International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

Target response electrical usage profile clustering using Big data

M.Thilagam[1], Ms.J.Kalaivani[2], Mrs.P.Hemalatha[3]

[1] *B.Tech (Information Technology), IFET College of Engineering, Villupuram,*

[2]*Associate Professor, Dept.of Information Technology, IFET College of Engineering, Villupuram,*

[3]*Asst Professor, Dept. of Information Technology, IFET College of Engineering, Villupuram.*

Abstract - *Data streams are very large, quick-changing, and unable to calculate. Clustering is a prominent task in mining data task; it can group same kind of objects in a cluster. The aim of choosing a Re-Cluster subset group of good characteristics with respect to the goal concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, accuracy learning, and improving outcome unambiguousness. While the effectiveness concerns the point in time necessary to find a re-cluster division of features, the efficiency is related to the value of the subset of features. In this, proposed clustering related to division selection algorithm works in two steps. In the first step, further are divided into clusters by using theoretic graph clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. To confirm the algorithm efficiency, we are working to use mRMR method with heuristic procedure. Heuristic algorithms used for solving a problem more quickly or for finding an approximate rearrange the cluster subset selection solution. Minimum Redundancy Maximum Relevance (mRMR) variety used to be more controlling than the extreme consequence selection. It will provide active way to expect the efficiency and success of the clustering based subgroup collection algorithm.*

Key Words: Cluster analysis, Load profiling, big data, Markov model, behavior dynamics, distributed clustering, demand response.

1. INTRODUCTION

All over the world have some set of goals to implement the power system in monopolistic area mainly focused on demand side. Now days the load serving entities (LSEs) is used development of high values. To have a better understanding of electricity consumption patterns and power managements are effective ways to enhance the competitiveness of LSEs. It has been revolutionizing the electrical generation and consumption by a two-way flow of power data. Most important data source from the demand side, advanced metering infrastructure (AMI), has gained increasing popularity worldwide; AMI allows LSEs to obtain electricity consumption data at high frequency, e.g., minutes to hours Large volumes of electricity consumption data[16] reveal .By the Research Report, the determine that smart meters will surpass 1.1 billion by 2022 . AMI will collect the electricity usage data profile in the range among 1 hour; This will increase in the amount of usage of electricity will processed in the past years. It means that by 2022 the electric utility of power in industry will be increase the data annually from smart meters. The primary and secondary value embedded in the high density and same data sets from power distribution systems. Aggregated load has already been successfully modeled using top-down methods. Singh et model distribution system load and Valverde et al. model load for load flow analysis with Gaussian mixture models to capture the probability density functions. However, autocorrelation found in electricity request of households was never combined. Bottom-up methodologies have in general good results because of the incorporation of a performance model. Top-down approaches have a lot of potential because of the lower modeling intensity: there is no need to model every appliance individually, which lowers the intensity of modeling significantly. The detection of behavior is in general done by pattern analysis. Techniques have been developed to find similarities within load profiles as between profiles within different domains such as clustering or classification of profiles forecasting selecting scenarios for load-wind combinations and selecting demand response policies a new short-term load forecasting framework based on big data technologies is proposed in this paper. In Section II, the framework and relevant techniques of the short-term load analysis and forecasting method are presented in detail. Section III introduces a technical framework of the proposed method using big data technologies. Section IV provides case study results. Section V concludes this paper. In general, short-term forecasting methods perform direct forecasting of the total system load using historical load data and weather data as inputs. However, since the grid consists of thousands of individual users and many time varying characteristics, a single forecasting method, such as those mentioned earlier, cannot adequately forecast individual loads, as well as the accompanying factors that influence the

variations in these loads. Therefore, current approaches, which treat all users as a single entity, sometimes may not be able to meet accuracy requirements under all circumstances. Another issue is that the load needs to be forecasted at the substation or bus level for calculation of the power flow. Most utilities do not process load forecasting at the substation or bus level because of the complexities involved in capturing the necessary information or because there is very little data available.

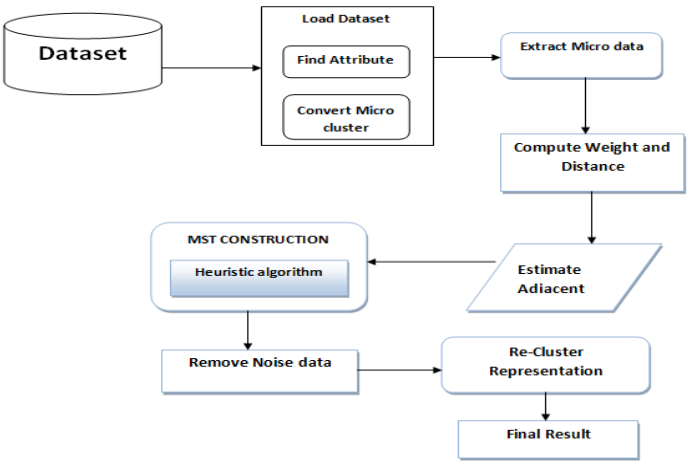
2. EXISTING SYSTEM

Data clustering is typically prepared as a two-stage process with a wired part which encapsulates the data into many micro-clusters or grid cells and then, in an offline process, these micro-clusters (cells) are re-clustered/combined into a smaller number of final clusters. Since the re-clustering is an offline process and thus not period critical, it is typically not discussed in detail in papers about new data stream clustering systems. Most papers suggest using an (sometimes slightly modified) existing conventional clustering algorithm (e.g., weighted k-means in CluStream) where the micro-clusters are used as pseudo point opinions. Another method used in Data Stream is to use reach ability where all micro-clusters which are less than a given distance from each other are connected together to arrange clusters. Grid-based algorithms typically merge adjacent dense grid cells to form larger clusters (see, e.g., the original version of D-Stream and MR-Stream). The number of clusters differs over period for some of the datasets. This needs to be considered when associating to clusters, which uses a stable number of clusters. This reduces the speed and accuracy of learning algorithms. Some existing systems doesn't removes redundant features alone

3. PROPOSED SYSTEM

In proposed system, the develop and determine a new method to give solution for this problem in micro-cluster-based algorithms. Here introducing the concept of a density graph which explicitly absorb the density of the original data between micro-clusters during clustering and then show how the graph can be used for re-clustering micro-clusters. In this project, proposed Clustering related to sub portion of selected method uses minimum spanning tree-based method to cluster characteristic. our proposed algorithm is not only focused on specific data type.. Thus, characteristic of divided part will able to identify and delete as much of the unwanted and repeated data as possible. Moreover, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other."In our proposed Cluster based subset Selection algorithm, it involves the construction of the minimum spanning tree from a weighted complete graph; the partitioning of the MST into a forest with each tree representing a cluster; and the selection of representative features from the micro-clusters.

4. ARCHITECTURE DIAGRAM



5. MODULES

A module is a part of a program. Programs are composed of one or more independently developed modules that are not combined until the program is linked. A single module can contain one or routines.

Our project modules are given below:

5.1 Load Data and Convert Micro Data

5.2 Compute Density Value

5.3 Estimate Adjacent Relevance between Each Data

5.4 Calculate Correlate and Remove Noise

5.5 Heuristic MST Construction

5.6 Cluster Formation

5.1 LOAD DATA AND CONVERT MICRO DATA

Load the data into the process. The data[16] has to be preprocessed for removing missing values, noise and outliers. Then the given dataset must be converted into the arff format which is the standard format for WEKA toolkit. From the arff format, only the attributes and the values are extracted and stored into the database. By considering the last column of the dataset as the class attribute and select the

distinct class labels from that and classify the entire dataset with respect to class labels.

5.2 COMPUTE DENSITY VALUE

Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation.[2] To find the relevance of each attribute with the class label, Information gain is computed in this module. This is also said to be Mutual Information measure. Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between feature values or feature values and target classes.

5.3 ADJACENT RELEVANCE ESTIMATION

The relevance among the feature Fi € F and the objective concept C is referred to as the T-Relevance of Fi and C, and represented by SU (Fi,C). If SU (Fi,C) is greater than a determined threshold , we say that Fi is a strong T-Relevance feature.

$$SU(X,Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)}.$$

After definition the relevance value, the redundant attributes will be removed with reverence to the threshold rate of data

5.4 CALCULATE CORRELATE AND REMOVE NOISE

The correlation between any two set of features Fi and Fj (Fi,Fj € ^ F ^ i ≠ j) is called the F-Correlation of Fi and Fj, and denoted by SU(Fi, Fj). The equation similar ambiguity which is used for identifying the relevance between the attribute and the class is again applied to find the comparison between two attributes with reverence to each label.

5.5 HEURISTIC MST CONSTRUCTION

With the F-Correlation value computed above, the heuristic Minimum Spanning tree is constructed. For that, we use heuristic algorithm which form MST excellently.

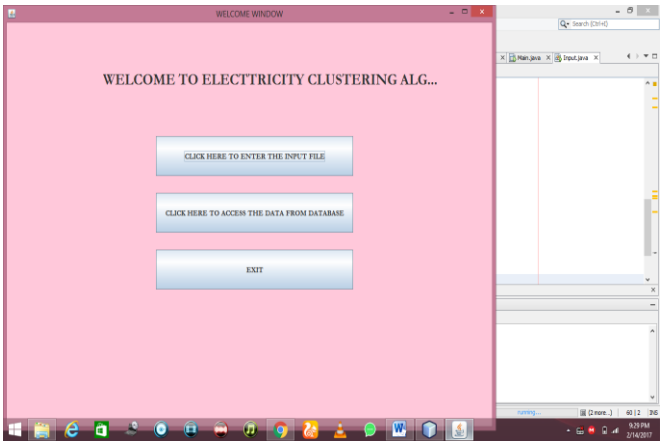
Heuristic algorithm is a greedy algorithm in graph model that finds a minimum spanning tree for a connected subjective graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where the complete weight of all the edges in the tree is minimized. If the graph is not linked, then it finds a minimum spanning forest (a minimum spanning tree for each connected component).

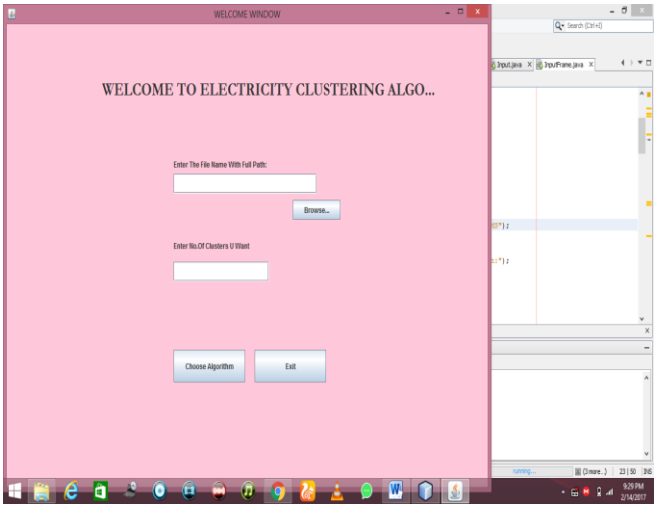
5.6 CLUSTER FORMATION

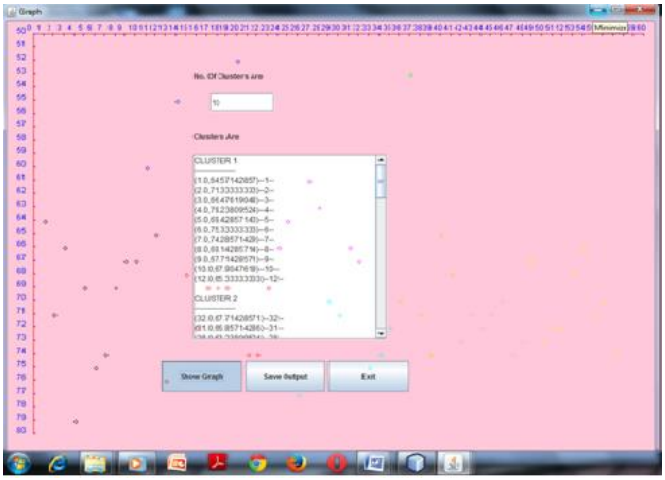
After construction the MST, in the third step, we first eliminate the edges whose weights are smaller than both of the T-Relevance SU (Fi`, C) and SU(Fj`, C), from the MST. After eliminating all the unnecessary edges, is obtained. Each tree Tj € Forest represents a cluster that is denoted as V (Tj), which is the vertex set of Tj as well. As illustrated above, the features in each cluster are redundant, so for each cluster V (Tj) we choose a characteristic feature Fj R whose T-Relevance SU(Fj R,C) is the greatest.

6. IMPLEMENTATION AND RESULT

In this system, here analyzing the data from electrical usage by daily basis and cluster into their usage profile and then they identified by graphical manner and fulfill the demand of Electric power to the user. Implementation is the phase of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in completing a popular new system and in giving the user, confidence that the new structure will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constrictions on implementation, designing of methods to achieve changeover and evaluation of changeover methods.







7. CONCLUSION

In this project, developed the first data stream clustering algorithm which clearly records the density in the part shared by micro-clusters and uses this information for reclustering. Experiments also show that shared-density reclustering already executes extremely well when the online data stream clustering element is set to produce a small number of large MCs. A heuristic algorithm used for solving a problematic more quickly or for finding an approximate re-cluster subset selection solution. Lowest Redundancy Maximum Relevance assortment used to be more powerful than the extreme relevance selection. It will provide effective way to predict the efficiency and effectiveness of the clustering based subset selection algorithm.

REFERENCES

[1] S. Tabibian, A. Akbari and B. Nasersharif, "Speech enhancement using a wavelet thresholding method based on symmetric Kullback–Leibler divergence," *Signal Processing*, vol. 106, pp. 184-197, 2015.

[2] G U Rui-Chun, J Y Wang. "A Parallel Clustering Model Based on MapReduce," *Computer & Modernization*, 2014

[3] Z Sun, G Fox, W Gu, "A parallel clustering method combined information bottleneck theory and centroid-based clustering," *Journal of Supercomputing*, vol. 69, pp. 452-467, 2014.

[4] Y Xiao, J Yang, H Que, "Application of Wavelet-based clustering approach to load profiling on AMI measurements," in *Electricity Distribution (CICED), 2014 China International Conference on. IEEE*, pp. 1537-1540, 2014

[5] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Science and Technology*, vol. 20, pp. 117-129, 2015.

[6] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of Low Voltage Network Templates-Part I: Substation Clustering and Classification," *IEEE Trans. Power Systems,* vol. 30, pp. 3036-3044, 2015.

[7] K. Zhou, S. Yang and C. Shen, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 103-110, 2013.

[8] G. J. Tsekouras, P. B. Kotoulas, C. D. Tsirekis, E. N. Dialynas, and N. D. Hatziargyriou, "A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers," *Electric Power Systems Research*, vol. 78, pp. 1494-1510, 2016.

[9] S. V. Verdu, M. O. Garcia, C. Senabre, A. G. Marin, and F. J. G. Franco, "Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps," *IEEE Trans. Power Systems*, vol. 21, pp. 1672-1682, 2006.

[10] G. Chicco and I. S. Ilie, "Support Vector Clustering of Electrical Load Pattern Data," *IEEE Trans. Power Systems*, vol. 24, pp. 1619-1628, 2009.

[11] M. Piao, H. S. Shon, J. Y. Lee, and K. H. Ryu, "Subspace Projection Method Based Clustering Analysis in Load Profiling," *IEEE Trans. Power Systems*, vol. 29, pp. 2628-2635, 2014.

[12] G. Chicco, O. Ionel and R. Porumb, "Electrical Load Pattern Grouping Based on Centroid Model with Ant Colony Clustering," *IEEE Trans. Power Systems*, vol. 28, pp. 1706-1715, 2013.

[13] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, pp. 68-80, 2012.

[14] I K. Fodor, "A Survey of Dimension Reduction Techniques," *Perpinan*, vol. 205, pp. 351-359, 2003.

[15] M. Abrahams and M. Kattenfeld, "Two-stage fuzzy clustering approach for load profiling,", in *Universities Power Engineering Conference (UPEC), 2009 Proceedings of the 44th International*. pp. 1-5, 2009.

[16] NasrinBanu.A, Sindhuja.K and **Suganthi.V**, "Survey on Secured Proxy Based Distributed Data Storage in Public Cloud Database," *International Journal of Science, Engineering and Technology Research,* vol. 4, no. 3, pp. 555-558, Mar 2015

[17] G. Chicco, R. Napoli and F. Piglione, "Comparisons Among Clustering Techniques for Electricity Customer Classification," *IEEE Trans. Power Systems*, vol. 21, pp. 933-940, 2006.

[18] E. D. Varga, S. F. Beretka, C. Noce, and G. Sapienza, "Robust Real-Time Load Profile Encoding and Classification Framework for Efficient Power Systems Operation," *IEEE Trans. Power Systems*, vol. 30, pp. 1897-1904, 2015.

[19] S. Zhong and K. Tam, "Hierarchical Classification of Load Profiles Based on Their Characteristic Attributes in Frequency Domain," *IEEE Trans. Power Systems,* vol. 30, pp. 2434-2441, 2015.

[20] J. Torriti, "A review of time use models of residential electricity demand," *Renewable and Sustainable Energy Reviews*, vol. 37, pp. 265-272, 2014.

[21] Y Xiao, J Yang, H Que, "Application of Wavelet-based clustering approach to load profiling on AMI measurements," in *Electricity Distribution (CICED), 2014 China International Conference on. IEEE*, pp. 1537-1540, 2014.

[22] A Notaristefano, G Chicco, F Piglione. "Data size reduction with symbolic aggregate approximation for electrical load pattern grouping," *Generation, Transmission & Distribution, IET*, vol. 7, pp. 108-117, 2013.

[23] A. Albert and R. Rajagopal, "Smart Meter Driven Segmentation: What Your Consumption Says About You," *IEEE Trans. Power Systems*, vol. 28, pp. 4019-4030, 2013.

[24] M Rodriguez, I González, E Zalama, "Identification of Electrical Devices Applying Big Data and Machine Learning Techniques to Power Consumption Data," in *International Technology Robotics Applications Springer International Publishing*, pp. 37-46, 2014.

[25] A Rodriguez, A Laio. "Clustering by fast search and find of density peaks," Science, vol. 334, pp. 1492-1496, 2014.