AN APPROACH FOR DEMAND SIDE MANAGEMENT USING K- MEANS CLUSTERING

Shraddha Sharma[1], Prof. Seema Pal[2]

[1] Dept. of Electrical Engineering, Jabalpur Engineering College, Jabalpur, Madhya Pradesh, India
[2]Assistant professor, Dept. of Electrical Engineering, Jabalpur Engineering College, Jabalpur, Madhya Pradesh, India

---***---

Abstract - *Smart meter is an advanced metering infrastructure (AMI) that includes a smart meter, a bidirectional communication network, and a data management system. Using data analytics and machine learning to evaluate high-frequency smart meter data yields important insights into home power consumption trends, as well as improved load forecasting and demand response management implementation. In this study, Principal Component Analysis (PCA) is employed as a dimensionality reduction technique to extract features from a dataset collected from the UMassTrace repository. The K-means unsupervised partitional clustering algorithm uses three distance metrics to cluster reduced data: Euclidean, Manhattan, and Pearson correlation distances. MATLAB programming software is used to do feature computation and clustering. The clustering model is evaluated by obtaining the average silhouette coefficient. Euclidean distance is obtained to perform best with better average silhouette coefficient, indicating that data points in a cluster are compact and far apart from other clusters, making distance measurement preferable for clustering consumer load profiles for better demand side management.*

Key Words: **Smart meters, Dimensionality reduction, PCA, K-means, Manhattan distance, Euclidean distance, Pearson correlation distance, Average silhouette coefficient, Demand response management**

1. INTRODUCTION

Advanced metering infrastructure (AMI) which comprises of smart meter, bidirectional communication network and data management system are being increasingly deployed in recent years. They have significant role by providing benefits to end consumers, network operators and energy suppliers. Smart meters offer range of functions such as advance metering, control, data storage and communication technologies .It helps consumers by providing them near real time consumption patterns which help them to manage their energy usage, reduce greenhouse gases emission and save money[1].It improves demand management, network planning and operation by providing accurate demand forecast ,locate outages and shorten supply restoration time, reduce operational and maintenance costs of network and improve asset utilization in distribution[2][3]. Smart meters generate enormous amount of high frequency data, which exhibits the characteristics of Big data i.e. velocity, volume, variability, variety and value thus, require a robust communication infrastructure for data processing and storage at utility end. This data being highly dimensional in nature, greatly impacts the analysis and deduction process making it less efficient due to curse of dimensionality. This challenge necessitates the use of dimensionality reduction algorithms or techniques.

Dimensionality reduction techniques convert high dimensional data to reduced dimension without the loss of significant information. These techniques when employed reduces the computational complexity associated with smart meter data, as every data obtained from smart meters are not helpful in drawing useful conclusions[4]. Once converted to lower dimensionality, these data can be used by consumers and utility operators to deduce important results and understand energy consumption trends, anomaly detection, energy theft and better demand side management.

Energy consumption behaviors of individual consumers are used by utility for improving better demand side management. It selects the appropriate number of consumers to participate and present precise data on peak energy consumers. Clustering is used to group the load profile of different types of consumers in a distribution network. The main basis of clustering is to group load profile in different clusters with minimum intra-cluster distance or maximum intra-cluster similarity and maximum inter-cluster distance or minimum inter-cluster similarity. The two broad categories of clustering methods are hierarchical and partitional clustering methods. Hierarchical clustering groups the load profile into different clusters by generating nested partitions [5].In Partitional clustering method each cluster is represented by its center which summarizes all the load profile present in the cluster. The main focus is to optimize the objective function, which is the distance between the center and all the load profiles.

In this paper, PCA has been used for dimensionality reduction and k -means partitional clustering method for clustering of different consumer profile. An evaluation index, silhouette coefficient is used to compare the clustering done using the three different distance measures Euclidean, Pearson correlation and Manhattan distance for better demand side management.

2. RELATED WORKS

A fairly comprehensive comparison and study of several clustering approach is available in[6]. The research [7] demonstrate the need for caution while obtaining data from time series in order to support the statements made in relation to the findings of an empirical assessment. Although[8] [9] [10] have examined clustering, no assessment of the quality of the resulting clusters which has to do with the clustering strategy chosen has been done, the distance measure under investigation, and a study and explanation of the forms of the resulting distinctive load profiles. In[11] , the daily and segmented load profiles are clustered using K-means clustering algorithm to offer a load estimation technique using four metrics for distance –the Pearson, the Euclidean, Manhattan and Canberra correlations are examined.[12] Compared the clustering findings from four different techniques-random forest, KNN, decision tree, and ANN in order to forecast which consumer would be suitable for demand response management based on the analysis of smart meter data. In[13]comparative analysis between k-means and k -medoids technique is done to identify different energy behavioral groups and apply different pricing rules based on consumption time weekend conditions. In another study[14],k- means along with other techniques using different distance measure such as cosine, Euclidean, correlation and Manhattan are used to cluster consumption patterns based on peak position which can be identified as hurtful moments of the day.

3. DIMENSIONALITY REDUCTION AND CLUSTERING ALGORITHM

In this section a review on dimensionality reduction techniques and clustering algorithm has been done.

3.1. DIMENSIONALITY REDUCTION TECHNIQUES

The process of transforming high dimensional data into a suitable representation with fewer dimensions is known as dimensionality reduction. These dimensionality reduction algorithms can be categorized as supervised, unsupervised and semi-supervised. Supervised algorithms involve labeling a training set of known data .a reliable prediction for the data classes is done using this. Algorithms under this category are linear discriminant analysis (LDA)[15], independent component analysis (ICA)[16], and support vector machine (SVM)[17] and kernel principal component analysis (PCA).

Unsupervised techniques use unlabeled data to find structure. unsupervised dimensionality reduction techniques include Singular value decomposition(SVD),PCA and ICA .Generally ,because labeling data is expensive ,the quantity of labeled data is constrained ,whereas unlabeled data is more readily available. Semi-supervised algorithms include efficient utilization of both labeled and unlabeled data[18].

Principal Component Analysis (PCA)

PCA is one of the most widely used algorithms and is regarded as the best linear dimension reduction method as it reduces the mean square error .PCA seek to locate a linear subspace of reduced dimension d from a given dataset of dimension D such that the data points primarily lie on it given a collection of data on D-dimension. The principal components (PC), a new coordinate system constructed by d orthogonal vector, make up the decreased dimension. The linear combination utilizing the vectors linked to the highest variance is the first PC. The second PC, which is either orthogonal to the first PC or uncorrelated with the second highest variance, is the linear combination of the vectors corresponding to that PC. The same linear formula is used to build other PCs different vectors that represent variations ranging from highest to lowest. Typically, a large number of PCs are obtained, but majority of the variance is explained by the first few PCs and less dominant PCs can be ignored. Hence more energy is concentrated in the lower subspaces.

The Eigen value decomposition of data covariance matrix is represented by[19]:

$$\boldsymbol{E\,X\,X^{T}=E\,\lambda\,I} \qquad (1)$$

To project the data into lower subspace, the Eigen vector corresponding to the most important Eigen value are used after decomposition as follows[19]:

$$\boldsymbol{X_{N\times d}{}^{PCA}=X_{N\times d}\;E_{D\times d}} \qquad (2)$$

Cumulative variance is given by[19]:

$$\frac{sum_{i=1}^{k}}{sum_{i=1}^{D}}=\frac{sum_{i=1}^{k}}{trace(\Sigma)} \quad \text{Here, D> k} \qquad (3)$$

3.2. CLUSTERING ALGORITHM

K- Means Clustering

An unsupervised learning approach called k-means divides ***N×D*** matrix into k clusters. The algorithm's objective is to reduce the distance between the cluster's core and all of its data points. This is referred to as "local optima within cluster sum of squares"[20].Every data point within a cluster is highly similar as indicated by the pairwise distances between each point and its center. The objective function is given as follows[21]:

$$J=\sum_{j=1}^{k}\sum_{i=j,i\in j}^{n}\parallel U_i-\mu_j\parallel^2 \qquad (4)$$

U_i = vector that represents the i^{th} user, i=1, 2, 3…..N

μ_j= vector representing the j^{th} cluster center, j = 1, 2, 3……N

Dimension of each user U_i = [U_i (k), k=1, 2, 3 ...D*]*. The j^{th} cluster center also has dimension of μ_j =[μ_j(k) ,k =1,2,3….D].Typically, the cluster μ_j for j=1,2,3….k are first inferred, ideally a random data points are chosen from the dataset .Every centroid μ_j categorizes the data points U_i such that the distance between data point U_i and all its k centroid is minimum. Euclidean, correlation, city block, hamming and other methods are used to estimate this distance and the center μ_j is updated to represent the average of U_i contained within the clusters.

4. METHODOLOGY USED

In this paper, the high dimensional smart meter dataset is reduced to obtain the important and pragmatic information from the reduced dimension. The reduced data is then used to group the residential consumers based on their consumption patterns for better analysis of load profile. The smart meter dataset used in this paper is from UMassTraceRepository [22]. This dataset contains energy consumption for 443 buildings over the same 24 hour period. The sampling rate is 1 sample per minute. In order to guarantee that data are similar, distance assessment is crucial. This is done in order to ascertain which systems data are supposed to be related to, whether they are similar or not, and what distance measurements are required in order to compare them. The ability to determine a quantitative score of the degree of similarity or dissimilarity of the data (proximity measure) plays a crucial role in the clustering process. Therefore, in order to determine which method is best, it is necessary to compare some of the commonly used methods, namely Euclidean, Manhattan and Pearson distance with a combination of min-max normalization.

- Euclidean Distance: One method for measuring the distance between two pieces of data in Euclidean space is the Euclidean distance (including fields Euclidean two dimensions, three dimensions, or more). Using the following formula [23] one can assess the degree of similarity:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}\left(x_{ik}-y_{jk}\right)^2} \qquad (6)$$

D= distance between *i* and *j*, I as the cluster data center j on the other attribute , *k* symbol of each data, *n* the amount of data , x_{ik} is the data in the cluster to be k, and y_{ik} is the data on the each data to *k.*

- Manhattan (City Block) Distance: The Manhattan (city block) distance is calculates the absolute difference between the coordinates of two objects. The formula used to calculate the distance is as follows-

$$d_{ij} = \sum_{k=1}^{n}\left|x_{ij} - y_{ik}\right| \qquad (7)$$

- Pearson correlation distance: this measure is a dissimilarity measure rather than an actual distance metric. It is derived from the Pearson correlation coefficient as follows[24]:

$$\text{Pearson distance} = 1-r \qquad (8)$$

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \qquad (9)$$

r = correlation coefficient

$\overline{x}$ = Mean of the values of the x variable

$\bar{y}$ =Mean of the values of the y variable

Cluster analysis Technique – process of grouping data is done through general stages of the k-means clustering algorithm which includes normalization of data. In this paper, the data has been normalized using min-max normalization technique.

According to Min-Max Normalization

$$\text{Normalized data (X')} = \frac{x-\min(a)}{\max(a)-\min(a)} \qquad (10)$$

X' is the normalized data, x is the data per column min (a) and max (a) are the minimum and maximum value of data per column. In k-means the number of clusters and each cluster is assigned a centroid (cluster center) randomly. Clusters are assigned to data, based on distance calculation between the data and the centroid of each cluster; here we have used Euclidean, Manhattan and Pearson correlation distance as distance metric. Every time when a data is assigned to a cluster, the centroid of the cluster is again updated and the same clustering process goes on till, the centroid is not changing anymore or same set of data are obtained in clustering process or max iteration have reached.

As a means of criteria to estimate the performance of each distance metric, average silhouette coefficient is calculated. Silhouette Coefficient can be calculated through following equation:

$$s(i) = \frac{b(i)-a(i)}{\max\left(a(i),b(i)\right)} \qquad (11)$$

S (i) is normalized silhouette coefficient

b (i) is the average distance of the data in one cluster to all the data points in other clusters

a (i) is the average distance of the data in a cluster to all the other data points in the same cluster

The value of silhouette coefficient varies from -1 to 1, with -1 representing wrong clustering, 0 representing the same clustering i.e. no variation in clustering even if different distance measures are used and 1 representing the best clustering.

5. RESULTS AND ANALYSIS

The approach in this paper entails decreasing the smart meter dataset and extracting useful information from the reduced dimensions. The paper focuses on reducing the dimensionality of smart meter data obtained from the UMassTrace Repository. This data set tracks the energy consumption in 443 buildings over a 24-hour period .The sampling rate is 1 sample per minute. Figure 1 shows the energy consumption plot of a random user.

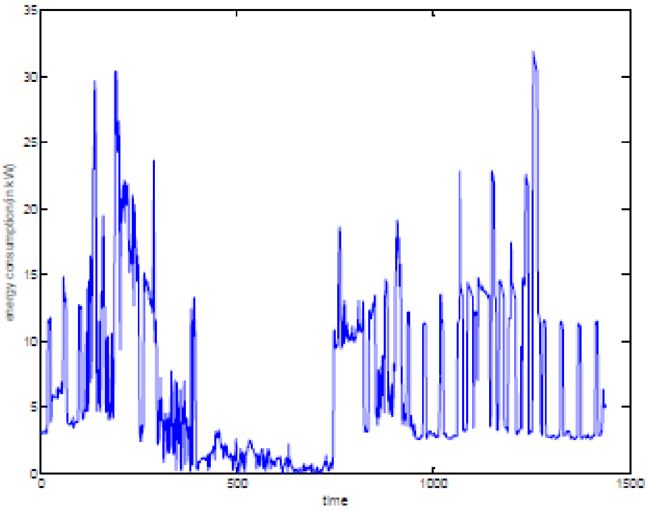

Fig-1: Energy consumption usage of a random consumer

The principal components range from 11400 to 1400 based on the Eigen vectors arranged in decreasing order. The test analyzes energy usage data using MATLAB. Figure 2 displays the cumulative variance obtained using PCA. The first PC preserves 93.6689% of the variance .At a PC of 350, 100% of the variance remains unchanged .This analysis suggests that a reduced dataset of matrix with 443 rows and 350 columns is sufficient for demand side management applications to give pricing information to individual consumers. This reduces the redundancy prior to clustering to gather important information. It is observed that PCA performs better in terms of accuracy and precision as the dimensionality reduction size increases.

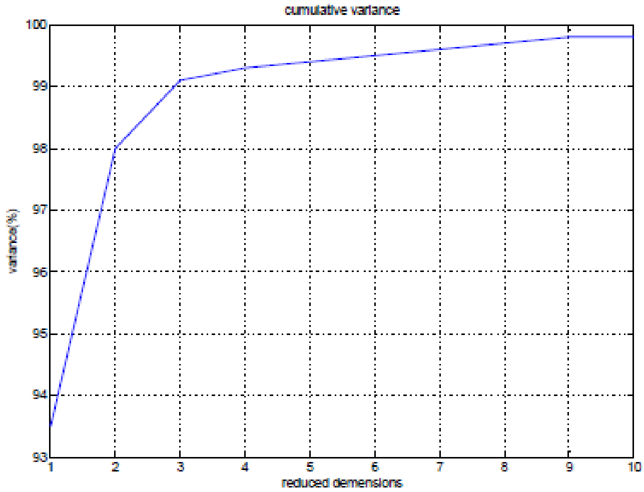

Fig-2: Cumulative variance of reduced dimensions

K –means technique is used for the clustering of consumer feature set, since clustering approaches use distance to calculate cluster sets, so high value characteristics are given higher weightage .To circumvent this, the data is standardized using min-max normalization techniques .Figure 3 and Figure 4 shows the plot of data before and after normalization of 52 consumer samples.

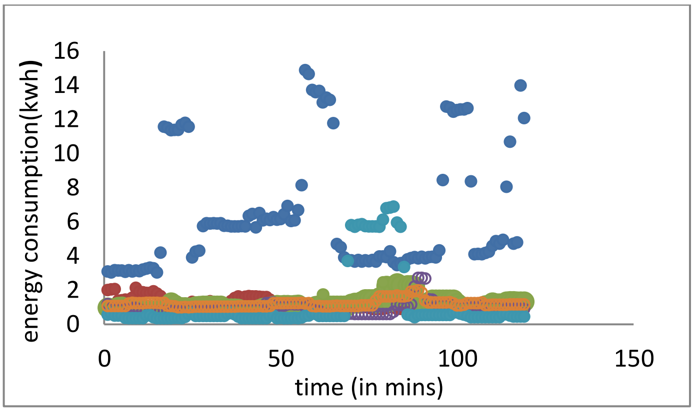

Fig-3: Scatter plot of data points before normalization

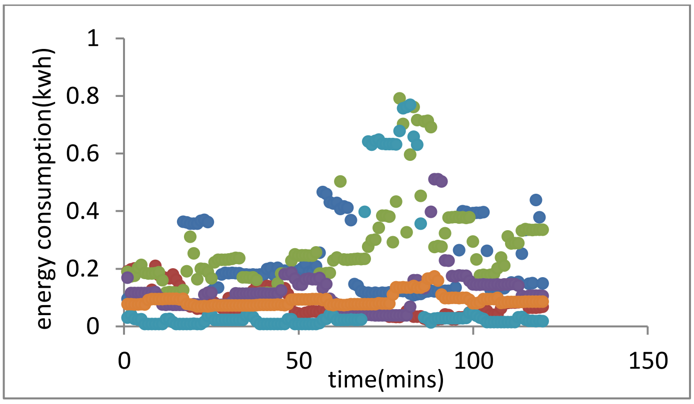

Fig-4: Scatter plot of data points after min- max normalization

Clustering is performed for k=7 using Euclidean, Manhattan and Pearson correlation distances. It is observed that same consumer profile is assigned to different cluster when the distance measure used for clustering is changed. To find out which distance measure results in better clustering of the consumer profile, an evaluation index, average silhouette coefficient is calculated using distance measures shown in table below-

Table-1: Average silhouette coefficient calculation

Distance	Without normalization	After normalization
Euclidean distance	0.3539	0.3760
Pearson correlation distance	0.3057	0.3546
Manhattan distance	0.2163	0.2422

It can be observed from the table that there is increase in the silhouette coefficients of the respective distance metrics after the normalization of data. Hence, results in better clustering. Also, the effect of outliers is decreased when data is normalized which increases the performance of PCA and clustering process and improved results are obtained. Figure 5, Figure 6 and Figure 7 represent the average silhouette coefficients for k=7 using Euclidean and Manhattan distance and Pearson correlation distance as distance metric respectively.

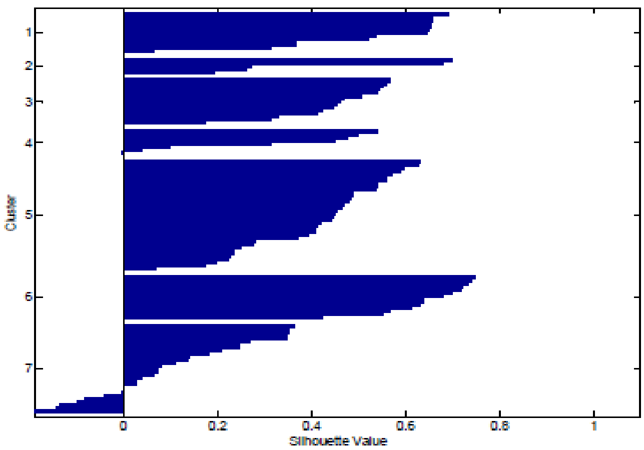

Fig-5: Average silhouette coefficient using Euclidean distance

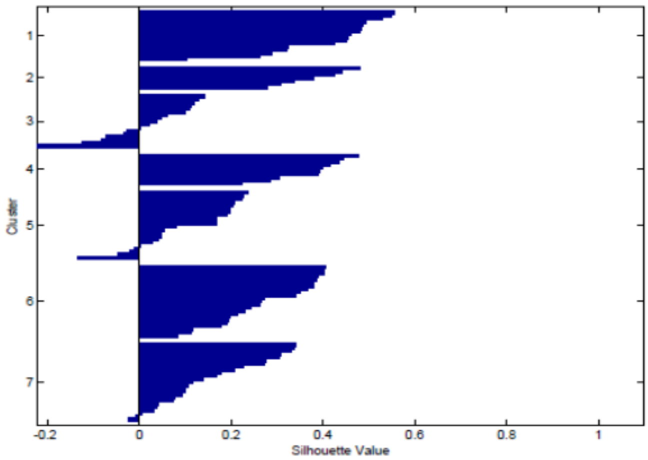

Fig-6: Average silhouette coefficient using Manhattan distance

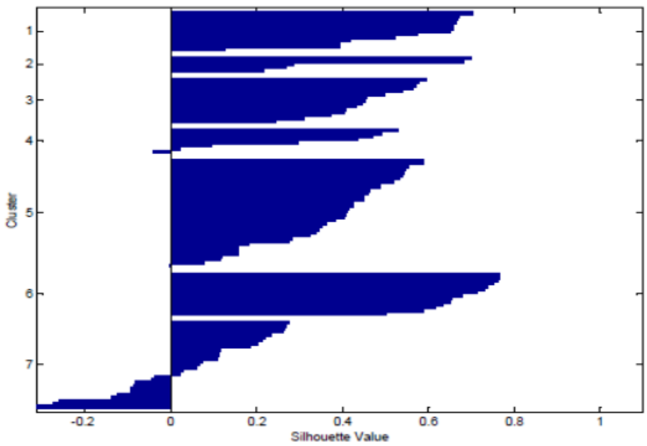

Fig-7: Average silhouette coefficient using Pearson correlation distance

The calculation of average silhouette coefficient for all the respective distance measures depicts an increase in the average silhouette coefficient after the normalization of data representing improved clustering. Higher the value of silhouette coefficient, higher is the intra-cluster similarity and inter-cluster dissimilarity. It can be inferred from the table that Euclidean distance shows the highest increase in the silhouette coefficient 0.3539 to 0.3760 while for Pearson correlation distance it increases from 0.3057 to 0.3546 and Manhattan showing the least average silhouette coefficient value from 0.2163 to 0.2422.

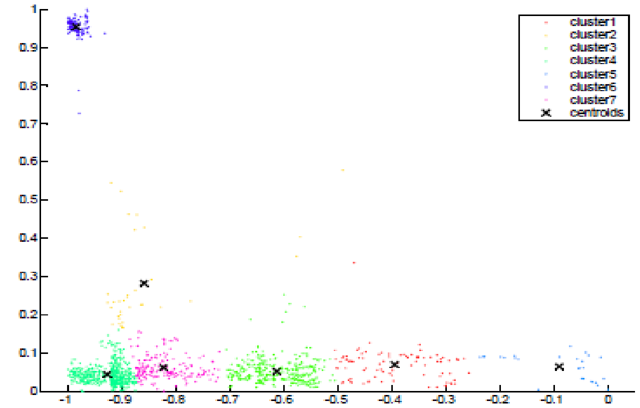

Fig-8: K-means clustering using Euclidean distance

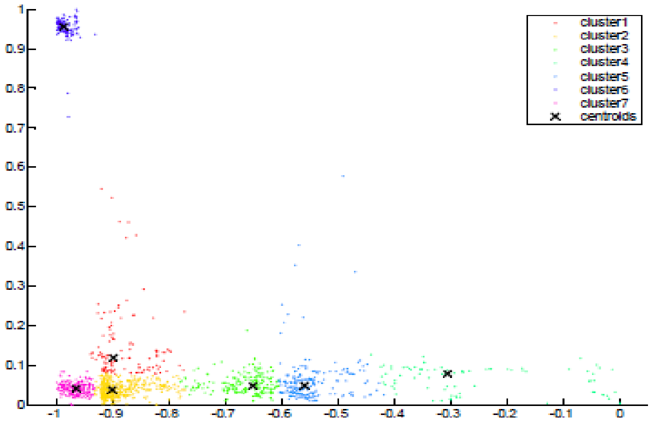

Fig-9: K-means clustering using Manhattan distance

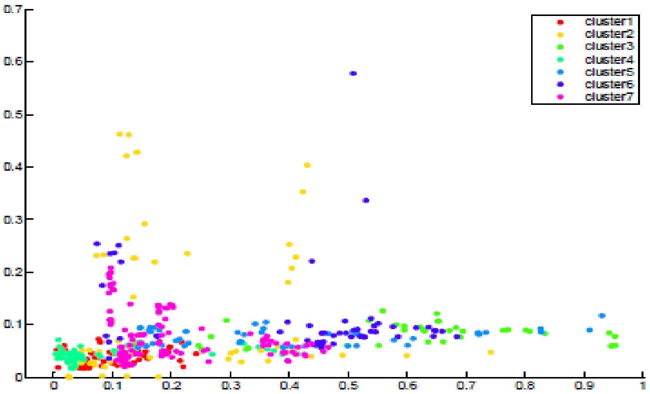

Fig-10: K- means clustering using Pearson correlation distance

To calculate the accuracy of the clustering the root mean square error is calculated using the silhouette coefficient obtained before and after normalization using both distance metric. RMSE and accuracy is found out to be 0.0624 and 93.75% for Euclidean, 0.15 and 85.47% for Pearson correlation and 0.119 and 82.02% for Manhattan distance respectively.

Table -2: Root mean square error calculation

Distance	Silhouette coefficient without normalization	Silhouette coefficient after normalization	RMSE	Accuracy
Euclidean distance	0.3539	0.3760	0.0624	93.75%
Pearson correlation distance	0.3057	0.3546	0.15	85.47%
Manhattan distance	0.2163	0.2422	0.119	88.02%

6. CONCLUSION

The available gap between demand and supply is expanding due to the increase of electrical equipment, resulting in an electricity deficit during peak hours. Demand side management techniques boost the possibility to capitalize on consumption fluctuation, lowering peak power demand. Shifting loads from peak to off-peak or turning off partial loads during peak hours can be problematic for some customers; however, leveraging their consumption patterns through clustering results in more flexible DSM techniques. In this study, we used PCA as a dimensionality reduction technique to reduce the dimensions of a smart meter dataset from UMassTraceRepository from 1440 to 350 while retaining all key information and the maximum variance of the data. The reduced dataset is clustered using k-means clustering (k=7), with Manhattan, Euclidean, and Pearson correlation distances used as distance metrics. However, clustering with the Manhattan distance as a distance metric enables robust clustering, particularly when data has a high dimensionality and the impact of outliers or extreme values must be minimized. The average silhouette coefficient serves as the clustering validation index. The Euclidean distance produces an average silhouette coefficient of 0.3760, an RMSE of 0.0624, and an accuracy of 93.75%, indicating that clustering using the Euclidean distance as distance metric results in better categorization of consumers based on the similarity of their typical electricity consumption behavior, better temporal feature extraction, and pattern identification of household consumption. Based on these findings, power suppliers can better understand their power consumers and target potential customers for effective and adaptable demand side management measures

REFERENCES

[1] “Smart meter, amart data smart growth - Google Search.”

[2]N. Jenkins, C. Long, and J. Wu, “An overview of the smart grid in Great Britain. Engineering, 1 (4), 413-421.” 2015.

[3] "Operating Britain's secure smart meter network | Smart DCC."

[4] I. K. Fodor, "A survey of dimension reduction techniques," Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States), 2002.

[5] R. Mena, M. Hennebel, Y.-F. Li, and E. Zio, "Self-adaptable hierarchical clustering analysis and differential evolution for optimal integration of renewable distributed generation," *Appl. Energy*, vol. 133, pp. 388–402, Nov. 2014, doi: 10.1016/j.apenergy.2014.07.086.

[6] F. Iglesias and W. Kastner, "Analysis of similarity measures in times series clustering for the discovery of building energy patterns," *Energies*, vol. 6, no. 2, pp. 579–597, 2013.

[7] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: a survey and empirical demonstration," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Edmonton Alberta Canada: ACM, Jul. 2002, pp. 102–111. doi: 10.1145/775047.775062.

[8] S. Ramos, V. Figueiredo, F. Rodrigues, R. Pinheiro, and Z. Vale, "Knowledge extraction from medium voltage load diagrams to support the definition of electrical tariffs," *Eng. Intell. Syst. Electr. Eng. Commun.* vol. 15, no. 3, pp. 143–149, 2007.

[9] R.-F.Chang and C.-N. Lu, "Load profiling and its applications in power market," in *2003 IEEE Power Engineering Society General Meeting (IEEE Cat. No. 03CH37491)*, IEEE, 2003, pp. 974–978.

[10] R. Fatima, D. Jorge, and F. Vera, "A comparative analysis of clustering algorithms applied to load profiling [C]," in *The Third International Conference on Machine Learning and Data Mining in Pattern Recognition, Leipzig, Germany*, 2003, pp. 73–85.

[11] A. Al-Wakeel, J. Wu, and N. Jenkins, "k -means based load estimation of domestic smart meter measurements," *Appl. Energy*, vol. 194, pp. 333–342, May 2017, doi: 10.1016/j.apenergy.2016.06.046.

[12] M. Martinez-Pabon, T. Eveleigh, and B. Tanju, "Smart meter data analytics for optimal customer selection in demand response programs," *Energy Procedia*, vol. 107, pp. 49–59, 2017.

[13] G. Shamim and M. Rihan, "Novel technique for feature computation and clustering of smart meter data," in *2019 International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, IEEE, 2019, pp. 1–5.

[14] H.-Â. Cao, C. Beckel, and T. Staake, "Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns," in *IECON 2013-39th annual conference of the IEEE industrial electronics society*, IEEE, 2013, pp. 4733–4738.

[15] E. I. G. Nassara, E. Grall-Maës, and M. Kharouf, "Linear discriminant analysis for large-scale data: Application on text and image data," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2016, pp. 961–964.

[16] I. Koch and K. Naito, "Dimension selection for feature selection and dimension reduction with principal and independent component analysis," *Neural Comput.*, vol. 19, no. 2, pp. 513–545, 2007.

[17] J. H. Ang, E. J. Teoh, C. H. Tan, K. C. Goh, and K. C. Tan, "Dimension reduction using evolutionary Support Vector Machines," in *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, Hong Kong, China: IEEE, Jun. 2008, pp. 3634–3641. doi: 10.1109/CEC.2008.4631290.

[18] Z. Yu *et al.*, "Incremental Semi-Supervised Clustering Ensemble for High Dimensional Data Clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 701–714, Mar. 2016, doi: 10.1109/TKDE.2015.2499200.

[19] A. Aleshinloye, A. Bais, and I. Al-Anbagi, "Performance analysis of dimensionality reduction techniques for demand side management," in *2017 IEEE Electrical Power and Energy Conference (EPEC)*, Saskatoon, SK: IEEE, Oct. 2017, pp. 1–6. doi: 10.1109/EPEC.2017.8286232.

[20] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.

[21] D. Hand, H. Mannila, and P. Smyth, "Principles of data mining. Massachusetts Institute of Technology," 2001.

[22] "Smart - UMass Trace Repository." Available: https://traces.cs.umass.edu/index.php/Smart/Smart

[23] H. Anton and C. Rorres, *Elementary linear algebra: applications version*. John Wiley & Sons, 2013.

[24] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson Correlation Coefficient," in *Noise Reduction in Speech Processing*, vol. 2, in Springer Topics in Signal Processing, vol. 2. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–4. doi: 10.1007/978-3-642-00296-0_5.