May – June 2020
ISSN: 0193-4120 Page No. 5690-5695

Smart Profile: Smart Meter Data Oriented Customer Profiling

Bhawna Dhupia[1], M. Usha Rani[2]
[1]Research Scholar, SPMVV, Tirupati, India
[2]Professor, Dept. of Computer Science, SPMVV, Tirupati, India
[1]bhawnasgn@gmail.com, [2]musha_rohan@yahoo.com

Article Info
Volume 83
Page Number: 5690 - 5695
Publication Issue:
May - June 2020

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *17 May 2020*

Abstract

Advanced metering infrastructure (AMI) enabled the collection of consumer status in real-time. That is the reason, this information is very critical and huge to take an important decision regarding many aspects of the smart grid. The data collected from smart meter helps to analyse consumer profiling, energy demand-response, optimization of energy generation and many more. This paper demonstrates the complete process of Energy Data Analytics starting from cleaning of real time data till clustering of customer cased on energy usage over a period of six months. Clustering is one of the techniques which grouped the data elements based on some common characteristics. This paper exhibits the implementation of the clustering technique to classify the raw data into more meaningful information using Machine Learning (ML) models. Determining the number of clusters for any dataset is one of the crucial tasks. In this paper, techniques for cluster optimization, types of clustering and its implementation are discussed in detail. The dataset for the implementation is real-time data of industrial sector from Himachal Pradesh, Solan (H.P.). Furthermore, the implementation of all the techniques is performed on the energy data set using python software.

Keywords: *AMI, Machine Learning, K-Means, Clustering, Energy Load Profile*

I. INTRODUCTION

The smart meter is one of the vital sources of AMI, which enables us to record and monitor the daily consumption data of the customer. It collects the information on data usage after every 15-30 minutes. In this way, it collects a huge amount of data for every customer. This granular collection of energy data helps to analyse the consumption pattern of the customer and the probability of energy demand. This can also help to identify the peak load usage of the energy and can help to plan the energy optimization policies. This is why research on energy data has gained momentum over a decade. While reviewing the literature on smart meter data analytics, it is found that the K-Means is a dominant clustering technique [1]. The clustering technique helps to categorize the consumer based on any criteria given by the user. When it comes to loading profiling, it is categorized based on the consumption of energy by each consumer. To get the load profile of the consumer, time-series consumption data is normally required with the time-stamp. As discussed above, the smart meter collects the data after every 15-30 minutes. So, normally it records 48-96 readings per day per consumer along with the time stamp details. This helps us to draw a load curve to show the energy consumption of the consumers.

This paper will demonstrate modern data analytics techniques to process the load profile of the customer. Although the paper will discuss a few clustering techniques, the proposed method will be K-means to cluster the customers depending upon the energy usage profile. Sometime finalizing the number of clusters is a confusing task, when the numbers of records are more. This paper will also discuss various techniques to estimate the numbers of clusters best suited to the data. The work will be implemented on the data collected from industries in India.

The paper aims at classifying the customers on the basis of smart meter energy usage data. To classify the customers, we used supervised clustering

May – June 2020
ISSN: 0193-4120 Page No. 5690-5695

algorithms, such as KMeans and agglomerative clustering techniques. Initially, section 2 describes the clustering techniques used to classify customers. It covers the detailed description of Agglomerative and KMeans clustering techniques. Section 3 deals with the techniques used to decide on a reasonable number of clusters, to classify the data. The Elbow Index Method and Average Silhouette Index Method (ASI) are the best methods used to decide on the number of clusters. Section 4 of the paper presents the complete implementation and results on the actual data followed by the discussion of the results, and section 4 concludes the paper with future work suggestions.

II. CLUSTERING TECHNIQUES

Clustering is a technique which groups data into most meaningful information depending on variable/ conditions assigned [1]. It is applied to the group of a consistent data set "where the within-group-object similarity is minimized and the between-group-object dissimilarity is maximized" [2]. Methods implemented for clustering the data in this paper are discussed in this paper are KMeans and KMediod, which are the best suited on electronics data

Hierarchical Clustering

clustering groups the data elements into a tree-like structure and gives the optimal clusters in a dataset. Hierarchical clustering is divided into two types namely; Top-Down clustering or Divisive method and Bottom-Up clustering or Agglomerative method. Top-Down or Divisive method assigns all the element to one cluster initially and then divide then into sub-clusters followed by the recursive process of selection till all the elements assigned to their most eligible cluster. Bottom-Up clustering or Agglomerative method assigns each element to its own cluster and then analyse all elements one-by-one and assign the element in the most similar cluster. It repeats this process until all the elements are assigned [3]. A dendrogram is used to represent this clustering [4].
estimate the proximity matrix, the distance between each point in the cluster is estimated before clustering. In hierarchical clustering, there are four prevalent methods used to check the distance between the points such as Single Linkage, Complete Linkage, Average Linkage, and Ward's Minimum-Variance Method. In this paper, Minimum-Variance Method of Ward's is used through agglomerative clustering using python programming.

K-means Clustering

As discussed in the introduction part also, K-means clustering is one of the most popular unsupervised machines learning clustering methods. This clustering method has proven records in the adequacy of the clustering in various applications of the smart grid [5,11]. In this method, 'K' means the number of centroids is chosen randomly initially for the data set. The iterative process of selecting the elements to the assigned centroids is performed. During the process, most suitable centroids are automatically assigned, to get the perfect result. Finally, it gives the most feasible clusters and grouped elements as a result. In this method, the number of clusters should be known in advance to assigned as centroids of the data. K-means algorithm goal is to minimize an objective function known as SEE or Squared Error function.

III. CLUSTER VALIDATION

Defining the adequate number of clusters for a dataset is a key issue especially in K-mean clustering where the specification of a number of clusters should be in advance. The requirement of evaluating the genuine number of clusters results in the discovery of various methods [6,13]. This article covers the two most popular methods for indexing an optimal number of clusters used in the smart grid namely, Elbow index [7] and Average Silhouette index [8]. All the indexes evaluate based on different properties of the data elements. So, it is advisable to apply more techniques to get the best results [9].

Elbow Index Method

The main aim of deciding on the number of clusters for any data set is to obtain the number which is optimal and can give the best clustering of elements. In the case of K-means clustering, the number of clusters should be defined so that the total Within-cluster Sum of Square (WSS) is minimized [7]. The total WSS measure the cohesion

May – June 2020
ISSN: 0193-4120 Page No. 5690-5695

and separation of the elements in the cluster which ensures the compactness of the cluster.

IV. RESULT & DISCUSSION

The categorization of the customer is one of the critical tasks, which helps in implementing the policies to improve the energy conservation [12]. Based on the energy consumption curve, customers are segmented [10]. In this section, we will present the implementation phase and results such as deciding on the number of clusters, consumer clustering based on the usage data extracted from the smart meter device. The clustering of customers is done by two methods, namely, K-Means and Agglomerative. The procedure and implementation details are presented in sections 4.1 and 4.2.

Data Summary & Preparation

Since the paper discussed all the methods and techniques applied for the procedure to calculate the energy load profile, now will discuss the information regarding the data set. This study is done on the data collected from the electricity department, Himachal Pradesh, India. This contains 69 records of the data collected by smart meter on daily electricity consumption. All meters are deployed in the industrial area; hence the data is of industrial category. The reading is recorded every 30 minutes interval. Every measurement is a load curve based on 24 hrs reading to understand consumer behavior. The fields in records were meter number, name of customer, timestamp (30min interval), and energy consumption in KVA and KW. The data was collected for a period of 8 months, starting from Dec'2018 till Jul'2019. The main aim of the study is to implement the clustering technique on the data and classify the dataset into the most suitable cluster. These clusters then get processed to calculate the average energy consumption and to find out the majority of user energy usage. This study can help to predict the load forecasting in further study. The data is cleaned and processed for further analysis using IBM SPSS analytical tools.

The data determine the consumption behavior of the companies. This data set is processed to get the energy daily consumption, average load and, average daily consumption for each cluster. The total number of meters was 69 initially, but after cleaning, identifying missing values and ignoring outlier, we have left with 61-meter records. As discussed above, to get the optimal cluster number we use Elbow Index. Figure 1 shows the result of the Elbow Index Method using Python, from which we decided to have 4 clusters as an optimal number.

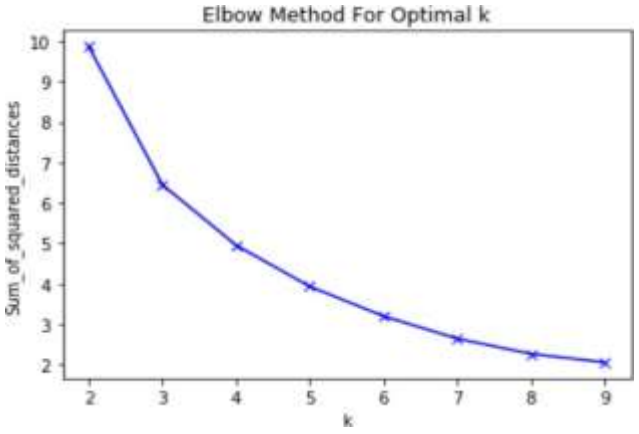

Fig 1. Optimal Cluster Validation – Elbow Method

Implementation of Clustering Technique

To process the data sets into clusters, the agglomerative clustering technique is implemented using Python. The dataset is processed to calculate the total energy consumption (KW) for 8 months and then averaged using MS-Excel analytics tools. As specified above, the Elbow index method suggested 4 clusters to group the data. The following figure (Fig 2) shows the dendrogram generated by Python programming. The X-axis shows the meter number and y-axis shows the energy consumption. This clearly shows 4 clusters in hierarchical order.

May – June 2020
ISSN: 0193-4120 Page No. 5690-5695

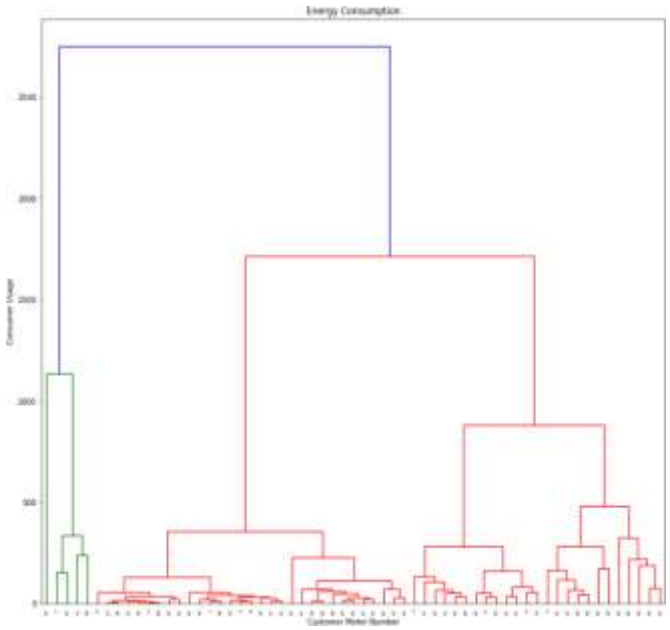

Fig 2. Dendrogram based on average power consumption in KW (8 Months)

Elements in clusters are extracted using K-Means clustering in python. Table-1 shows the elements on a particular cluster, average energy consumption for the duration of 8 months. The total consumer before cleaning of data was 69 and after processing 61. The distribution of the customers in the most suitable cluster is given in table 1 as follows.

Table 1. Details of clusters formed on the complete dataset

Cluster	No. of Consumers	Average Energy Consumption (KW)	Average Daily Consumption (KW)
C1	12	105.15	0.04
C2	31	27.40	0.01
C3	5	411.88	0.15
C4	13	191.07	0.07

Load profiles for each cluster is shown with the help of scattered chart to visualize the consumption of customers in each cluster generated. Table-1 shows the number of elements in each cluster, average consumption for 8 months, and average daily consumption in each cluster. This result helps to identify the categories of industries according to their energy load consumption. Further, we can utilize this information to estimate the peak load for each company and plan to optimize the energy during peak load hours. This result will prove to be beneficial to process the long-term energy demand prediction also. Now, the following figures (3-6) are representing the energy consumption of each customer in the respective cluster.

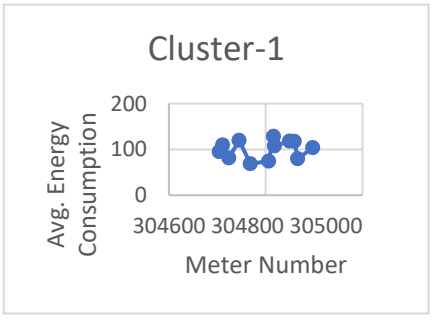

Fig 3.

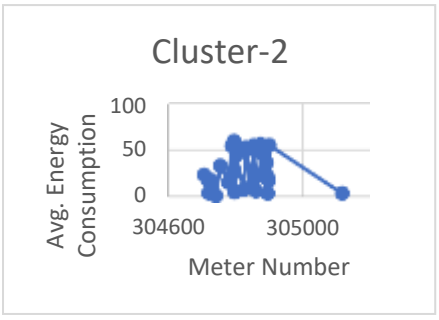

Fig 4.

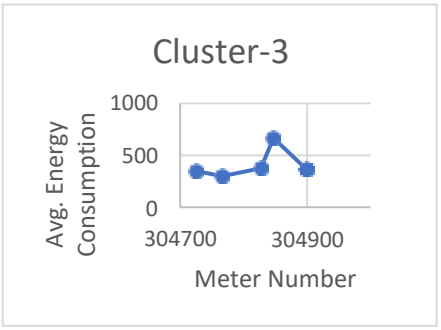

Fig 5.

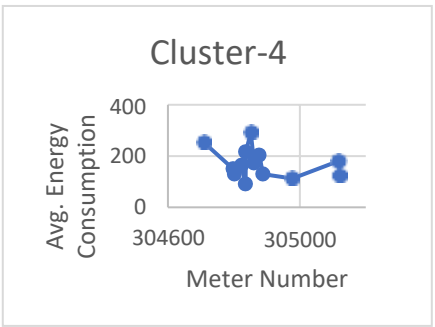

Fig 6.

Fig 3. – Fig 6. Showing distribution of customers in four clusters

May – June 2020
ISSN: 0193-4120 Page No. 5690-5695

From the above visualization, it is quite evident that clustering and comparison of customer energy consumption are very helpful in view of designing tariff structure, optimal energy planning, demand-side management, and many other planning decisions. From the above shown graphical result, it is clear that there are four categories of the company depending upon the load consumption profile. The average minimum load category is cluster-2 shown in figure-2, which is the majority rate of usage. The average maximum category is cluster-3 shown in figure-4. This cluster consists of a minimum number of industries also.

V. CONCLUSIONS & FUTURE WORK SUGGESTION

This paper demonstrates a complete step to process a dataset to calculate the average energy load consumption of the consumers. It also gives an idea about the most popular methods used for estimating the number of clusters for any data set and types of clustering techniques used for data mining. Taking into consideration, an energy consumption data set all the steps are performed using python. The processing of raw data is done using IBM SPSS software. To visualize results for a better understanding scattered graph is used. We have done this implementation based on the consumption of the customers. For further research, the timestamp data will be considered to process the consumption based on the specific time of the day. It will also be processed to differentiate the consumption of the customers based on weekdays and weekend. Since the data is from the industrial sector, we expect to have the maximum usage in weekdays. We plan to use this result for predicting the energy demand for the future and optimization of energy during peak load hour by offering better energy policies and incentives. We will be exploiting timeseries models and algorithm for future work such as ARIMA[15], LSTM[14], MLP[16] Timeseries to get accurate prediction.

REFERENCES

[1] A. M. Tureczek, and P. S. Nielsen, “Structured literature review of electricity consumption classification using smart meter data,” Energies, vol. 10, no. 5, pp. 584, 2017.

[2] J. Kwac, J. Flora, and R. Rajagopal, “Household energy consumption segmentation using hourly data,” IEEE Transactions on Smart Grid, vol. 5, no. 1, pp. 420-430, 2014.

[3] C. Chelmis, J. Kolte, and V. K. Prasanna, "Big data analytics for demand response: Clustering over space and time." pp. 2223-2232.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” ACM computing surveys (CSUR), vol. 31, no. 3, pp. 264-323, 1999.

[5] J. L. Viegas, S. M. Vieira, J. M. Sousa, R. Melicio, and V. Mendes, "Electricity demand profile prediction based on household characteristics." pp. 1-5.

[6] R. Granell, C. J. Axon, and D. C. Wallom, “Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles,” IEEE Transactions on Power Systems, vol. 30, no. 6, pp. 3217-3224, 2014.

[7] P. Bholowalia, and A. Kumar, “EBK-means: A clustering technique based on elbow method and k-means in WSN,” International Journal of Computer Applications, vol. 105, no. 9, 2014.

[8] A. M. Ferreira, C. A. Cavalcante, C. H. Fontes, and J. E. Marambio, “A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector,” International Journal of Electrical Power & Energy Systems, vol. 53, pp. 824-831, 2013.

[9] A. M. Tureczek, P. S. Nielsen, H. Madsen, and A. Brun, “Clustering district heat exchange stations using smart meter consumption data,” Energy and Buildings, vol. 182, pp. 144-158, 2019.

[10] R. Al-Otaibi, N. Jin, T. Wilcox, and P. Flach, “Feature construction and calibration for clustering daily load curves from smart-meter data,” IEEE Transactions on industrial informatics, vol. 12, no. 2, pp. 645-654, 2016.

[11] J. Kang, and J.-H. Lee, “Electricity customer clustering following experts’ principle for demand response applications,” Energies, vol. 8, no. 10, pp. 12242-12265, 2015.

[12] R. Al-Otaibi, N. Jin, T. Wilcox, and P. Flach, “Feature construction and calibration for clustering daily load curves from smart-meter data,” IEEE Transactions on industrial informatics, vol. 12, no. 2, pp. 645-654, 2016.

[13] Gouveia, João Pedro, and Júlia Seixas. "Unraveling electricity consumptionprofiles in households through clusters: Combining smart meters and door-to-door surveys". Energy and Buildings 116 (2016): 666-676.

[14] Kong, Weicong, Zhao Yang Dong, Youwei Jia, David J. Hill, Yan Xu, and Yuan Zhang."Short-term residential load forecasting based on LSTM

May – June 2020
ISSN: 0193-4120 Page No. 5690-5695

recurrent neural network." IEEE Transactions on Smart Grid 10, no.1 (2017): 841-851.

[15] Contreras, Javier, Rosario Espinola, Francisco J. Nogales, and Antonio J. Conejo. "ARIMA models to predictnext-day electricity prices." IEEE transactionson power systems 18, no. 3 (2003): 1014-1020.

[16] Huang, Dongliang, Hamidreza Zareipour, William D. Rosehart, and NimaAmjady. "Data mining forelectricity price classification and the application todemand-side management." IEEE Transactions on Smart Grid 3, no. 2 (2012): 808-817.