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

FEATURE EXTRACTION FROM END-USERS LOAD PROFILE TO CLASSIFY CUSTOMERS INTO DIFFERENT GROUPS

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

Abstract- As the technology has been excelled and the traditional grid is moving towards the smart grid. As every day much new equipment is being installed like sensors, smart meters and the volume and velocity of data is increasing rapidly. With the increase in the smart meters, the opportunities for data analysis have been increased but also increased the computation timings and the cost for hardware. So, to overcome these above problems the dimensionality reduction is a solution. For dimensionality reduction, there are many techniques available in the literature. For this thesis, feature extraction has been chosen to reduce the dimensions from the dataset. The dataset of 200 households is taken which provides the load consumption for every 10 mins. The three features are extracted i.e. means, standard deviation, and symmetry of profile. Using the above feature, the dimension reduction has been achieved. The cluster inertia has also been reduced using the above-extracted features.

Keywords: Mean, standard deviation, and symmetry analysis of households

INTRODUCTION

Function extraction is a process of dimensionality depletion by which a preliminary set of raw data is decreased to more manageable groups for processing. A typical of these huge data sets is a large number of variables that require a lot of computing resources to the procedure. Feature extraction is the name for methods that select and /or combine fickle into features, effectively decreasing the amount of fact that must be a procedure, while still accurately and completely reporting the original data set.

Why is this Useful?

The procedure of feature extraction is practical when you need to decrease the number of resources needed for proceeding without losing necessary or relevant information. Feature extraction can also decrement the amount of unnecessary data for a given analysis. Also, the reduction of the data and the machine's attempt in building variable combinations (features) facilitate the speed of learning and generalization pace in the machine learning process.

Image Processing

Algorithms are used to find attributes like shapes, edges, in images or videos. Load profiles express the degree to which the system resources (CPU, memory, and network capacity) are loaded in actuality. The loading is usually shown in terms of the number of users or the frequency of times that a transaction is conveyed out at a particular time. Usually, the loading of a system is not continuously even, but changes over some time: there are peaks and valleys within a 24-hour extend. Often, weekends will express a unique loading from weekdays. And during holiday periods and public holidays, the loading of a set-up may look unique again.

For the formation of a load profile, information from the following origin is merged:
Measuring the loading of the set-up using particular tools (monitors).

- The responsibility for this resides with a department for "Technical System control" Interviewing users
- This amounts to the following questions: "Which deal do you convey out? How often, and when?"

The testing of load profiles comes less than what is often mention to as "performance testing "and is a testing forte in itself. While it is feasible to do it manually, tools are usually working that generate a particular loading of the set u-p. Using the tools, a practical loading is *simulated*, such as:

Creation of virtual users
A virtual person is a tiny program that simulates a person. On one PC, not some such packages can run without delay. This avoids the use of the bodily presence of a distinct PC for each user. This is mainly implemented for subjecting the entire machine, including the community, to a particular loading

Offering transactions through the database-control interface
This creates a sure loading of the back-end of the set-up without overloading the front-end or the network. It eases direct measurement of whether the database server has suitable dimensions.

Decreasing the number of character to need during a statistical analysis can lead to several benefits such as:

- Accuracy improvements.
- Overfitting risk reduction.
- Speed up in training.
- Improved Data Visualization.
- Increase in explain ability of our model.

It is statistically proven that when carrying out a Machine Learning task there exists the best number of features that should be old for every particular task. If more attributes are contributed than the strictly critical ones, then our model overall performance will just lessen (because of the added noise). The real challenge is to find out what is the excellent quality of attributes to use (this is, in fact, dependent on the amount of data we have at hand and by the complexity of the work we are trying to attain). That's where attributes Selections techniques come to our rescue.

1.4 Feature Selection

Many separate methods can be applied for the Feature chosen. Some of the primes are:

1. Filter Method filtering our dataset and taking only a subdivision of it containing all the to the point features (e.g. correlation matrix using Pearson Correlation).

2. Wrapper Method follows the same objective of the Filter procedure but uses a Machine Learning version as its assessment criteria. We feed a few features to our Machine Learning model, evaluate their products and then decide if increment or decrement the feature to increment accuracy. As a result, this method can be most accurate than filtering but is most computationally costly.

3. Embedded Method Similar Filter Method the Embedded Method made use of a Machine mastering version. The dissimilarity among the two strategies is that the Embedded approach examines the precise training iterations of our ML model and then ranks the vital of every feature based on how a whole set of each of the attributes contributed to the ML model training

1.5 Types of Load Profiling
Load data preparation,
Load curve clustering,
Clustering evaluation,
Customer segmentation
Result application.

1.6 Clustering Techniques
The five most popular clustering techniques including hierarchical clustering, k-means, follow-the-leader, fuzzy k-means, and fuzzy classification, and then analyzed the dissimilar among the Techniques
The main contributions include:
(1) Extending new mentioned clustering methods especially for the indirect clustering to expose a extra complete summary;
(2) Proposing a evaluation on the categorization strategies of client which may be very vital for patron segmentation;

(3) Focusing on the utility of load profiling to demand response;
(4) Addressing a few challenges and possibilities of load profiling in the huge statistics
1.8 Feature Extraction using PCA

1.9 Feature Extraction using CNN

II.PROBEM DEFINING

A large amount of data is entered every second from smart meters. To analyze and store this much amount of data will require a lot of additional computations and hardware. It will increase the installation and operating costs for data analysis. Therefore, the reduction in computation and hardware is much required. One of the dimensionality reduction techniques is to extract features. Therefore, feature extraction for load profiles of customers should be studied.

III. Objectives

Based on the above discussion the following objectives are formed:
Data science is a comparatively new branch and does not have roots in various fields. Power system analysis is also one of a similar branch. The applications of data analysis in power systems are still under a fledgling stage. There is not so much literature available on feature extraction for demand response. Therefore, an intensive study of feature extraction techniques for demand response should be done.

1. To study and analyze different feature extraction techniques for demand response.
2. To study and analyze different techniques for extracting features from customers' load profiles based on their patterns.
3. To classify customers in different groups based on the feature extraction.
4. To create a foundation for multi-stage clustering of customer's load profiles

IV. Research Methodology

For the extraction of the features, the most used features which are used in literature and most favorites of most of the researchers of this field are, mean, standard deviation skewness. Therefore, to reduce the dimensions in this thesis, the following parameters or features of different customers are extracted:

- **Mean**

$$m = \frac{\sum d_n}{N} \qquad (1)$$

It is an average of the load consumption for a given period. It is simple yet very useful criterion.

- **Standard deviation**

$$\partial = \sqrt{\frac{1}{N}\sum(c_i - m)^2 . n_i} \qquad (2)$$

Standard deviation is the measure of the variation in consumption from average consumption. It is also simple to calculate but also useful as it shows the rigidness of a load profile

Skewness

Skewness is the measure for the degree of symmetry for the degree of consumption's symmetry for a customer's load profile. It is measured by two indices namely Skew and Kurt. Skew is the third degree movement whereas Kurt is the fourth degree moment.

$$Skew = \frac{1}{N\partial^3}\sum(c_i - m)^3 . n_i \qquad (3)$$

$$Kurt = \frac{1}{N\partial^4}\sum(c_i - m)^4 . n_i \qquad (4)$$

After extracting these features, the feature vectors are formed. Based on these feature vectors the clustering is performed using the K-means clustering technique.
In k-means nearest neighbor, the number of clusters is not defined. To check the number of optimal clusters the elbow method is used. In this method, the program starts iteration from 1 cluster and goes up to the max no of clusters. At each iteration, the error is calculated and when the error decrement decreases to a limit then it is considered the optimal number of clusters.

V. Result &Discussion

A dataset shown in table 1 has been selected for analysis in this thesis, the dataset contains 200 Household electrical load profile. The electrical consumption is given every 10 mins.

The dataset of 200 households is chosen and the dataset is converted from 10 to one hour. Then maximum load consuming households is selected which is household number 92 with the mean consumption of 26786.58 Watts Hour. The load consumption profile for household 92 is shown in table 3

House Number	Mean	Standard deviation	Skew	Kurt
Household 1	3555.831250	1595.308103	1.153587	0.462750
Household 2	6513.447500	2249.696071	-0.215181	-1.078383
Household 3	9707.082083	3602.791242	0.561726	-0.775356
Household 4	5671.462500	3295.761937	0.999882	0.047010
...	...	...	...	...
Household 196	6977.748333	4850.040552	1.733616	2.836122
Household 197	4294.911667	2409.987992	1.458054	3.096514
Household 198	13434.912500	1813.135251	1.393651	1.257262
Household 199	17503.191667	4220.794700	0.773752	-0.478644
Household 200	5736.235417	5383.996453	3.259506	12.279809

Table 1: Mean, standard deviation, and symmetry analysis of households

Table 3: Mean, Standard Deviation, And

Symmetry Analysis for household 92

Table 3: Mean, Standard Deviation, And Symmetry Analysis for household 92

Feature	Value
MEAN	7055.03
STD	4549.05
MIN	1380
25%	3875.80
50%	6017.39
75%	8850.13
MAX	40911.17

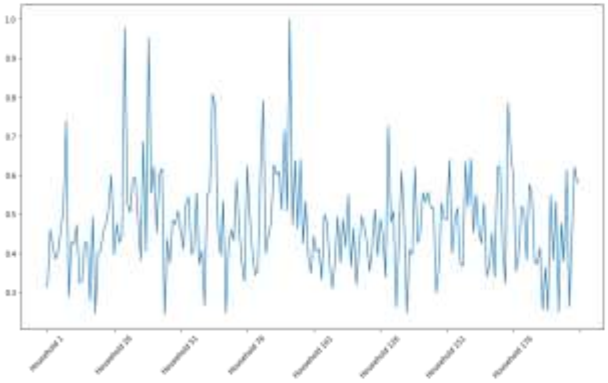

Fig 3 Standard deviation of 200 households

Here, it can be noted that the standard deviation of most of the households is not very different. For most of the households, it varies between 0.32 to 0.63. The customer with a standard deviation below 0.3 is very rigid and should not be targeted for demand response.

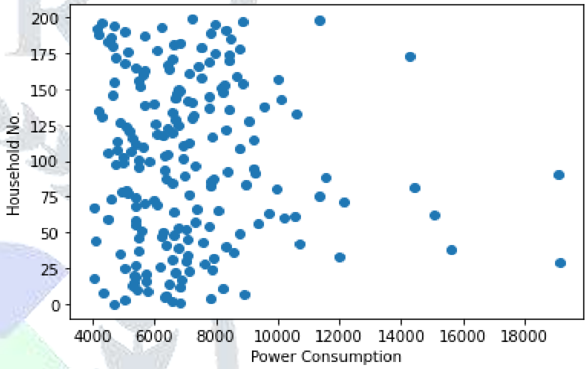

Fig 4: Scatter plot of the mean of 200 customers

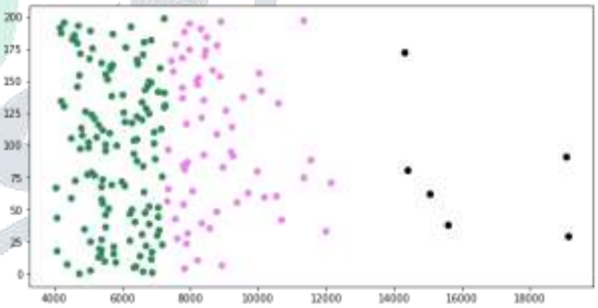

Fig 5. Scatter plot of different clusters using K-Means clustering technique

It can be noticed that the customers can easily be classified into different categories which can be further clustered into different groups.

VI.Conclusion & Future Scope

A dataset of 200 households was taken in which the power consumption for very 10 mins was given. Using this dataset, the three features were extracted. The features were mean, standard deviation, and symmetry analysis. For symmetry analysis, two indices that are skewness and kurtosis were chosen. Accordingly, these three features were extracted for all the 200 households and a new dataset of 200 households with these three features as a column has been created.

Plotting means and standard deviation of the household has proved that they are very essential features. The graph of means and standard deviation has shown that 80 % of the households were very similar to each other. The dimensionality reduction has also been achieved by replacing 28 customers with means consumptions within 5000W to 5500W for an hour. The window of 500W can be taken for replacing more customers starting from 3000 to 15000W and a great amount of dimensionality reduction can be achieved.

The feature selection has also improved the clustering. The clustering done with features gives better results, it has been observed in this thesis that while clustering with K-means with these features reduces te cluster's inertia5.4×10^{10}.

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.