International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 5, September- October 2017 ISSN 2278-6856

Analyzing Electricity Consumption Behavior Of Households Using Big Data Analytics

B. Mohan Krishna[1], K. Bala Chowdappa[2]

[1]PG Scholar, Dept of CSE, G. Pulla Reddy Engineering College (Autonomous), Kurnool, AP, India,

[2]Assistant Professor, Dept of CSE, G. Pulla Reddy Engineering College (Autonomous), Kurnool, AP, India,

Abstract

In a competitive retail market is looking to analyze their customer each and every power usagedata into a smart meters that provide lots of opportunity for power board to gain more knowledge on customer's power usage in a large volumes of smart grids meters data. There are a lot of analytical solutions to provide the power usage of households but these solutions are not provided that power of utilization there is a lot of need for analyzing power usage of households data analytics, which includes the relation with type of power usage(laundry, kitchen, heaters and AC etc).The system proposes in finding the power usage by implementing levels of clustering algorithms that fits the best models to locate the households of power usage in varioussituation and areas. The proposed methods find the various power expenditure behaviors like day type, time and season in a large data set.

Keywords: Big Data, Power usage, Analyzing household data, Smart meters, Data Analytics, Day type, Season and Sub meters.

1. INTRODUCTION

NATIONS around the world are having the aggressive targets to restrict the monopolistic power system towards liberalized markets especially on the demand side. In a real-time power supply market, Electricity Load Serve Entities (ELSEs) are being developed in great number.The competitiveness of ELSEs [1] can be enhanced by understanding the patterns and realizing personalized power usage managements.In the meantime, thehigher smart meters (smart grids) method is popularity increasing in a worldwide both weight serve higher smart meters infrastructure and are capability day type hours to minutes at high frequency. Conservation power usage data is large volumes expose information of customers that can potentially used by LSEs to manage their generation and demand resources efficiently provide their personalized service.

Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
16/12/2006	17:24:00	4.216	0.418	234.84	18.4	0	1	17
16/12/2006	17:25:00	5.36	0.436	233.63	23	0	1	16
16/12/2006	17:26:00	5.374	0.498	233.29	23	0	2	17
16/12/2006	17:27:00	5.388	0.502	233.74	23	0	1	17
16/12/2006	17:28:00	3.666	0.528	235.68	15.8	0	1	17
16/12/2006	17:29:00	3.52	0.522	235.02	15	0	2	17
16/12/2006	17:30:00	3.702	0.52	235.09	15.8	0	1	17
16/12/2006	17:31:00	3.7	0.52	235.22	15.8	0	1	17
16/12/2006	17:32:00	3.668	0.51	233.99	15.8	0	1	17
16/12/2006	17:33:00	3.662	0.51	233.86	15.8	0	2	16
16/12/2006	17:34:00	4.448	0.498	232.86	19.6	0	1	17
16/12/2006	17:35:00	5.412	0.47	232.78	23.2	0	1	17
16/12/2006	17:36:00	5.224	0.478	232.99	22.4	0	1	16
16/12/2006	17:37:00	5.268	0.398	232.91	22.6	0	2	17
16/12/2006	17:38:00	4.054	0.422	235.24	17.6	0	1	17

Fig 1 Daily power usageweight profiles of various smart meters

The power usage of customers more than a specific period are given by load profiling ,e.g., one day, summer, and winter can help ELSEs to understand how the power is actually used for different customers and obtain the customers' load patterns. The weight profiling data is clusters are classified into two categories direct and indirect clustering [2] direct cluster methods are applied to load data. The wide spread and high-frequency collection of stylish meters encounters the challenges for data storage, communication and analysis. So, the indirect clusters are providing methods to decrease the size of the load data before clustering and dimensions.

The easy to get system are passed out on the power usage of different time delays. On the each day the powerutilization patterns may vary even for same customer in a smart meters. To expose the actual power usageshowof the customers the collected load patterns are not enough to analyze the actions. The other brave that encounters is the handling data collected on every day of high frequency and dimensionality.

The proposed system implements the time-based MarKov model[3] for handling dynamics of customer'spower usage and transferring of the daily load profiling data customers into a state transaction matrixes. The large information of the customers can be handled by the clustering fast search and find density peaks technique[4]is integrated with the divide-and-conquer approach,to makedata processingmore efficient.

2. BASIC METHODOLOGY

The proposed methods are dynamic discovery of the power usage can be separated into six stages. The first stage conducts few load data scheduling, including data hygienic and weight curve normalization. In the second stage SAX technique is used to decrease the weight profiling and dimensionality. In third stage by using the Markov model the each individual customer power energyutilization.In the fourth stage the expanse matrix is dissimilarityamong any two Markov models are extract calculatedby the K-L distance technique. The fifth stage performs a modified clustering fast search and find density peaks algorithm thatdetermine usage of powerutilization. Finally, in the sixth stage the analysis of the households electrical data targeting are obtain as the results.

Data Normalization

The normalization process transform the utilization data of arbitrary value $x = \{X_1, X_2, \ldots X_H\}$ to the variety of [0, 1]. The normalization is used to decrease the impact of anomalous data and effect of regular changes in the maximum values.

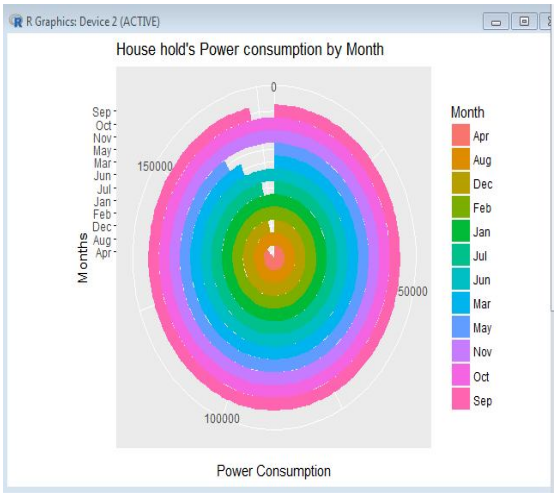

Fig 2 Power usage per year in a different months

Simple API for XML load curves

Simple API for XML (API) technique is a time series data with lower bounding of the Euclidean distance [5] and dimensional reduction. The discreteness into symbolic strings was done by SAX in two steps: symbolizing the PAA version into a discrete string of numeric time series weight data transforming into a piecewise aggregate approximation (PAA). The basic idea of PAA is instinctive and simple, replacing the amplitude values falling in the similar time intermission with their mean values.

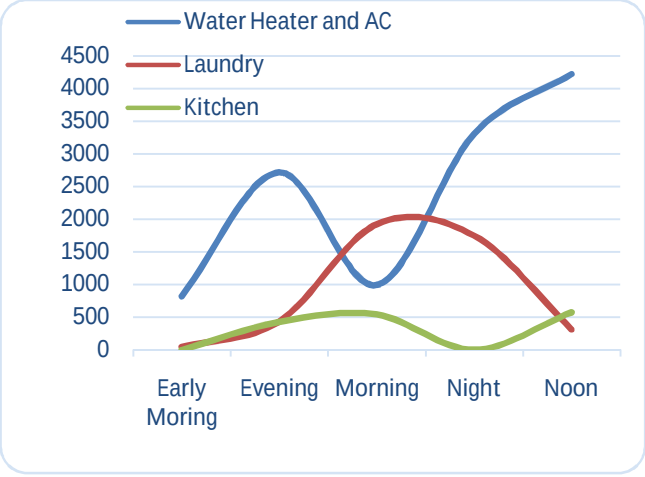

Fig 3 Customersof power usage data in various day type over one week.

The time axis is divide into five periods each day. These data to represented as "**Early morning, Evening, morning, night and noon**", with five symbols and a total of 35 day type periods. The time sphere is separated into regular period and inside each reserve, the common of the amplitude values is calculated.

Time-based MarKov model

The full usage of the customers' near and past states to be done predict the stages of power usage for the each customer. The MarKov property or Morkov chain is described as the extraction of the future consumption state from the current state.

Distance calculation

The dissimilarity among the two probabilistic distributions can be quantified effectively using the Kullback-Liebler(K-L) [6] distance technique. The discrimination connecting two Markov model with the state transition matrices **Pti**and **Ptj**, the K-L distance is definedas

$$KLD\left(\mathrm{P_i^t}, \mathrm{P_j^t}\right) = \frac{1}{N}\sum_{m=1}^{N}\sum_{n=1}^{N} p^t_{imn} \log \frac{p^t_{imn}}{p^t_{jmn}}$$

a. Clustering Fast Search and Find Density Peaks (CFSFDP) Algorithm

The algorithm allows detecting of non round clusters with diverse densities to decrease the usage complexity for huge data sets and hard threshold in employment tocompute the local density.

$$\rho_i = \sum_{j=1}^{N} \chi\left(D_{ij} - d_c\right)$$

3. PROPOSED SYSTEM

A. Framework

A divide-and-conquer framework for circulatedcluster, where **Li** denote the original data on the *i*thspread local site; **Mi** denotes the delegate objects preferred from the *i*th distributed local site; and **R** denotes the global cluster results. Each aimcorrespond to a customer described by transition probability matrixes. The system consists of three ways:

Step 1: The Simple API for XML (SAX) and time-based Markov model for individual customers are handle separately. Separate the big data set into k parts, each marked as **Li**. Note that the data on one distributed site can be further partitioned to create the size of the data sets on each site more even.

Step 2: An adaptive k-means method is performed for each individual part to acquire a certain number of cluster centers. Each cluster center can represent all the objects belonging to this cluster with a small error. All these cluster centers of **Li** are selected as the representative objects **Mi**, which are defined as a local model.

Step 3: A Clustering Fast Search and Find Density Peaks (CFSFDP) technique is applied to all the representative objects (local models) are centralized and gather to

organize them into several groups **R,**defined as a global model. Then final clustering result, the cluster label of each local site would be updated.

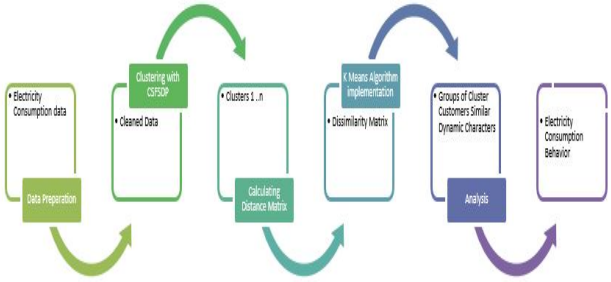

Fig 4 Analyzing power usage of households using big data analytics.

4. IMPLEMENTATION

A. Description of the Data Set

The data used in this research contains 2,075,259 (@ Million) records measurements gathered between December 2006 and November 2010 (47 months). The measurements are gathered from households ho are using different meters in deferent place. Beloware the descriptions of the meters used in this research.

Electricalsmart meters :

1) sub_metering_1: An power usage water-heater and an air-conditioner.
2) sub_metering_2: kitchen and dishwasher and a microwave.
3) sub_metering_3:laundry room, washing apparatus, a plummet drier, a refrigerator and a light.

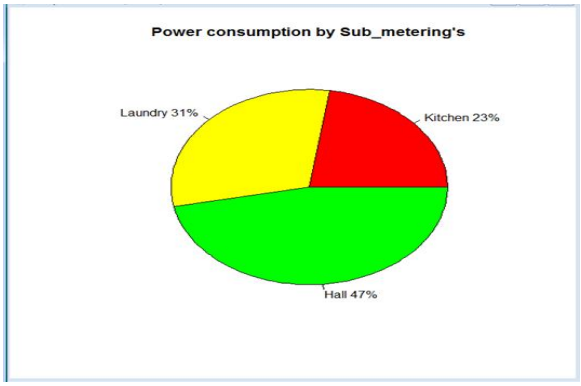

Fig 5 Powerusage separated into three sub meters

Measurement description:

1. Smart meters gives inspiration to active power every minute (in watt hour) in the householddata in submeters and global active power.
2. In the dimensionsbased on the timestamps slightly represented bydatasets values andseparatedbetween two repeatedsemi-colon attribute.

Data Set Attributes information:

1. time: time in format hh:mm:ss
2. date: Date in format dd/mm/yyyy
3. voltage: minute avg voltage (in volt)
4. global intensity: household global minute avg current intensity (in ampere)
5. global_active_power: household global minute avg active power (in kilowatt)
6. global_reactive_power: household global minute-avg reactive power (in kilowatt)
7. sub_metering 1: power usage sub-metering(in watt hour) related kitchen, dishwasher and a microwave .
8. sub_metering2: power usage sub-metering(in watt hour) laundry room, washing engine, a tumble drier, a refrigerator and a light).
9. sub_metering3: power usage sub-metering(in watt hour) water warmer and air conditioner.

The data set records used in power utilization energy of total customers over 20,75,259four years (1410 days) at a granularity of 60 seconds. The poorweightprofile are roughly identified by detecting weight profiles with absent values or all zeroes.

B. Modeling the Power Usage Dynamics for Each Customer

Each and every customers are regular routine of power usage based on their utilization reasonably day type defined as below :[8]

Early Morning (00:00-06:30),
Morning (6:31-11.30),
Noon (11:31-14:30),
Evening (14.31-19.30) and
Night (19:31-24:00).

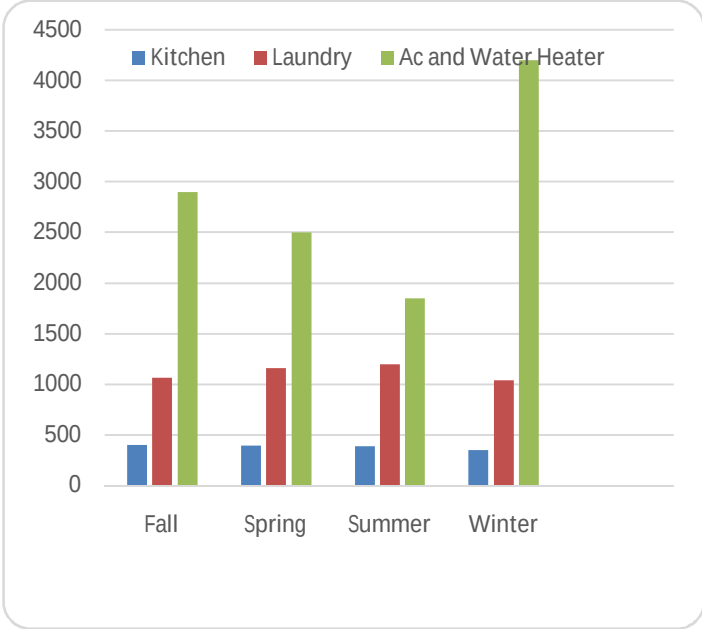

Fig 5 2Dplane map for full periods of customers their usage type.

On this basis, the load data are transformed into PAA representations which also vary from 0 to 1.It can be see that the higher power usage of lower density. For further analysis separatedinto a four seasons. Below are the season defined based on weather conditions.

1) Fall
2) Winter
3) Summer
4) Spring

These seasons to show howhouseholds power usage varies base on weather condition. This performance helps the power usage distributers to manage the electricity loads based on the demand.

C. Clustering for Full Periods

The typical dynamic characteristics of power usage and segment customers into several groups clustering technique is applied to the full periods. The plotting of the local density and expanseof each customer considered according to decision graphchoose the density peaks, where a total clusters can be obtained, which have been marked with diverse colors.

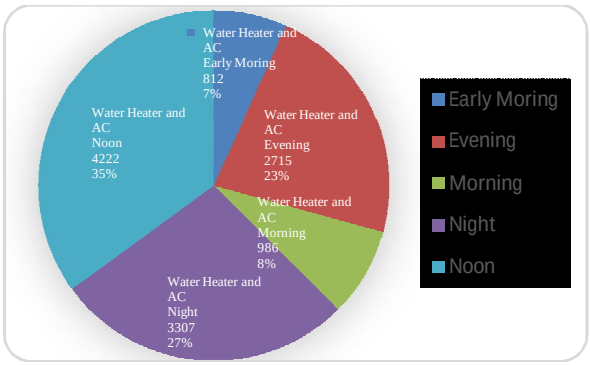

Fig 6 Decision graph to find density peaks for full periods.

The allocation of mapped customers into a 2-D plane according to their dissimilarity matrix by Multi Dimensional Scaling (MDS) [7]is a very effective dimensional decrease way for visualizing the stage of similarity among different objects of a data set. Each object is placed in N-dimensional, amid object distances to beconserved as close as achievable. Each point in the plane stands for a customer. Points in the equal cluster are marked with the similar color. It isobserved that the users of dissimilar groups are irregularlydispersed. Approximately 90% of the customers belong to the 10 larger clusters, whereas the other 10% are distributed in the other 30 clusters. In this way, these 6445 customers separated into diverseclubs in accordance to thepower energyutilizationactivedistinctiveness for complete span of time.

D. Clustering of Each Adjacent Periods

The full periods of vibrant characteristics and instead a certain time period. The demand response potential in no peak shaving of each customer to evaluate the dynamics from Periods day type 1 to 2 are much more vital to measure the potential change of wind power at midnight, the dynamics from Periods day type 4 to 1 should be emphasized. It is necessary toconduct customer segmentation into different adjacent periods.The four period of customers distributedby produced such as bells and clustering methods are efficiently address the no spherically data. Conventional the dynamics from day type is more diversity because people more active during the day type in season less power usage.

E. Distributed Clustering

The ration between thecompressed data and original data volumes is called as the Compression Ratio(CR).

$$CR = \frac{\text{No.of local models}}{\text{No.of the whole objects}}$$

To estimatethe presentation of the algorithm used in both the distributed and centralized clustering processes. The high consistency indicates good act of the distributed algorithm. The matching rate of the algorithm with centralized that can be high in 96.47%. Higher clustering quality with a lower CR. In accumulation of time and space complexity of the modified FSFDP in global model is $O((CR \bullet N)^2)$. It means that the efficiency of the worldwidecluster has improved by $(1/CR)^2$ times, where CR<1 holds. In this case, the efficiency has been boosted to approximately $(1/0.065)^2 \approx 235$ times.

5. CONCLUSION

Our main idea is the analyzing of power usage of households data. Simple API for Xml and time-based Markov models are different from static prospective, there utilized activedistinctiveness of every clustering technique. Clustering Fast Search and Find Density Peaks (CFSFDP)is performed to discover theclassic dynamics of power usage and segment customers into different groups. Finallya time area analysis and day type is conducted on the result of the dynamic analyzing to identify the requiredpotential response of each group's customers. The challenges of massive high spatial power utilization information are marked in smart meters. Simple API for Xml technique is used to ease the cost of data communicate, storage and reduce the numerical power data. Markov model to several transition matrixes and transform long-term data. Data sets are distributed by clustering algorithm and analyzed the power usage based on various meters (Kitchen, Hall, and Laundry), Times (for every minute) and day types. Further work will extendon the temperature and dynamic data analytics.

REFERENCES

[1]. P. Panapakidis, M. C. Alexiadis, and G. K. Papagiannis, "weight profiling in the deregulating power usage retail markets: A review of the applications," in Proc. 9th IntConfEur. Energy Market (EEM), Florence, Italy, 2012, pp. 1–8.
[2]. Y. Wang et al., "weight profiling and its application to requirereaction: A review," Tsinghua Sci. Technol., vol. 20, no. 2, pp. 117–129, Apr. 2015.
[3]. D. Niu, H. Shi, J. Li, and C. Xu, "Research on power weight forecasting based on united model of Markov and BP neural networks," in Proc.8th World Congr. Intell. Control Autom. (WCICA), Jinan, China, 2010, pp. 4372–4375.

[4]. A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks,(cfsfdp)" Science, vol. 334, no. 6191, pp. 1492–1496, 2014.
[5]. J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic illustration of time series, in Proc. 8th ACM SIGMOD Workshop Res. Issues Data Min. Knowl. DiscovSan Diego, CA, USA, Jun. 2003, pp. 2–11.
[6]. S. Kullback and R. A. Leibler, "in order and adequacy," Ann. Math. Stat., vol. 22, no. 1, pp. 79–86, 1951.
[7]. J. D. Leeuw, "Multidimensional scale," in Handbook of Statistics, vol. 31. Amsterdam, The Netherlands: Elsevier, 2001, pp. 285–316.
[8]. J. Torriti, "A review of households power usage order of time models." Renew. Sustain. Energy Rev., vol. 37, pp. 265–272, Sep. 2014.

AUTHOR

B. Mohan Krishna, pursuing M.Tech in Computer Science and Engineering from Jawaharlal Nehru Technological University, Anatapuramu (JNTUA).

K. BalaChowdappa, working as an Assistant professor in Computer Science and Engineering Department. G Pulla Reddy Engineering College (Autonomous), Kurnool.