Data Mining Applied to Customer Categorization Based on Load Profiling

Shreyas Karnick[1], Dr. Shivakumara Aradhya[2]
[1]*Mtech-Power System Engineering Acharya Institute of Technology,Student*
[2]*Dept. of Electrical and Electronics Engineering Acharya Institute of Technology, Professor*

Email:shreyas.mtps.14@acharya.ac.in[1],shivakumararadhyars@acharya.ac.in[2]

Abstract-Load Profiling, a method where load consumption patterns of different electricity consumers are identified using the daily/monthly load curves is used in Distribution System planning activities like peak load management and time of use tariff. The load profiling identifies various customers with similar load patters and groups them into clusters. This method of customer categorization helps the utilities by eliminating the tiresome task to collect load information data of individual customers' continuously over time and analyze each of them to be applied for planning activities. Instead with defined categories of customers exhibiting similar consumption patterns, the utilities can then efficiently plan the distribution of power without any difficulty. Efficient distribution system load data processing and analysis arises as one of the main concerns in Load Profiling. Data mining a process to derive interesting and intelligent knowledge from large databases and hence analyze the data to obtain patterns of similarity or dissimilarity can be employed for data processing and analysis. Thus this paper aims to utilize the Tadpole method in Dynamic Time Warping of Data Mining implemented in the R-tool to effectively process and analyze load data and hence obtain different categories of customers. This will in-turn aid in efficient distribution system planning with regards to demand side management programs.

Index Terms-Load Profiling, Customer Categorization, Data Mining, Tadpole Dynamic Time Warping.

1. INTRODUCTION

The unprecedented increase in demand for electricity rises many concerns over efficiently supplying reliable power to large number of consumers. Setting up of new generating stations may be technically and economically infeasible. The best alternative to do away with installing new generating stations is Demand Side Management (DSM), which employs changing the pattern of power consumption at the end-user side rather than the generation side. In the Smart Grid environment DSM activities mainly revolve around the Load Data collected from meters. The Smart Grid has led to a number of revelations in the Electrical Sector wherein the Smart Meters/Automated Meter Readings pave way to two way communication between the utilities and the consumers in addition to short time load data logging. Thus there is necessity to make best use of this data to be applied in DSM and Demand Response programs such as peak load management, time of energy usage tariff, and incentives for use of energy during off peak times. One of the key aspects of DSM include categorizing customers on basis of similarity or dissimilarity of their load consumption patterns. This categorization of customers depends on the load profiles obtained from the Smart meter logged load data. Data mining techniques can be employed in order to recognize patterns of similarity or dissimilarity in the load data of various customers.

One of the most widely used pattern recognition techniques-K-means along with two modified forms of the K-means have been used for optimal grouping of demand patterns [1]. A number of other algorithms such as Hierarchical clustering [2], Fuzzy k-means[3], Modified Follow-the-leader [4], Adaptive vector quantization [5], Self-organized maps [6] have been used to cluster the load data. These aforementioned algorithms are limited in the sense that they cannot process high-dimensional, complex and dynamic time series data.

This paper thus utilizes the Time Series Algorithms implemented in R-Tool to efficiently categorize the electrical load data, which basically vary with time, to categorize various customers into groups depending on the similarity/dissimilarity of their load consumption patterns.

2. CUSTOMER CATEGORIZATION BASED ON LOAD PROFILING

In Power Systems, the load or consumption of energy varies with time, and the generation, transmission and distribution companies must respond to the customers demand with minimum time lag. Therefore, the time series load information becomes a prime necessity for a number of power system activities such as Tariff Design, System Planning, Load Forecasting, Peak Load Management and many such Demand Side Management activities.

A Load Profile is basically a graph of variation in the electrical load versus time that provides a visual depiction of the load information. The load profiles vary in accordance with the type of customers like residential, industrial, commercial etc., along with other important factors such as time, climate, customer composition and supplynetwork structure.

In the presence of an extensive distribution system, collecting continuous load information of individual customers over time comes across as impractical. Also certain categorization of the load information based on voltage levels, economic activities etc., would fail to provide the daily, weekly and seasonal variation of demand patterns.

Thus, in order to facilitate efficient analysis and the applications of load profiling, customers exhibiting similar consumption or demand patterns are grouped into differentcategories. With well-defined categories, it becomes an easy task for the utilities to implement the demand side management and demand response mechanisms and planning activities for the categories of similar loads rather than individual loads which might prove a tiresome task that requires more resources and also is time consuming.

3. DATA MINING APPLIED TO CUSTOMER CATEGORIZATION

The recent drift towards the Smart Grid has paved way to many novel technologies of which the Smart Metering and Automated Meter Reading (AMR) are few of the many to be mentioned in the scenario of customer categorization based on load profiling. The Smart Meters/AMR provides the utilities with continuous, accurate and up-to-date data ofconsumption of individual customers that can be effectively used for load profiling followed by categorization of the customers based on similar consumption patterns. The data thus obtained from these meters combine to form enormous amount of data that face hindrance in analysis and processing. A solution to this problem can be found by applying the Data Mining concepts to effectively analyze and process the load information data.

Data Mining is an inter-disciplinary subfield of computer science that involves analytical processing of large amounts of data (big-data) in search of consistent patterns among the existing variables and hence aid in prediction and analyses of future variables. It basically extracts the intelligent and interesting knowledge from large datasets and further transforms into understandable manner for future use. A number of techniques and algorithms exist in Data Mining that aid effective extraction of knowledge and also specific patterns are recognized in the existing enormous data.

With the load data being time series in nature, the Time Series Clustering algorithms used in Data Mining are used to obtain various groups of customers with similar load consumption patterns. The demand response mechanisms are applied to different classes of customers grouped used Time series clustering algorithms. This paper mainly concentrates on the Dynamic Time Warping[7] that initially used for the purpose of voice recognition has been employed to cluster load data from sixteen individual meters based on the similarity/dissimilarity depicted by their respective load profiles.

4. DYNAMIC TIME WARPING

Dynamic time warping (DTW) is a well-known technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions. DTW recovers optimal alignments between sample points in the two time series. The alignment is optimal in the sense that it minimizes a cumulative distance measure consisting of “local” distances between aligned samples. This procedure is called time warping due to the fact that it warps the time axes of the two time series in such a way that corresponding samples appear at the same location on a common time axis.

Any distance (Euclidean, Manhattan, etc.) which aligns the *i*-th point on one time series with the *i*-th point on the other will produce a poor similarity score as shown in Fig-1

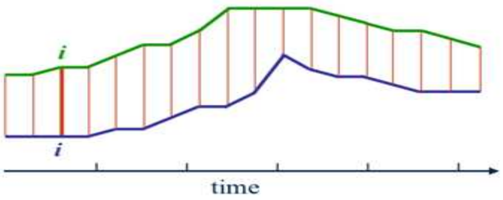

Fig-1: Euclidean/Manhattan Distance Measure

A non-linear (elastic) alignment such as DTW produces a more intuitive similarity measure, allowing similar shapes to match even if they are out of phase in the time axis as shown in Fig-2.

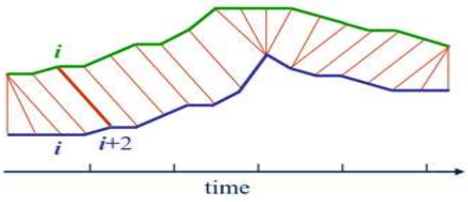

Fig-2: Dynamic Time Warping Distance Measure

Originally, DTW has been used to compare different speech patterns in automatic speech recognition. In fields such as data mining and information retrieval, DTW has been successfully applied to automatically cope with time deformations and different speeds associated with time-dependent data. Thus, with load data being time series in nature, DTW is appropriate to be applied to categorize customers based on their load profiles. The TADPole clustering algorithm [8] is an improvisation of the DTW algorithm that employs pruning strategy to reduce the number of calculations that exists in the conventional DTW algorithm. R-Tool has been used to implement the TADPole, algorithm to cluster the load data obtained from individual meters into groups depicting patterns of similarity/dissimilarity.

5. METHODOLOGY INVOLVED IN TIME SERIES CLUSTERING OF LOAD DATA

- Load Data collected from sixteen individual meters and at 30-minute interval making up 48 blocks per day for a period of 3-months that have been recorded using the Smart meters are used for the purpose of clustering of load data.
- This collected load data stored in csv (comma separated variables) is read using the R-Tool, a statistical tool that is well suited for data mining applications.
- Once the data is read, it is processed such that the rows depict the individual meters while the columns depict the time series data recorded at an interval of thirty minutes for a period of three months for the individual meters.
- The control function is then initialized with a windowing constraint that is necessary to warp an observation falling within the window. This is basically used to speed up the DTW calculation.
- The TADPole clustering algorithm where in the cluster centers are always elements of the data and is deterministic on the value of cut-off distance that is indicative that any distance measure less than this distance can be considered for grouping into the same cluster.
- The cluster centers are obtained and the load profiles of individual meters against the cluster groups they belong are plotted.

6. RESULTS AND DISCUSSIONS

The algorithm is converted accordingly into a code in the R-Tool using the “dtwclust” package. Once the program is run, the cluster centers that act as template data are obtained which are elements of the data only and the rest of the data are compared with the cluster centers on the basis of the cut off distance. Any distance below the cut off distance is used as a deterministic quantity to obtain the clusters. The results thus obtained involve efficient grouping of load data into clusters wherein a cluster center in initially obtained and the remaining load data are assigned to respective clusters.The DTW clustering process in case of electrical load data occurs by checking whether a certain event has occurred earlier in the cluster centers and if the pattern matches with an earlier event in the given windowing constraint, thus speeding up the clustering process.

The following results depict individual load profiles of meters that belong to a particular cluster followed by the plot of the cluster they belong. The number of clusters were initialized to be 10 and the groupings are as follows:

- Cluster-1: The cluster-1 comprises of load profiles from three meter reading from three individual meters which have similar consumption patterns though shifted in timeand are shown in Fig-3 while Fig-4 shows the entire Cluster-1 plot.

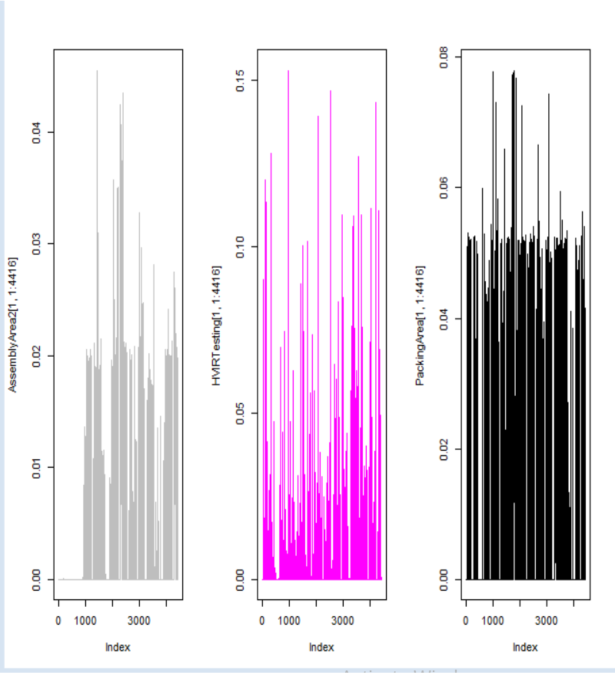

Fig-3:Individual Load Profiles of Load Data Belonging to Cluster-1

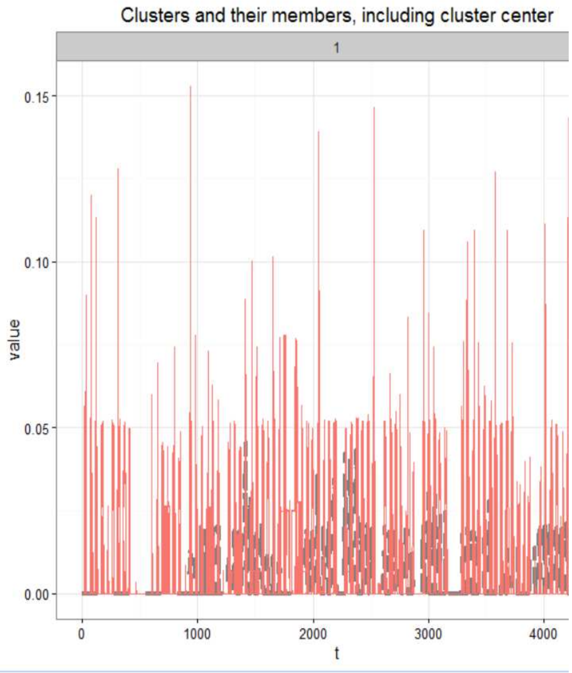

Fig-4:Cluster-1

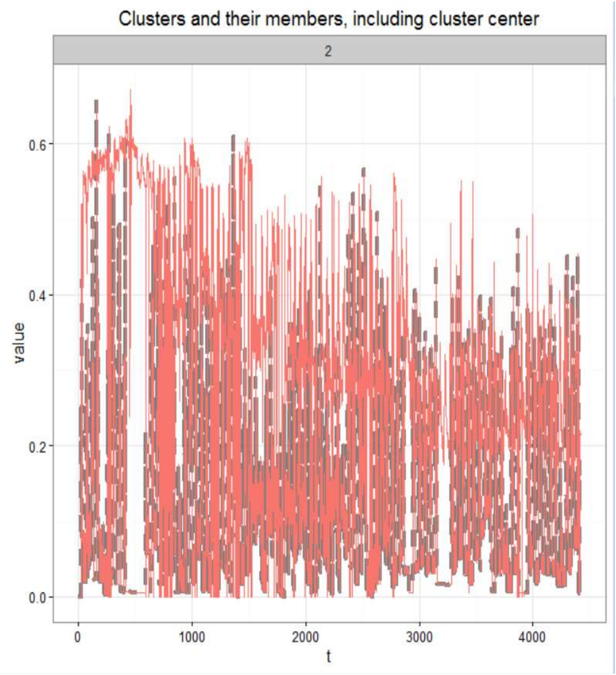

Fig-4:Cluster-2

- Cluster-2: The cluster-2 comprises of load profiles from two meter reading from two individual meters which have similar consumption patterns though shifted in timeand are shown in Fig-5 while Fig-6 depicts the entire Cluster-2 plot.

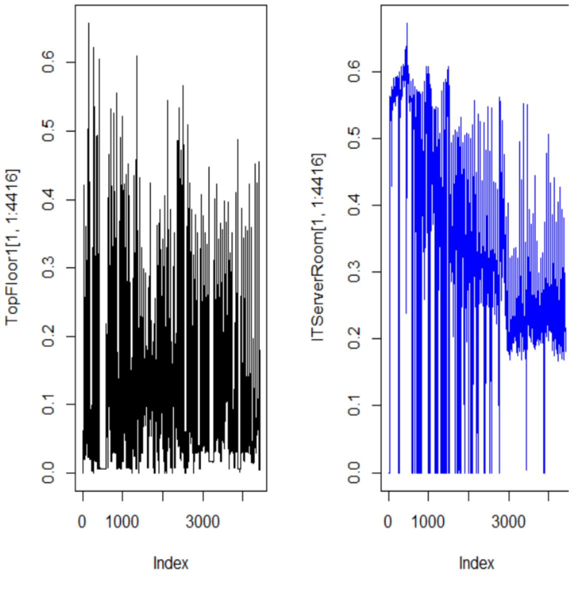

Fig-5:Individual Load Profiles of Load Data Belonging to Cluster-2

- Cluster-6: The cluster-6 comprises of load profiles from two meter reading from two individual meters which have similar consumption patterns though shifted in timeand are shown in Fig-7 while Fig-8 illustrates the entire Cluster-6 plot

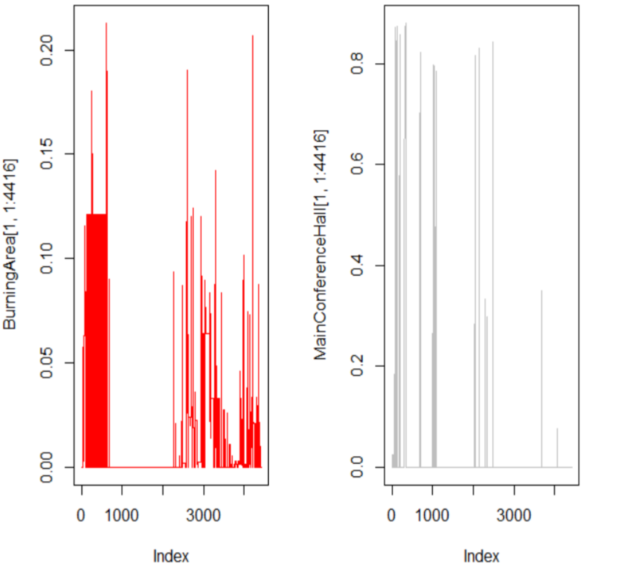

Fig-7:Individual Load Profiles of Load Data Belonging to Cluster-6

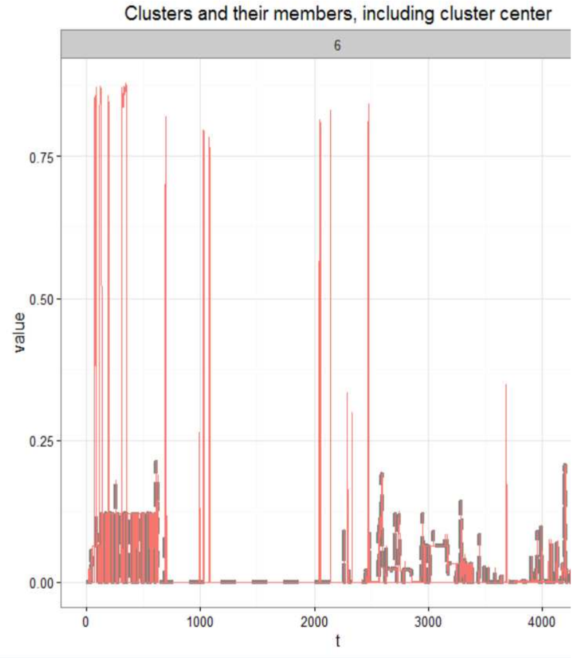

Fig-8:Cluster-6

- Cluster-8: The cluster-8 comprises of load profiles from two meter reading from two individual meters which have similar consumption patterns though shifted in timeand are shown in Fig-9 while Fig-10 showss the entire Cluster-6 plot

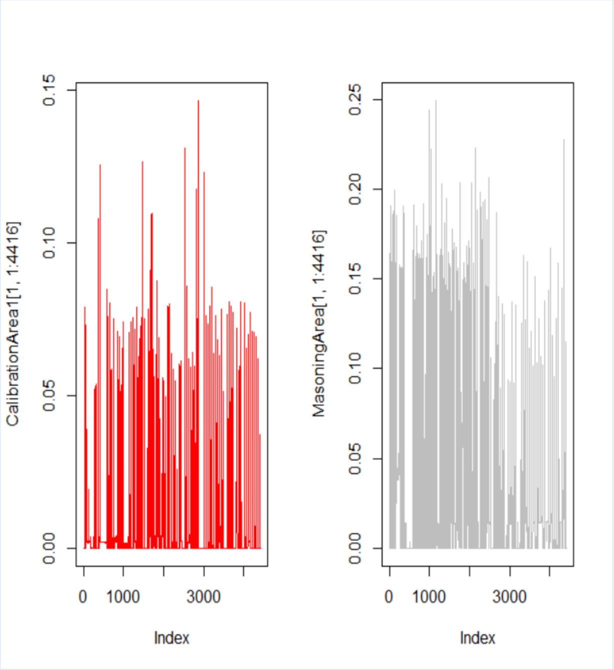

Fig-9:Individual Load Profiles of Load Data Belonging to Cluster-8

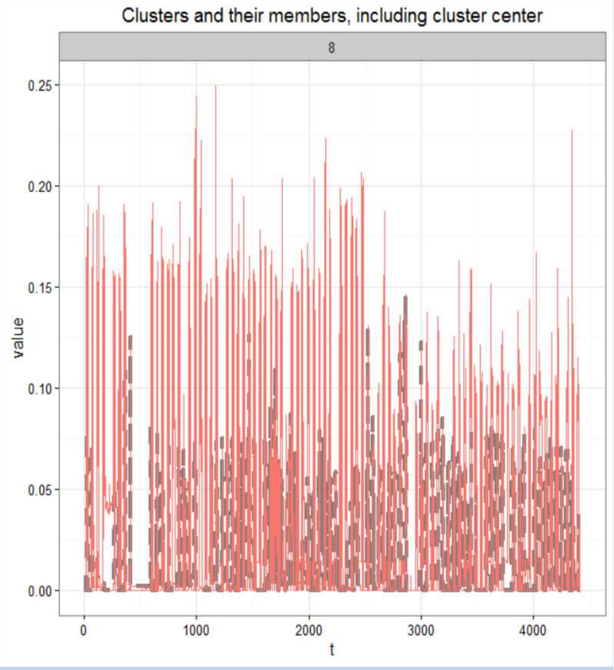

Fig-10:Cluster-8

The rest of the meters do not fall under any grouping and depict unique load consumption patterns. They are by themselves a cluster, these include Clusters-3, 4, 5, 7,9,10.

7. CONCLUSION AND FUTURE SCOPE

The clusters of load data thus obtained using the TADPole algorithm implemented in R-Tool produce satisfactory grouping of individual load data from smart meters. There is efficient processing and analysis of huge amount of load data obtained from Smart Meters at an interval of 30-minutes for a period of 3-months using appropriate Data Mining techniques.

The clusters thus obtained can be used in two ways, one to apply appropriate Demand Side Management Techniques to a given group of customers and hence doing away with the need to design individual DSM programs to individual customers. Another use of clustering the load data involves, with the identification of cluster centers, it becomes a simple task to assign a new customer to a respective group easily and access their consumption patterns for future leading to effective demand side and supply side planning activities in this world of fast growing demand.

REFERENCES

[1] Ioannis P. Panapakidis, Minas C. Alexiadis, Grigoris K. Papagiannis, “Enhancing the clustering process in the category model load profiling” , IET Generation Transmission and Distribution, Vol. 9, Issue 7, Pages:655-665, April-2015.

[2] M. R. Anderberg, *Cluster Analysis for Applications*. New York: Academic,1973.

[3]J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*.New York: Plenum, 1981.

[4] Y.-H. Pao and D. J. Sobajic, “Combined use of unsupervised and supervisedlearning for dynamic security assessment,” *IEEE Trans. PowerSyst.*, vol. 7, no. 2, pp. 878–884, May 1992.

[5] G. J. Tsekouras, F.D. Kanellos, V.T. Kontargyri, I.S. Karanasiou, A.D. Salis, N. E. Mastorakis, “A New Classification Pattern Recognition Methodology for Power System Typical Load Profiles”, WSEAS TRANSACTIONS on CIRCUITS and SYSTEMS, Vol.7, Issue 12, December 2008.

[6] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed.Berlin, Germany: Springer-Verlag, 1989.

[7] Hiroaki Sakoe and Seibi Chiba, “Dynamic Programming Algorithm Optimization for Spoken Word Recognition”, IEEE Trans. On Acousitcs,Speech and Signal Processing, Vol. ASSP-26, No. 1, February-1978.

[8] Begum N, Ulanova L, Wang J and Keogh E (2015). “Accelerating Dynamic Time Warping Clusteringwith a Novel Admissible Pruning Strategy.” In Conference on Knowledge Discovery andData Mining, series KDD ’15. ISBN 978-1-4503-3664-2/15/08.