JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BIG DATA AGGREGATION AND ANALYSIS IN DEMAND-SIDE MANAGEMENT IN CONTEMPORARY POWER SYSTEM

Daler, Er Gurpreet Kaur
M.Tech Student, Guru Kashi University, Talwandi Sabo
Assistant Professor, Guru Kashi University, Talwandi Sabo
dalerallied@gmail.com
ergurpreet88@ gmail.com

Abstract- In this data-driven world, there are various sectors of industries that are being benefited from data analytical business. Similarly, the power industry can also be benefited from the data collected from various customers. The most difficult part is residential sector as the number of load profiles are very large in number. Analyzing these huge numbers of loads is next to impossible task, so to reduce the number of profiles we should remove the redundant profiles. It can be done by replacing similar load profiles with a single load profile. But the challenge here is to identify similar load profiles. For that in this paper, the k-means clustering procedure is been used. In this paper, an example of the dataset of 200 households is taken, whose 10-mint load consumption is given. This dataset is clustered using k-means clustering and 4 clusters are extracted. The best number of clusters are decided using the elbow method.

Keywords:loadprofiles,clusters,centroids

I INTRODUCTION

Data aggregation is the operation of collecting data and representing it in a condensed format. The data may be collected from many data sources with the goal of combining these value sources into a summary for data examination. Cluster data refers to numerical or non-numerical facts that are gathered from multiple origins and/or on multiple measures, variables, or single and compiled into data summaries reports, typically for the motive of public reporting or analytical examine i.e., examining trends, producing comparisons, or divulge detail and insights that would not be observable when data elements are viewed in isolation. For example, facts about whether single students graduated from high college can be a cluster that is, compiled and summarized into an individual graduation rate for a graduating college, and annual college graduation rates can then be clustered into graduation rates for districts, states.

For purposes of this study, the term “big data” is defined as large volumes of unstructured and heterogeneous data units that are complex about conventional techniques.

Big Data and Analytics in Demand Response focuses on interfaces with data sources, the data sources character, and data analytics for Demand Response programs, with the goal of updating utility operations and grid reliability. We reviewed matter grid smart meter and PMU values to recognize the feature of the utility and the value of large-data analytics for the issue grid.

The identification of season types can be done through clustering methods. Numerous clustering methods were grown which can be

effectively used to cluster the load profiles. K means clustering, the most widely known and the most effective clustering methods, belongs to the class of centroid-based totally clustering. In this method, every cluster is denoted by the central vector, which is not an important part of the cluster. The K means method takes K as the excitation to take part the data objects into K groups such that the intra cluster same is high, while the inter-cluster same is low. K needs to be particular in advance, it is appraised one of the largest weaknesses of the clustering method. Moreover, it is detrimental in terms of collecting vacant clusters. Fuzzy C-means algorithm is another method belonging to the same class, where each data object has a degree of membership to every cluster, unlike K-means clustering where each data objectives into one cluster only. Hierarchic clustering is a class of clustering methods that do not provide a single partitioning of the data set rather create a hierarchy with a tree-like structure based on distance and similarity between them. The users still need to select an appropriate number of clusters. These techniques are not very robust to outliers which can insert as extra clusters. The arrangement of the clusters is illustrated by tree diagrams called Dendron grams. This algorithm works through grouping the data one in accordance to the nearest displacement measure of all the pair-wise distance between the values points. The bottom-up approach of hierarchical clustering is called agglomerative clustering, while the top-down approach is called divisive clustering.

1.2 Types of Data Analytics

1.2.1 Descriptive analytics

Descriptive analytics mine and cluster values to give awareness into the past (what happened).

1.2.2 Predictive analytics

Predictive analytics employ a variety of statistical, modelling, data-mining, and machine-learning techniques to learn recent and historical data as a basis for forecasting the future.

1.2.3 Prescriptive analytics Prescriptive analytics use optimization and simulation procedures to suggest possible results and recommend the fine path of action for any pre-unique final results.

We present to provide insights into advanced situational and prescriptive analytics as well as technologies for the pre-emptive resolution of field challenges. Situational analytics combine descriptive, predictive, and prescriptive analytics to understand real-time intelligence about the condition of the grid

II.PROBEM DEFINING

There is a large amount of data with high volume, velocity and data have been generated every second. The number of households, having load profiles is also increasing day by day. Therefore, the number of load profiles that have to be analysed is increasing. And computing such a number large number of profiles differently is a very complex task. So, similar profiles should be identified and replaced with one class representative. So, identifying similar types of profiles and groups is known as clustering. There is a need to identify and implement the techniques of clustering specific to load profiles of electrical consumers.

2.2 Objectives

Based on the above discussion the objectives of this paper are:

To study and analyse big data analysis in demand-side management of a power system.

To study the analyse different techniques to gather and aggregate big data from smart meters.

To study and analyse different techniques for reducing dimensions and computational burden.

III Research Methodology

There are various techniques available in the literature for clustering. To find the similarity between different profiles the distance between them is measured. There are various types of distances following:

a) Vector to vector distance

$$d(x,y) = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(y_h - x_h)^2} \quad (1)$$

A load profile of the day can be a vector x and another load profile vector y.

b) Vector to set distance: Computed by using the distances between the vector y and each of the M members of set X.

$$d(y,X) = \sqrt{\frac{1}{M}\sum_{x\epsilon X} d^2(y,x)} \quad (2)$$

c) Average set to set distance, given by the mean distance between all pairs of members x_q of the set X (with Q members) and y_j of the set Y (with j members)

$$d(X,Y) = \frac{1}{QJ}\sum_{q=1}^{Q}\sum_{j=1}^{J} d(x_q, y_j) \quad (3)$$

d) Steps for Clustering

e) The data of 200 households are taken with the frequency of 10 minutes of load consumption of 1 year. The panda's data frame is created in Jupiter notebook is taken.

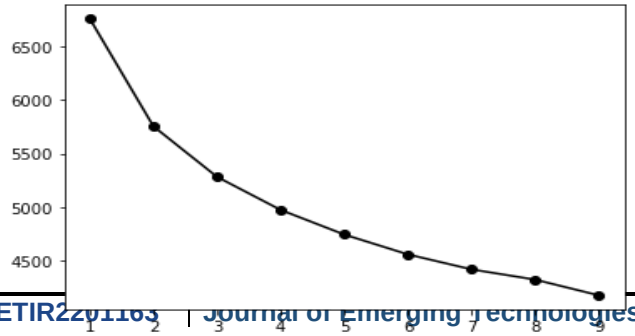
Time	House 1	House 2	….	…..	House 199	House 200
2010/01/01 00:00:00	3271	2776	….	…..	13013	2827
2010/01/01 01:00:00	2548	4093	…..	…..	14443	3164
……	…..	…..	…..	…..	….	….
2010/12/31 23:40:00	6217	6579	…..	…..	5786	4756
2010/12/31 23:50:00	6943	5275	…..	…..	8175	6968

Then the data is resembled into hourly sampled data. And the resembled sum () method of pandas in python as shown in the table above. As the aim of this thesis is to cluster the data into different clusters to ease the analysis. For that purpose, the data frame is needed to be turned into a time-indexed data frame as shown in the table.

Then K-means clustering is performed to cluster similar types of load profiles into different groups.

Firstly, the random number of centroids is declared. Then the clusters are assigned as the distance between different centroids and the load point using the following equation.

It's a minimization hassle of two parts. We first limit J w.r.t. w_{ik} and treat μ_k fixed. Then we minimize J w.r.t. μ_k and deal with w_{ik} constant. Technically talking, we differentiate J w.r.t. w_{ik} first and update cluster assignments (E-step). Then we differentiate J w.r.t. μ_k and recomputed the centroids after the cluster assignments from the previous step (M-step).



After assigning all the load points to different clusters and the centroids are calculated again.

Using those centroids then the distance from each load point is calculated and again new clusters are performed accordingly.

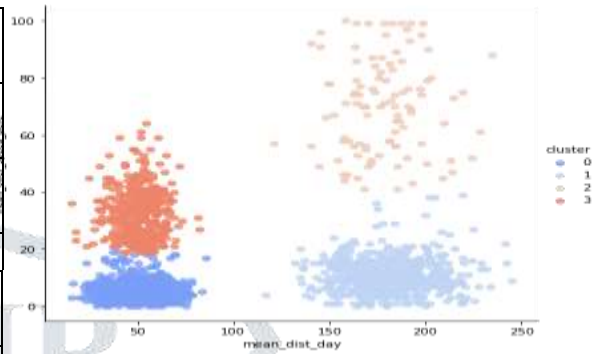


Figure 3:Elbow method to find no of cluster

- Then the sum of square error is to be calculated against each cluster
- Again, the Euclidean distance between each point and the new cluster centroids should be calculated and these steps are doing again and again until the sum of square error becomes almost constant or large number of iterations has been reached

IV. Result &Discussion

The clustering has been performed on the dataset of 200 households, where the energy consumption for every 10 minutes is given. If the raw data of the consumers are drawn on a single graph, then it will look like a mess as shown in figure 4.

As it can be noted that, analyzing this is a cumbersome task and even impossible to some extend to analysis the profiles individually. So to analyze the data properly it has been clustered using k means clustering.

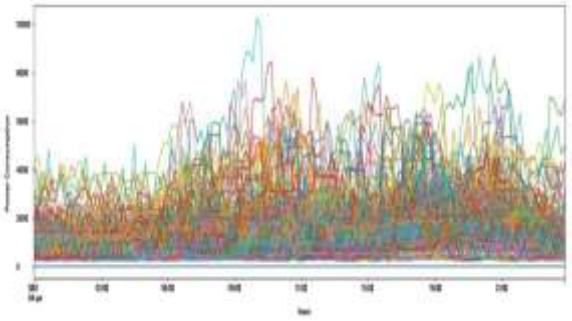


Figure 5: Error v/s number of clusters graph

And to find the optimum number of clusters elbow method is used. In which it is started with 1cluster and check for the error and draw the graph. So, from the curve, it can be noted that an optimum number of clusters can be 4 or 5. Using k means clustering in this thesis 4 clusters have been chosen and are shown in figure 4.

Here, various features can be extracted from figure 4, which was impossible in figure 3. The features are as follow:

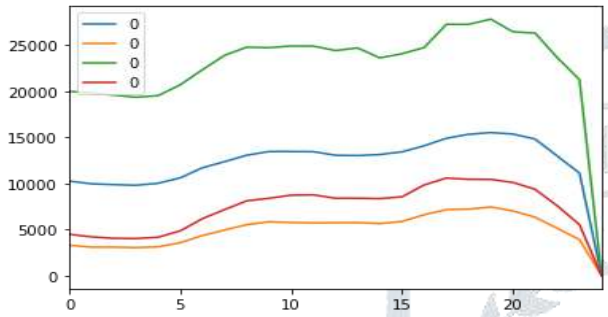
Features of the clusters

Most revenue generating Households

Maximum density cluster

Minimum density cluster

Time of peak and valley of a cluster



The above shown the four load profile for all the 200 households. Any analysis can be conducted on the class representative profile on the behalf of all the households of the particular cluster. So, it could be computational free and cost of hardware could also be reduced.

Table3 : Four Cluster's Class Representative

Hours	1st Cluster Class	2nd Cluster Class	3rd Cluster Class	4th Cluster Class
0	11506	4794	21369	5890
5	11917	4931	21780	6301
10	14657	7260	25890	10136
15	14520	7266	24931	9726
20	16712	8493	28219	11643

V.Conclusion & Future Scope

In this paper, a dataset of 200 households is taken in which a load of every 10 minutes is given. The datasets have been plotted on the graph and observed that it was hard to analyze and take any decision on that behalf. So, it is extremely important to cluster the load profiles into groups of similar consumption patterns so that data analysis could be possible. So, the first challenge to be faced during clustering is to find out how many clusters are to take, which can be decided by many techniques provided in the literature. However, in this thesis, the elbow method has been chosen to find out the maximum number of clusters. So, in this case, 4 clusters are chosen and the clustering technique is k-means clustering. The clear and distinct 4 clusters have been formed and their class representative is chosen. The whole group of households is represented with a single load profile i.e. their class representative. This has resulted in the reduction of computation to 50 times. As instead of 200 load profiles only 4 profiles are to be analyzed now.

The future scope covers a very wide area and a lot of new researchers have been attracted to this field. The clustering can be done on various bases. The basis could be trend based, variability based and so on. Each basis has its own application

VI REFERENCES

[1] J.Luo (VSTLF), "Real-time anomaly detection for very short-term load forecasting," SGEPRI, 2018.

[2] M. Yue, "Anomaly Detection Based on Long Short-Term Memory Neural Network," IEEE, 2017.

[3] A. lazaris, "An LSTM Framework For Modeling Network Traffic," IEEE, 2019.

[4] v. k. prasanna, "A short-term building cooling load prediction method," 2019.

[5] P. Zhao, "Advanced correlation-based anomaly detection method for predictive maintenance," IEEE, 2017.