and Shujie Li¹

¹ College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
² School of Automation Science and Engineering, South China University of Technology, Guangzhou, 510641, China

a Corresponding author: lihaibo@hqu.edu.cn

Abstract. Accurately grasping electricity consumption behaviour and trend in different period can provide decision support for electric resource scheduling. It is necessary to analyze the electricity consumption time sequence. To serve the purpose, a clustering method called VDSI (Variance of Difference between two Sequence Items) is presented by analyzing the typical daily electricity consumption. In the method, time sequence of electricity consumption is split into sub-sequences based on a predefined time window. Firstly, any two time sub-sequences are converted to a new one, in which any item is the difference between items of the two sub-sequences. Then, the variance is used to measure the distance of the two sub-sequences. Finally, hierarchical clustering algorithm is used to cluster sequences in different time window. By comparing with traditional sequence clustering, the VDSI can obtain more accurate clustering results.

1. Introduction
Electricity consumption, as an important indicator for an enterprise business situation, can be used to measure the activity level of industrial production in an enterprise or a region[1-2]. In addition, the differences of industry development in different regions can also be reflected by electricity consumption. All these provide a basis for electric power companies to optimize resource allocation between different regions. Moreover, electricity consumption also reflect the economic development trend of a region[3-4]. During the period of economic growth, industrial production is active, and electricity consumption will rise, and vice versa. Therefore, monitoring electricity consumption behaviour is of great significance for both macroeconomic forecasting and transmission and distribution facilities.

Generally, with the widespread use of intelligent communication technologies (ICTs) and intelligent terminal devices, such as smart meter in smart grids, the amount of data accumulated by the electric power system is growing[5]. It is a new challenge to deal with the massive electricity consumption data and analyze electricity consumption behaviour in different period[6-7]. Data mining is one of the important approaches[8-11]. To cluster sequences of electricity consumption, traditional distance algorithms, including Euclidean distance, often cause class cluster too strict, so the result has a large deviation from reality.

For the reason mentioned above, we present an improved distance algorithm, called variance of difference between two sequence items (VDSI). We take hierarchical clustering algorithm as
clustering algorithm to resolve the trend of electricity consumption in different periods. Tackling the problem well can provide accurate and scientific data support for operating strategy of electric power companies, including electric power resource scheduling and electricity pricing, etc., and further improve the operational level of electric power companies.

The rest of this paper is organized as follows. In Section 2, the characteristic of electricity consumption behaviour is analyzed. In Section 3, an improved distance algorithms based on VDSI is presented. The design of the experiments and the results evaluation are presented in Section 4. Finally, in Section 5, a summary and a discussion of the future work are presented.

2. Problem Definition

To monitor the electricity consumption of users in real time, data collecting is necessary. Generally, some additional data collecting devices are connected to electricity meters, or electricity meters have the function of data collecting. The electricity consumption data is periodically transmitted to computer servers by wireless network. The electricity consumption data collected by the devices in time order is called the electricity consumption sequence (PCS), denoted as 

\[S = \langle e_1, e_2, \ldots, e_L \rangle, \]

in which \(L\) is the length of the sequence.

To analyze the electricity consumption behavior, the sub-sequences of PCS should be analyzed so that all the distribution trend can be discovered. Given a PCS \(S\), and a time window \(w\), let \(w = N\), in which \(N < L\), cluster \(S_1, S_2, \ldots, S_{L-N+1}\), that are all the sub-sequences of \(S\), as shown in Figure 1. Our concern is to obtain all class clustering \(\text{ClusterSet} = \{C_1, C_2, \ldots, C_k\}\) and their clustering centers \(\text{Cluster}\) = \(\{S_1 = \langle e_{i+1}, e_{i+2}, \ldots, e_{i+L-1} \rangle\mid 1 \leq i \leq K\}\). These clustering centers reflect electricity consumption behaviour.

![Figure 1. Electricity consumption sequence and sub-sequences](image)

3. The Clustering Algorithms based on VDSI

Generally, data directly collected by data collecting devices are digits series of electricity meter, denoted as \(a_1, a_2, \ldots, a_{L+1}\). All the data should be preprocessed and converted to electricity consumption, denoted as \(S = \langle e_1, e_2, \ldots, e_L \rangle\), in which \(e_i = a_{i+1} - a_i (1 \leq i \leq L)\).

Traditional distance algorithm is too strict for our analysis object, i.e., sequence of electricity consumption because similar trend of the sequences should also be included in a class cluster, and not just similar items of a sequence. For this reason, given two time sequences \(S = \langle x_1, x_2, \ldots, x_N \rangle\) and \(S_i = \langle y_1, y_2, \ldots, y_N \rangle\). Let \(d = |x_i - y_i|, 1 \leq i \leq N\), the distance algorithm can be defined as follows.

\[\delta(d_i) = \frac{\sum_{i=1}^{N} (d_i - D)^2}{N} \]  

(1)

In (1), \(D\) is the average value of \(d_i, 1 \leq i \leq N\).

Traditional clustering algorithm such as k-means is used wildly. However the number of clustering centers \(K\) is difficult to be determined. Especially when the amount of data is large, it is very difficult to do. According to the characteristics of sampled data, hierarchical clustering algorithm is used to cluster sequences.
In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally divided into two types: agglomerative hierarchical clustering and divisive hierarchical clustering. The former is a bottom-up approach, in which each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. The latter is a top-down approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a tree.

In this paper, we use agglomerative hierarchical clustering. In an agglomerative hierarchical clustering, each branch of the built hierarchy is a cluster. This method has several steps:

1. Calculation of distances between individuals.
2. Choice of the two nearest individuals.
3. Aggregation of the two nearest individuals in a cluster. The cluster is now considered an individual.
4. Go back to the step 1 and loop while there is more than one individual.

The results of a hierarchical agglomerative clustering can be showed as a tree which represents the distance between the individuals. The process of clustering is shown as Figure 2.

![Figure 2. The hierarchical agglomerative clustering](image)

An algorithm is presented as follows to get all class clusters from a sequence of electricity consumption.

Algorithm 1: GetClusters

Input: \( S = \langle e_1, e_2, \ldots, e_L \rangle, w, \text{threshold} \) // the sequence of electricity consumption, time window, and the threshold of similarity between clusters.

Output: \( (C, CCtrSet) \) // the set of cluster and their centres

1: \( C \leftarrow S, CCtrSet \leftarrow 0 \)
2: \( maxs \leftarrow 0 \) // maxs, to save the max similarity
3: for \( i \leftarrow 1 \) to \( L-w \)
4: for \( j \leftarrow i+1 \) to \( L-w \)
5: \( \varepsilon \leftarrow \text{CalSimilarity}(S_i, S_j) // S_i = \langle e_i, e_{i+1}, \ldots, e_{i+w} \rangle, \) //is sub-sequence of \( S \)
6: \( \text{cluster}_r \leftarrow S_i, \text{cluster}_s \leftarrow S_j \)
7: \( \text{Table} \leftarrow (\text{cluster}_r, \text{cluster}_s, \varepsilon) \)
8: endfor
9: endfor
10: while \( (\maxs \leq \text{threshold}) \)
11: for \( k \leftarrow 1 \) to \( \text{abs}(\text{Table}) \)
12: if($Table_k.\varepsilon > threshold$) then
13: $Table_k. cluster, \leftarrow Table_k.$
$cluster_r \cup Table_k. cluster_s$
14: $CCtrSet \leftarrow CalCCtrSet (Table_k. cluster_r)$
//calculate new cluster centres
15: delete $Table_k. cluster_s$
16: endif
17: if($Table_k.\varepsilon > maxs$) $maxs \leftarrow \varepsilon$
18: endfor
19: endwhile
20: return ($C, CCtrSet$)

In algorithm $GetClusters$, the function $CalSimilarity$ is used to calculate the similarity between sub-sequence of $S$, and $CalCCtrSet$ is used to calculate cluster centre. All clusters are combined if their distance $\varepsilon > threshold$. If there are $t$ clusters are combined on average every time, the clustering number is $w/t$, time complexity is $O(n^2/t)$.

4. Experiment and evaluation

4.1 The data set of the experiment

The data set of the experiment is collected from a college. To achieve energy saving and emission reduction, in this college, more than 2,000 sensors are installed on the electricity meters distributed more than 20 buildings. The data is collected at 30 min-intervals from January 10, 2015 to May 10, 2016. From all the data set of electricity consumption, we choose one. There are 19828 data of electricity consumption. The data format is shown as Table 1.

| time(Y-M-D H:M:S) | digits of electricity meter |
|-------------------|-----------------------------|
| 2015-01-10 12:57:42.0 | 81400.0                     |
| 2015-01-10 13:27:43.0 | 81407.0                     |
| 2015-01-10 13:57:44.0 | 81415.0                     |
| 2015-01-10 14:27:45.0 | 81422.0                     |
| 2015-01-10 14:57:46.0 | 81430.0                     |
| 2015-01-10 15:27:47.0 | 81437.0                     |
| 2015-01-10 15:57:48.0 | 81449.0                     |

All the data should be preprocessed and converted to electricity consumption. Because the times in Table 1 represent order relationship among the digits of electricity meter, they need not be converted only if the order relationship can be ensured. After preprocessing of data in Table 1, the sequence $S$ should be $<7,8,7,8,7,12,13,15,14>$. 

4.2 Experimental results

Using the same data set, the VDSI-based experiment was done three times. Let time window be 16, and set different distance thresholds 0.9, 1.9, 2.9 respectively, as shown in Figure 3, 4 and 5.
Mapping the major clusters to the original data, we find that the time of similar electricity consumption is almost from 22 o'clock every day to 6 o'clock the next morning. In Figure 4 and 5, some daytime of winter and summer holidays are included. We select some samples in some major clusters, as shown in Figure 6. The electricity consumption behavior is very obvious.
4.3 Comparison of experiment

We compare our clustering algorithms based on VDSI with Euclidean distance in hierarchical clustering algorithm. Firstly, using the same data set, the Euclidean distance-based experiment was done four times. Let time window be 16, and set different distance thresholds 0.01, 0.03, 0.05, 0.1 respectively, as shown in Figure 7, 8 and 9.

Figure 7 illustrates that there are several major clusters to distinguish each other distinctly. All the three figures illustrate that the major clusters are bigger and bigger with the increasing threshold. In the biggest clusters in Figure 7, 8 and 9, the sequences of electricity consumption are similar.
Finally, selecting 3 pairs of proper thresholds, comparison between the VDSI-based method and the Euclidean distance-based method is done, as shown in Table 2. The numbers of major clusters is smaller in the VDSI-based method than in the Euclidean distance-based method, because more samples are clustered in in the VDSI-based method. The comparison illustrates that, for analyzing electricity consumption behavior, the VDSI-based clustering method is more effective than the traditional clustering.

Table 2. The comparison between the VDSI-based and the Euclidean distance-based method.

|                        | the VDSI-based | the Euclidean distance-based |
|------------------------|----------------|------------------------------|
| numbers of clusters over 50 samples | 49  | 70  |
|                        | 45  | 61  |
|                        | 18  | 48  |
| average numbers of samples in major clusters | 1975 | 1246 |
|                        | 2497 | 1381 |
|                        | 5894 | 4002 |

5. Conclusions
To identify typical daily electricity consumption behavior, a method based on time sequence clustering is present to analyze electricity consumption collected from smart meters. Agglomerative hierarchical clustering is used, in which we adopt variance of difference between two sequence items to calculate distance between two sequences. The advantage is that not only similar sequence items but also similar trends of sequences are included in a class cluster. Therefore, clustering results are more accurate than traditional clustering. Recommended future work includes correlation analysis between different time sequences to discover the influence relationship between different electricity consumption users.

Acknowledgements
This work was supported in part by a grant from JZ160409 of the Natural Foundation Key Program for Young Scholars in the Universities of Fujian Province, and the Central Universities (Grant No. JB-ZR1147).

References
[1] G. Ray, S.B. Minna. Ecol. Econ. 153, 78 (2018)
[2] R.M. Tetlow, D.C. Van, C.P. Beaman, A.A. Elmualim and K. Couling. Build. Environ. 92, 75 (2015)
[3] L. Lars, A. Peter A., K. Magnus. J. Lightware Technol. 35(11), 2132 (2017)
[4] X.L. Bian, G.Q. Wang, H.D. Wang, S. Wang and W.D.Lv. Miner. Eng. 105, 22 (2017)
[5] A.L. Klingler, F. Schuhmacher. Energ. Effic. 1 (2018)
[6] O. Mohamed, I. Mohamed and M.Phil. Energ. Build. 128, 530 (2016)
[7] O. Motlagh, P. Paevere, T.S. Hong and G. Grozev. Appl. Math. Comput. 270, 165 (2015)
[8] N. Ranjbar, S.A. Zaki, N.M.Yusoff and A. Hagishim, Aya. Chem. Eng. T. 56, 379 (2017)
[9] Y.W. Chen, S.Y. T, N. Bougulia. Pattern Recogn. 83, 375 (2018)
[10] Y.W. Chen, S.Y. T, S. W. Pei, Trans. Syst., Man, Cybern., Syst. 48, 649 (2018)
[11] Y.W. Chen, S.Y. T. Inform. Sciences. 433, 510 (2018)