RESEARCH ARTICLE OPEN ACCESS

Analyzing the Behavior of Electricity Consumption Using Hadoop

Dr. Mohammed Abdul Waheed[1], Samreen Sultana[2]

[1]Assosciate Professor, Department of studies in Computer Science & Engineering,
[2]P.G.Student, Department of studies in Computer Science & Engineering,
VTU PG Centre, Kalaburgi, Karnataka, India.

Abstract:

In a competitive retail market, large volumes of smart meter data provide opportunities for load serving entities to enhance their knowledge of customers' electricity consumption behaviors via load profiling. Instead of focusing on the shape of the load curves, this paper proposes a novel approach for clustering of electricity consumption behavior dynamics, where"dynamics" refer to transitions and relations between consumption behaviors, or rather consumption levels, in adjacent periods. First, for each individual customer, symbolic aggregate approximation is performed to reduce the scale of the data set, and time-based Markov model is applied to model the dynamic of electricity consumption, transforming the large data set of load curves to several state transition matrixes. Second, a clustering technique by fast search and find of density peaks (CFSFDP) is primarily carried out to obtain the typical dynamics of consumption behavior, with the difference between any two consumption patterns measured by the Kullback–Liebler distance, and to classify the customers into several clusters. To tackle the challenges of big data, the CFSFDP technique is integrated into a divide-andconquer approach toward big data applications. A numerical case verifies the effectiveness of the proposed models and approaches.

Keywords — **Load profiling, big data, Markov model, electricity consumption, behavior dynamics, distributed clustering, demand response.**

I. INTRODUCTION

Nations around the globe have set forceful objectives for the rebuilding of monopolistic power framework towards changed markets particularly on the request side. In a focused retail advertise, stack serving elements (LSEs) will be created in incredible numbers. Having a superior comprehension of power utilization designs and acknowledging customized control administrations are viable approaches to upgrade the aggressiveness of LSEs. In the interim, savvy frameworks have been upsetting the electrical era and utilization through a two-route stream of force and data. As a vital data source from the request side, progressed metering framework (AMI), has increased expanding prominence around the world; AMI permits LSEs to get power utilization information at high recurrence, e.g., minutes to hours. Huge volumes of power utilization information uncover data of clients that can possibly be utilized by LSEs to deal with their era and demand assets productively and give customized benefit. Stack profiling, which alludes to power utilization practices of clients over a particular period, e.g., one day, can help LSEs see how power is really utilized for various clients and get the clients' heap profiles or load designs. Stack profiling assumes a fundamental part in the Time of Use (ToU) duty outline, nodal or client scale stack estimating, request reaction and vitality proficiency focusing on, and non-specialized misfortune (NTL) discovery.

Countries around the world have set aggressive goals for the restructuring of monopolistic power system towards liberalized markets especially on the demand side. In a competitive retail market, load serving entities (LSEs) will be developed in great numbers [1]. Having a better understanding of electricity consumption patterns and realizing personalized power managements are effective ways to enhance the competitiveness of LSEs [2]. Meanwhile, smart grids have been revolutionizing the electrical generation and consumption through a two-way flow of power and information. As an important information source from the demand side, advanced metering infrastructure (AMI), has gained increasing popularity worldwide; AMI allows LSEs to obtain electricity consumption data at high frequency, e.g., minutes to hours [3].

Large volumes of electricity consumption data reveal information of customers that can potentially be used by LSEs to manage their generation and demand resources efficiently and provide personalized service. Load profiling, which refers to electricity consumption behaviors of customers over a specific period, e.g., one day, can help LSEs understand how electricity is actually used for different customers and obtain the customers' load profiles or load patterns. Load profiling plays a vital role in the Time of Use (ToU) tariff design [4], nodal or customer scale load forecasting [5], demand response and energy efficiency targeting [6], and non-technical loss (NTL) detection [7].

The core of load profiling is clustering which can be classified into two categories: direct clustering and indirect clustering [8]. Direct clustering means that clustering methods are applied directly to load data. Heretofore, there are a large number of clustering techniques that are widely studied, including k-means [9], fuzzy k-means [10], hierarchical clustering [11], self-organizing maps (SOM) [12], support vector clustering [13], subspace clustering [14], ant colony clustering [15] and etc. The performance of each clustering technique could be evaluated and quantified using various criteria, including the clustering dispersion indicator (CDI), the scatter index (SI), the Davies-Bouldin index (DBI), and the mean index adequacy (MIA) [16].

The deluge of electricity consumption data with the widespread and high-frequency collection of smart meters introduces great challenges for data storage, communication and analysis. In this context, dimension reduction methods can be effectively applied to reduce the size of the load data before clustering, which is defined as indirect clustering. Such clustering can be categorized into two sub-categories, feature extraction-based clustering and time series-based clustering. Feature extraction which transforms the data in the high-dimensional space into a space of fewer dimensions [17], is often used to reduce the scale of the input data. Principal component analysis (PCA) [18], [19] is a frequently used linear dimension reduction method. It tries to retain most of the covariance of the data features with the fewest artificial variables. Some nonlinear dimension reduction methods including Sammon maps, curvilinear component analysis (CCA) [20], and deep learning [21] have also been applied to electricity consumption data. Moreover, as electricity consumption data are essentially time series. A variety of mature analytical methods such as discrete Fourier transform (DFT) [22], [23], discrete wavelet transform (DWT) [24], symbolic aggregate approximation (SAX) [25], and the hidden Markov model (HMM) [26] have been discussed in the literature. These methods are capable of reducing the dimensionality of time series and of maintaining some of the original character of the electrical consumption data.

II. RELATED WORK

The existing studies on load profiling mainly focus on individual large industrial/commercial customer, medium or low voltage feeder, or a combination of small customers, load profiles of which shows much more regularity [25]. It should be noted that although these dynamic characteristics are always “deluged” in a combination of customers, they could be described by several typical load patterns. However, with regard to residential customers, at least two new challenges will be faced. One challenge is the high variety and variability of the load patterns. As indicated by Fig. 1, there are clear differences in the electricity consumption patterns of the two residents. Peak loads have different amplitudes and occur at different times of day, for example. Electricity consumption patterns also vary on a daily basis even for the same customer. In this case, several typical daily load patterns are not fine enough to reveal the actual consumption behaviors. The daily profile should be decomposed into more fine-grained fragments, which are dynamically changed and identified. Moreover, as the consumption behavior of a specific customer is essentially a state-dependent, stochastic process, it is important to explore the dynamic characteristics, e.g., switching and maintaining, of the consumption states and the corresponding probabilities. The other challenge is that of “big data”. Considering the high frequency and dimensionality of the data contained in the load curves, data sets in the multipetabyte range will be analyzed [27]. Traditional clustering techniques are tricky to be executed in a “big data world”.

The existing studies on load profiling mainly focus on individual large industrial/commercial customer, medium or low voltage feeder, or a combination of small customers, load profiles of which shows much more regularity. It should be noted that although these dynamic characteristics are always “deluged” in a combination of customers, they could be described by several typical load patterns. However, with regard to residential customers, at least two new challenges will be faced. One challenge is the high variety and variability of the load patterns. There are clear differences in the electricity consumption patterns of the two residents. Peak loads have different amplitudes and occur at different times of day, for example. Electricity consumption patterns also vary on a daily basis even for the same customer.

III.PROPOSED SYSTEM

The proposed methodology for the dynamic discovery of the electricity consumption can be divided into six stages. The first stage conducts some load data preparations, including data cleaning and load curve normalization. The second stage reduces the dimensionality of the load profiles using SAX. The third stage formulates the electricity consumption dynamics of each individual customer utilizing time-based Markov model. The K-L distance is applied to measure the difference between any two Markov model to obtain the distance matrix in the fourth stage. The fifth stage performs a modified CFSFDP clustering algorithm to discover the typical dynamics of electricity consumption. Finally, the results of the analysis of the demand response targeting are obtained in the sixth stage. The details of the first five stages will be introduced in the following, and the demand response targeting analysis part will be further explained in the case studies.

Advantages:

The proposed clustering method has the following advantages so that we adopt it to our study. First, CFSFDP is so elegant and simple that fewer parameters are needed with low time complexity, and it has shown high performance in classifying several data sets. After finding the density peaks, the assignment is of each object can be performed in a single step without iteration, in contrast with many other clustering methods like k-means

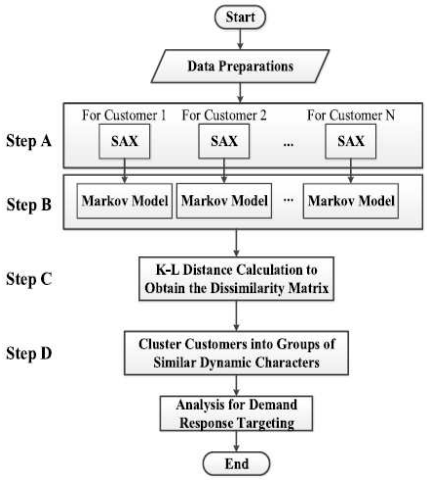


Fig. 1 Clustering of electricity consumption behavior dynamics processes.

IV. BASIC METHODOLOGY

The proposed methodology for the dynamic discovery of the electricity consumption can be divided into six stages, as shown in Fig. 2. The first stage conducts some load data preparations, including data cleaning and load curve normalization. The second stage reduces the dimensionality of the load profiles using SAX. The third stage formulates the electricity consumption dynamics of each individual customer utilizing time-based Markov model. The K-L distance is applied to measure the difference between any two Markov model to obtain the distance matrix in the fourth stage. The fifth stage performs a modified CFSFDP clustering algorithm to discover the typical dynamics of electricity consumption. Finally, the results of the analysis of the demand response targeting are obtained in the sixth stage. The details of the first five stages will be introduced in the following, and the demand response targeting analysis part will be further explained in the case studies.

V. IMPLEMENTATION

Modules:

1. Framework
2. Local Modelling-Adaptive k-means
3. Distributed algorithm for large data sets

MODULES DESCRIPTION:

Framework:

A divide-and-conquer framework for distributed clustering, where Li denotes the original data on the ith distributed local site; Mi denotes the representative objects selected from the ith distributed local site; and R denotes the global clustering results. Each object corresponds to a customer described by transition probability matrixes.

Local Modelling-Adaptive k-means:

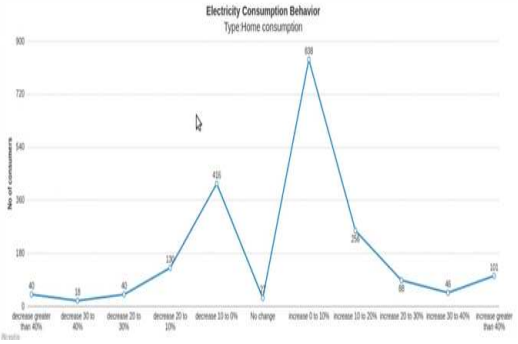
A set of clustering centres will be obtained by k-means, where the sum of the squared distances between each object is minimized. These centroids can be used as a "code book": each object can be represented by the corresponding centroid with the least error. This is called vector quantization (VQ). We try to establish a local model by finding the "code book" that guarantees that the distortion of each object by VQ satisfies the threshold condition.

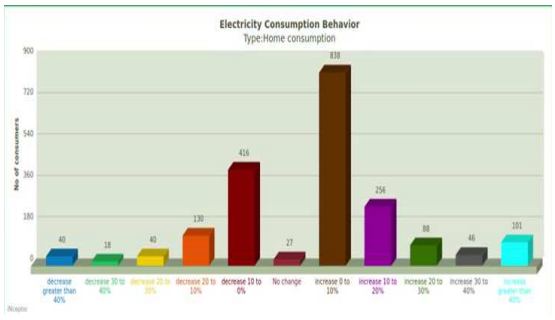
Distributed algorithm for large data sets:

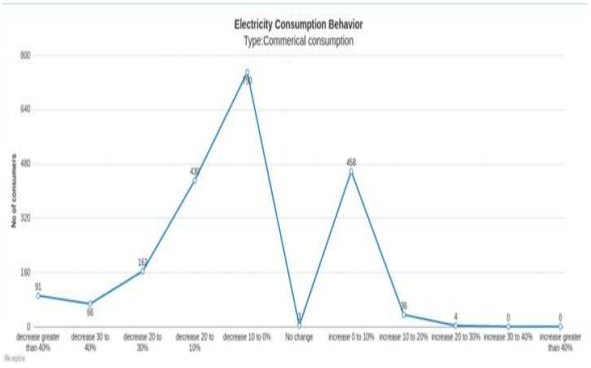
The electricity consumption data skyrocketing for population-level customers, is challenging the storage, communication and analysis of the data. Although SAX and time-based Markov model have largely reduced the dimensionality of the load profiles, the centralized clustering technique is not effective for big data challenges. On one hand, the electricity consumption data are collected and distributed on different sites. The electricity consumption data of customers are collected and stored on different substations they belong to. It is costly and time consuming to transmit whole data from each distributed site to a central site. On the other hand, the analysis and clustering of large data sets gathered from each distributed site need a very large time and memory overhead. When applying the CFSFDP, the dissimilarity matrix of all the customers should first be obtained, which accounts for most of the computation time.

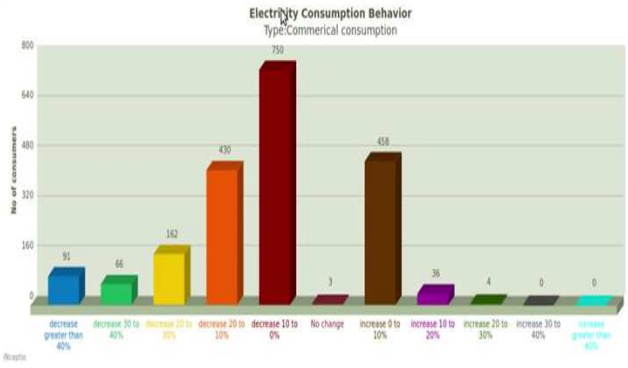
Exist many works on parallel clustering for big data applications. For these algorithms, the whole data set should reside on the same data centre and then be distributed to different clients like map-and-reduce in Hadoop. It is not satisfied with the practical situation of electricity consumption data collecting and storing. Besides, some fully distributed clustering algorithms are also proposed to tackle these challenges by aggregating the information of local data and then sending to a central site for central analysis.

VI. RESULTS









VII. CONCLUSIONS

In this paper, a novel approach for the clustering of electricity consumption behavior dynamics toward large data sets has been proposed. Different from traditional load profiling from a static prospective, SAX and time-based Markov model are utilized to model the electricity consumption dynamic characteristics of each customer. A density-based clustering technique, CFSFDP, is performed to discover the typical dynamics of electricity consumption and segment customers into different groups. Finally, a time domain analysis and entropy evaluation are conducted on the result of the dynamic clustering to identify the demand response potential of each group's customers. The challenges of massive high-dimensional electricity consumption data are addressed in three ways. First, SAX can reduce and discretize the numerical consumption data to ease the cost of data communication and storage. Second, Markov model are modelled to transform long-term data to several transition matrixes. Third, a distributed clustering algorithm is proposed for distributed big data sets. Limited by the data sets, the influence of external factors like temperature, day type, and economy on the electricity consumption is not considered in depth in this paper. Future works will focus on feature extraction and data mining techniques combining electricity consumption with external factors.

REFERENCES

[1] USA Department of Energy, Smart Grid / Department of Energy, http://energy.gov/oe/technology-development/smart-grid, 2014.

[2] I. P. Panapakidis, M. C. Alexiadis and G. K. Papagiannis, "Load profiling in the deregulated electricity markets: A review of the applications," in European Energy Market (EEM), 2012 9th International Conference on the, 2012, pp. 1-8.

[3] R. Granell, C. J. Axon and D. C. H. Wallom, "Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles," IEEE Trans. Power Systems, vol. 30, pp. 3217-3224, 2015.

[4] N. Mahmoudi-Kohan, M. P. Moghaddam, M. K. Sheikh-El-Eslami, and E. Shayesteh, "A three-stage strategy for optimal price offering by a retailer based on clustering techniques," International Journal of Electrical Power & Energy Systems, vol. 32, pp. 1135-1142, 2010.

[5] P. Zhang, X. Wu, X. Wang and S. Bi, "Short-term load forecasting based on big data technologies," CSEE Journal of Power and Energy Systems, vol. 1, no. 3, pp. 59-67, 2015.

[6] N. Mahmoudi-Kohan, M. P. Moghaddam, M. K. Sheikh-El-Eslami, and S. M. Bidaki, "Improving WFA k-means technique for demand response programs applications," in Power & Energy Society General Meeting, 2009.PES '09. IEEE, 2009, pp. 1-5.

[7] C. Leon, F. Biscarri, I. Monedero, J. I. Guerrero, J. Biscarri, and R. Millan, "Variability and Trend-Based Generalized Rule Induction Model to NTL Detection in Power Companies," IEEE Trans. Power Systems, vol. 26, pp. 1798-1807, 2011

[8] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," Tsinghua Science and Technology, vol. 20, pp. 117-129, 2015.

[9] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of Low Voltage Network Templates-Part I: Substation Clustering and Classification," IEEE Trans. Power Systems, vol. 30, pp. 3036-3044, 2015.

[10] K. Zhou, S. Yang and C. Shen, "A review of electric load classification in smart grid environment," Renewable and Sustainable Energy Reviews, vol. 24, pp. 103-110, 2013.