International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 16, No.1 (Jul-24)

DOI:http://dx.doi.org/10.12785/ijcds/160103

DIC2FBA: Distributed Incremental Clustering with Closeness Factor Based Algorithm for Analysis of Smart Meter Data

Archana Chaudhari[1], Preeti Mulay[2], Ayushi Agarwal[2], Krithika Iyer[2] and Saloni Sarbhai[2]

[1]*Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India*
[2]*Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India*

Received 29 Apr. 2023, Revised 26 Mar. 2024, Accepted 6 Apr. 2024, Published 1 Jul. 2024

Abstract: Due to increased civilization, smart cities, and the advent of technology, lots of buildings including commercials, residential, and other types are populating in numbers in the recent past. The electricity consumption is also affecting due to increased occupancy in these buildings. The analysis of the electricity consumption patterns will be helpful for consumers and electricity generation units to know about consumption and future requirements of electricity. As per the literature, the Incremental clustering algorithm is the best choice to handle ever-increasing data. In this research work, in the first phase, the electricity consumption data was extracted from smart meter images, and then in the second phase, the data was taken from extracted .csv files merging data from various sources together. This research proposes Distributed Incremental Clustering with Closeness Factor Based Algorithm (DIC2FBA) to update load patterns without overall daily load curve clustering. The proposed DIC2FBA has used Amazon Web Service(AWS) and Microsoft Azure HDInsight service. The AWS EC2 instance, along with the AWS S3 bucket and HdInsight, operates by clustering data from numerous sites using an iterative and incremental approach. The DIC2FBA first extracts load patterns from new data and then intergrades the existing load patterns with the new ones. Further, we have compared the findings achieved using the DIC2FBA with IK means and NFICA based on time, features, silhouette score, and the Davis Bouldin index, which indicate that our method can provide an efficient response for electricity consumption patterns analysis to end consumers via smart meters.

Keywords: Distributed Incremental Clustering, CFBA, Smart Meter Data Analysis, Microsoft Azure, AWS, CFBA

1. INTRODUCTION

Incremental Clustering is an effective unsupervised machine learning technique. Various algorithms can be employed to segregate data of a similar nature. An Incremental Clustering Algorithm is an uninterrupted process that assigns data points based on existing clusters formed [1]. It performs in a streaming fashion making the process efficient and optimized. Distributed Incremental Clustering Algorithm works one step ahead, making the entire system in a cloud environment. Working in a cloud environment provides many benefits, such as providing multiple virtual machines and storage space to make the business application collaborative.

Electricity Smart Meters (ESM) are electronic devices that monitor consumer energy usage at hourly or fewer intervals. The authors in the paper [2], [3] presented a bibliometric survey on ESM, which provides a brief overview of the current status of electricity smart meter data analysis using an incremental clustering approach and possible future work in this field. From the literature survey, it becomes clear that a large amount of ESM data gets accumulated over time. Therefore to get a proper analysis, it is efficient and easy to improve clustering results incrementally based on both old and new data rather than reclustering all data from scratch as new data arrives. Most of the data available for ESM are in Images, which becomes taxing to cluster based on visual representations.

India has installed over 1.3 million ESM in major cities in a few years [4]. It has become vital for us to manage electrical power consumption as its consumption has been increasing over time. Thus Smart meter analysis can help to minimize the cost based on usage patterns. Further, it can also predict usage based on past trends. Smart Meter Data Analysis (SMDA) has proved to be useful for electricity theft detection, which has been a major concern over the years.

Unsupervised learning helps identify the power consumption patterns of different consumers based on their lifestyle uses [5]. Nevertheless, it has also been found that the usage pattern of the same consumers also changes over the period. With the help of clustering algorithms, we can cluster the consumers with similar patterns, and the consumption of the same users can further be analysed to suggest efficient electricity usage.

This paper proposed a new approach to SMDA. Distributed Incremental Clustering with Closeness Factor Based Algo-

Corresponding Author: [1]*archanac.dit@gmail.com*
[2]*preetimulay45@gmail.com*

rithm (DIC2FBA) efficiently handles the incremental data and extracts the electricity consumption pattern of the consumer for effective use of the power system. The proposed system extracts the electricity load pattern of the consumer for effective use of the power system and reduces electricity consumption. The main contributions of this research are summarized as follows:

- Distributed Incremental Clustering with Closeness Factor Based Algorithm (DIC2FBA) is proposed for continuously updating load patterns based on ESM data.
- The closeness factor is used as a similarity measure to optimize the incremental clustering performance. Parameter updating is considered for performing incremental clustering continuously with an influx of new data.
- In the first phase, the electricity consumption data was extracted from the UFPR-AMR dataset [6], which consists of 2000 images from the Energy Company of Paraná, and then in the second phase, the data was taken from IIT Bombay Indian Residential Energy Dataset [7].
- Compared the findings achieved using the DIC2FBA with IK means [8] and NFICA [9] based on time, features, silhouette score, and Davis Bouldin index.

This paper is divided into four sections. Section 1, the introduction provides an overview of Distributed Clustering Algorithms and provides the current state of ESM Analysis. Section 2, which is related to work and research gaps, provides insight into work done until now in these areas. The third section proposed methodology gives an overview of the experiment performed using the DIC2FBA system. This is followed by section 4, which is the result evaluation and which discusses the comparisons of the proposed algorithm with K means and NFICA. Finally, we have conclusion sections and references at the end.

2. RELATED WORK

Technological advancements have now allowed each one of us to learn new skills at home or through various workshops conducted, and one of the ways to award your skill is by providing Certificates. Digital and Handwritten certificates have enormous data and provide numerous analysis. The problem with these types of datasets is they are usually in an unstructured format, mostly as images. But once the data is structured, we can use this information to provide analysis on which subject has recently gained popularity and how to improve the field of study at different universities.

In the literature, various unsupervised machine learning techniques, specifically clustering analysis is commonly used to categorize the pattern of the load, analyze residential electricity consumption data, and extract consumption patterns [3], [10]. In incremental K-means (IK means) clustering algorithm is applied to a dynamic database where the data may be frequently updated. And this approach measures the new cluster centers by directly computing the new data from the means of the existing clusters instead of rerunning the K-means algorithm [11].IK means produces k number of clusters effectively [8].

In [12], they developed a clustering method to analyze the pattern of electricity consumers, which supports targeted demand-side management and efficient operation of the intelligent grid. In [13], investigated suitable customers for demand response management and pattern recognition modeling. In [14] used, a hierarchical algorithm to cluster 27900 daily load curves of residential dwellings. The thirty cluster patterns are visible such as morning peak, mid-day peak, late night peak, night peak, and dual peak, and stability of load curve identified. In [15], the research used two levels of clustering, i.e., intra-building clustering and inter-building clustering. The intra-building clustering used a Gaussian mixture model-based clustering to identify the typical daily electricity usage profiles of each individual building. The inter-building clustering used an agglomerative hierarchical clustering to identify the typical daily electricity usage. In [16], the author used a Bayesian Information Criterion to select the number of clusters for the constrained-based Gaussian mixture model. However, the cited reference fails to explicitly disclose the analyses of the load-shedding patterns in both planned and unplanned scenarios. In [6] uses a hierarchical structure to help the negative impacts when various demand response plans act simultaneously on a similar system. The features considered for a hierarchical clustering are consumption value, time of utilization, and temperature to detect the points which are the most viable option for demand response program application. In [17] studies, various aspects of customer electricity consumption, like magnitude, duration, and variability, are used for the establishment of a demand response program. In [18] used, a model-based clustering algorithm to identify the potential of wet appliances (dishwashers, tumble dryers, and washing machines) and the willingness of customers to participate in the demand response program actively. Several National and International projects [19] have been carried out on pattern recognition and demand-side management. The accuracy of the clustering algorithm is influenced by the multi-dimensional of electricity consumption data. Therefore, various pre-clustering techniques have been investigated in pattern recognition, such as Discrete Fourier Transform [20] and Principle Component Analysis [21]. Much work on the potential of electricity consumption patterns has been carried out [22], [23], [24]. However, there are still some critical issues, like user input for a number of clusters is required. To remove the dependency of the user input, Preeti et al.[25] developed a parameter-free Closeness Factor-Based Algorithm(CFBA). It is further extended by TBCA (Threshold-based Clustering Algorithm) to analyze diabetic patient's clinical parameters [26] and diabetic Mellitus [27]. TBCA uses the threshold value 0 to 1 only. To further enhance the threshold range from -1 to +1 Correlation-Based Incremental Clustering Algorithm

(CBICA) [28], a new variant of CFBA has taken shape. CBICA uses Pearson's coefficient of correlation similarity measures. Their approach is not well suited for a distributed system. Archana et al.[29] developed a real-time data analysis using Cassandra and Spark. In the paper [30] proposed a Log Likelihood-based Gradational Clustering Algorithm to identify the consumption patterns of the consumer. However, the cited reference suffers from the order sensitivity issue. It has now been hypothesized that an incremental clustering approach is an essential way to overcome the issue related to clustering with growing intelligent electricity meter datasets [9], [31]. The author [32], [2] proposed an android application named 'PowerStats', which gives the statistics of mobile phone charging patterns of users as per the model of phones, Plugged in/out battery percentage, Plugged in/out timestamp, Voltage & Current.

A. *Research Gaps*

With the extensive literature survey on the smart meter data analysis, the identified research gaps were highlighted as follows:

- he incremental clustering algorithm can be effectively applied in the field of load profiling for the reduction of electricity consumption
- The incremental learning via incremental clustering is achieved to utilize updated and clustered smart meter data as knowledge for further mining.
- Given the volume of data and the number of data types involved makes, smart meter data analytics are highly complex. Incremental learning of Smart Electricity Meter data is not made explicit in literature, which blurs their conceptual contours and constrains the efficacy of using the approaches in research and practice

3. PROPOSED METHODOLOGY: DIC2FBA SYSTEM

The proposed Distributed Incremental Clustering with Closeness Factor Based Algorithm (DIC2FBA) has been implemented on Amazon AWS cloud. The proposed system efficiently handles the influx of new smart meter data and extracts the required knowledge or hidden insights to improve energy management for both utility providers and customers. The DIC2FBA relates the new scenarios with the previous ones (empirical data), remembers the outcomes, and considers the impacts caused by the learning.
The development tasks and evaluation of the result divide the whole process into different phases. Figure 1 illustrates the system architecture of the proposed system comprises a collection of smart meter data from different geographical locations, preprocessing to normalize data, resulting in patterns of loads of a typical day followed by the cluster analysis to extract hidden patterns of electricity consumption.

A. *Data Preprocessing*

Most people think that your insights and analyses are only as good as the data you're using while working with data. In other words, if you put garbage data in, you'll get trash analysis out. If you want to build a culture around quality data decision-making, data cleaning, also known as data cleansing and data scrubbing, is one of the most crucial tasks for your organisation to take. In this research work, in the first phase, the electricity consumption data was extracted from smart meter images, and then in the second phase, the data was taken from extracted .csv files merging data from various sources together. The resultant csv datasets needed to be cleaned to achieve the best results. The following steps are performed to perform data preprocessing:

- Remove unnecessary variables: The data feature some inconsistencies, such as reading errors and outliers. So,we removed those customers whose data were not reliable.
- The null values which were present in the Smart Meter Dataset were removed.
- Removing the repeated customer ID so all data points are now identified based on their ID number

B. *Feature Extraction*

Feature extraction is a process in which prominent features from the dataset are extracted for further processing. The first and foremost step was to extract the features from the smart meter dataset. The features from the dataset were extracted using PCA [30]. The extracted feature are ID, hHour, kWh, with customer number, timestamp, and electricity consumed. In addition, we calculated peak and off-peak electricity percentage consumption.

C. *Exploratory Data Analysis (EDA)*

With the use of summary statistics and graphical representations, exploratory data analysis refers to the crucial process of doing first investigations on data to uncover patterns, uncover anomalies, test hypotheses, and check assumptions.
The correlation matrix of the smart meter dataset was calculated. Correlation is a type of statistical relationship that involves reliance. It refers to the degree to which two variables have a linear relationship with one another, as measured by correlation coefficients. corr() method in python was used to calculate the correlation matrix. Figure 2 shows the correlation matrix of the Smart Meter Dataset and the function used to implement the same. As we can see in figure 2, the correlation coefficient for the cells [1,1], [2,2], and [3,3] of the matrix is 1 because the features are the same. Also, the correlation coefficient for 'TS' and 'V1' is less than the coefficient for 'TS' and 'W1', which means 'TS' and 'V1' are less related or similar than 'TS' and 'W1'.

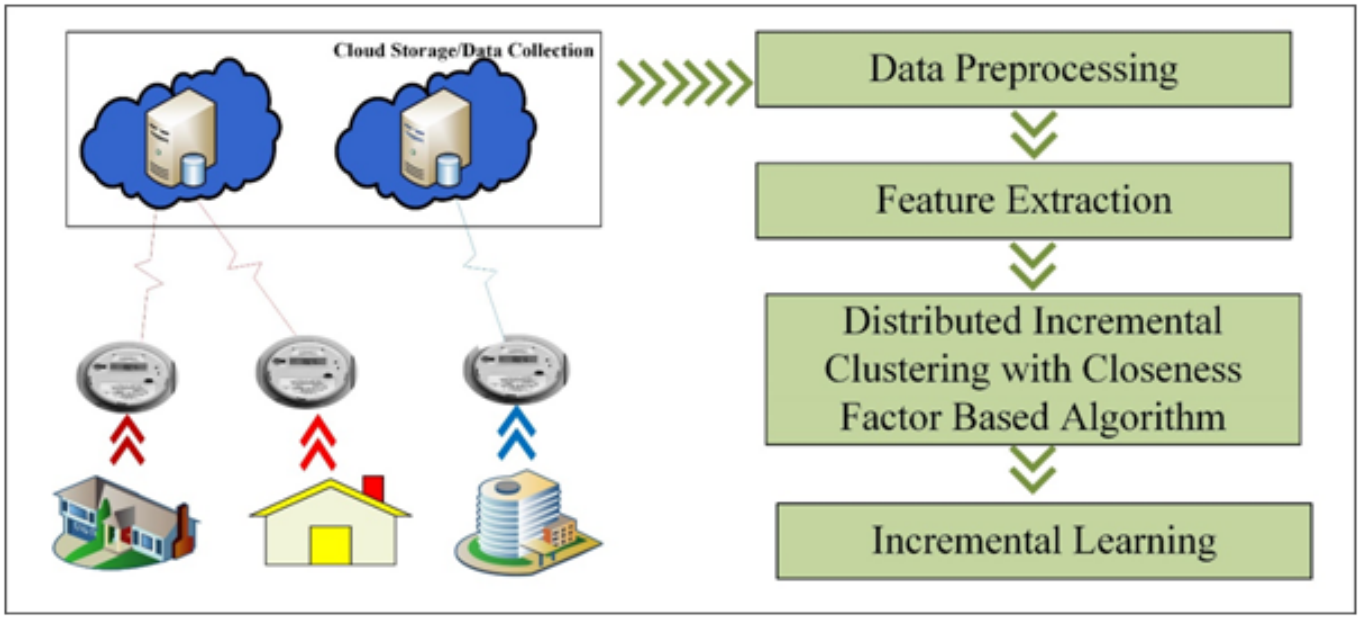

Figure 1. The proposed DIC2FBA System

```
corrMatrix = df.corr()
sn.heatmap(corrMatrix, annot=True)
plt.show()
```

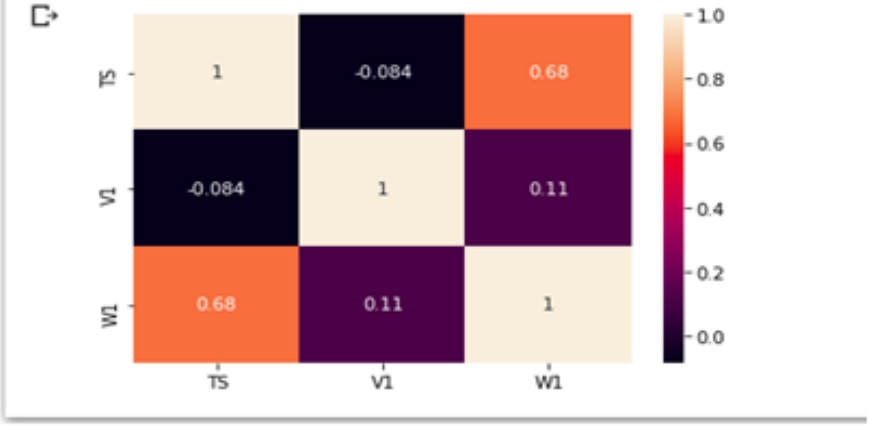

Figure 2. Correlation Matrix of ESM Dataset

D. Data Transformation

Data transformation is a data mining approach that entails converting raw data into a format that can be understood by our machine learning algorithm. The Smart Meter dataset was transformed using the Min-Max Scaling technique. In this approach, the values from the dataset are converted in a range of 0 to 1. Because of this conversion, the dataset will have smaller standard deviations hence decreasing the outliers.

E. Proposed DIC2FBA Algorithm

The proposed Algorithm 1 will be an enabler for electricity management in empowering consumers to save and manage their electricity consumption.

F. Proposed DIC2FBA Algorithm on Amazon Web Services

After data preprocessing, the data was ready for the algorithms to be applied. The proposed Distributed Incremental Clustering with Closeness Factor Based Algorithm applied to the smart meter dataset. Clustering is based on

Algorithm 1 Distributed Incremental Clustering with Closeness Factor Based Algorithm (DIC2FBA)

- **Input:** Smart meters datasets from consumer
- **Output:** A series of the cluster, which represents the electricity consumption patterns
- **Method:**
  1) Preprocessed data has been processed by computing the sum of attributes, log-likelihood, error, and weight. Using these computed values, the basic clusters are formed
     a) Consider two data series S_1 and S_2. $S_i(j)$ is the point i in series j.
     b) Calculate the total of the corresponding parameters of series considered $T_i(j)$, where $i = 1, 2, \cdots, m$ number of attribute and $j = 1, 2, 3, \cdots, n$ number of instances in the dataset
     c) calculate the probability ratio(P) [1] $P = \frac{\sum_{j=1}^{n} S_1(j)}{\sum_{j=1}^{n} T(j)}$
     d) Compute error $err(j) = \frac{P*T(j)-S_i(j)}{\sqrt{T(j)*p*(1-p)}}$
     e) Weight of each series $w(j) = \sqrt{(T(j))}$
     f) Closeness of the series is calculated as $Cr(j) = \sum_{j=1}^{n} \frac{err(j)^2*w(j)}{w(j)}$
     g) Repeat steps a to f until all the series have been processed.
  2) CreateClusters()
     a) Create clusters using the closeness method
     b) For all the series "i" to "n",
        i) Get the Cr value of S(i)
        ii) The lower the value of Cr means closer are the series and grouped in a cluster.
     c) In this way, basic clusters formed and stored into a cluster database along with their elements.
  3) UpdateClusers()
     a) For each existing clusters,
        i) Get each Series in this cluster, S(i)
        ii) Get the g value of S(i)
        iii) For each newly added series
           A) Get the Series S(j)
           B) Get the C_r value of S(j)
           C) If (S(i) – S(j) ≤ closeness-factor) : Add S(j) to cluster
           D) Continue to the next Series
        iv) For all the incremental series: Follow steps in 3(a)iii
  4) The obtained clusters are written in the output file.
- C_r values guide users on which data series are close to each other and can be a part of the same cluster.

Closeness Factor (CF) values. The CF value 1 indicates that data series are comparable, whereas 0 suggests that they are dissimilar. First, the basic clusters are formed in the first iteration, and in the second iteration, some incremental clusters are formed above the basic clusters.

One of the salient features of Distributed Clustering Algorithms is to cluster data coming from multiple sites. When applying a problem statement in a real-world scenario, data is bound to come from multiple sites, and reclustering information based on initial data becomes complex, time-consuming, and inefficient. Thus, we have applied the model developed in the local machine to support data from multiple sites without reclustering the initially clustered data.

To achieve the above stated, we have used AWS S3 bucket and EC2 instance. Amazon provides public cloud storage called 'Simple Storage Service' (S3), which can store enormous amounts of data and can be accessed from anywhere and everywhere as describe in Algorithm 2.

Elastic Cloud by Amazon helps us launch virtual machines at different locations with the desired specification for testing the model. Figure 3 demonstrates an AWS based DIC2FBA implementation. We have launched two virtual machines with the following additional configuration and the rest set to default:

- Platform: Ubuntu
- Availability Zone: ap-south-1a
- Security Group: Full-Access

G. DIC2FBA Algorithm on Azure HDInsight

Azure HDInsight is a customizable, enterprise-grade service for open-source analytics. It helps to distribute the DIC2FBA with the global scale of Azure, as shown in figure 4.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The results of the DIC2FBA algorithm on smart meter datasets are shown in this section.

A. Dataset Description

In the first phase, the electricity consumption data was extracted from the UFPR-AMR dataset, which only contained readings and no other features to be extracted. We used Opencv and Pytesseract to extract the reading, but we discovered that the proposed incremental clustering algorithms couldn't be run on just one function. To compare electricity usage, we need either time and voltage or power and voltage. As a result, we combined the UFPR-AMR dataset with IIT Bombay Indian Residential Energy Dataset [7]. The final dataset contains Unix TimeStamp, Voltage, and Active Power, as shown in Table 1. The dataset contains electricity consumption data from a high-rise residential building on the IIT Bombay campus from December 2016 to January 2018.

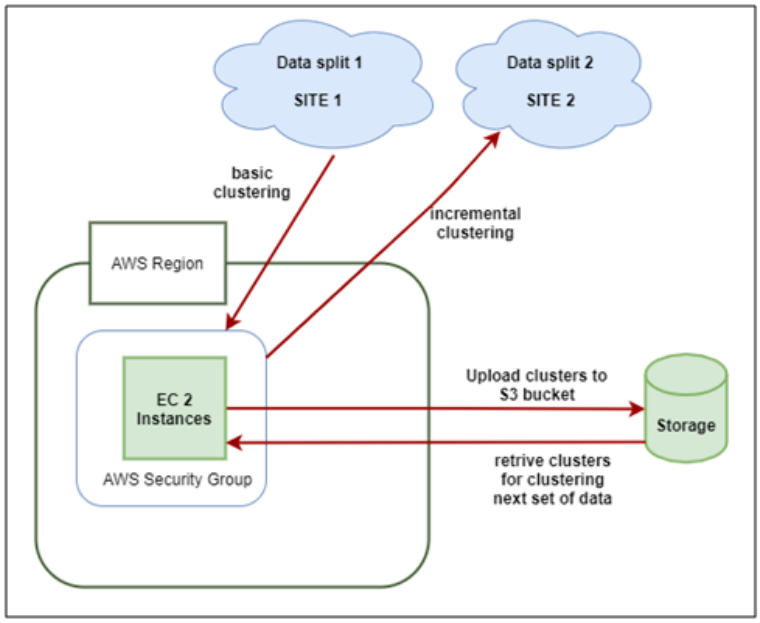

Figure 3. Proposed AWS based DIC2FBA

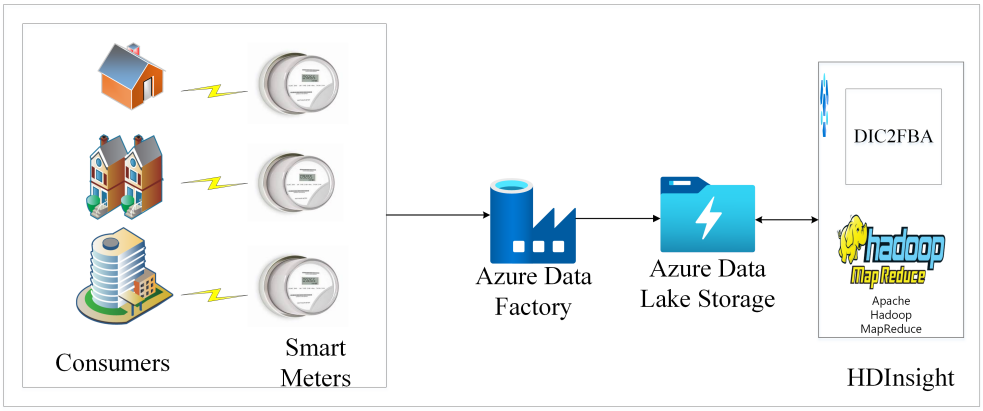

Figure 4. DIC2FBA Algorithm on Azure HDInsight

TABLE I. Description of Indian Residential Energy Dataset

Variables	Description
Country	India
No. of customers	60
Monitoring period	December 2016 to June 2018
Resolution	1 hour
No. of Instances	1080000
Building Type	Residential
Attributes	TS – Unix Time stamp (epochs), V1 – Voltage of phase 1 (V) V2 – Voltage of phase 2 (V), V3 – Voltage of phase 3 (V), W1 – Electricity consumption of phase 1 (Wh), W2 – Electricity consumption of phase 2 (Wh), W3 – Electricity consumption of phase 3 (Wh)

B. DIC2FBA Clustering Result

The results of the DIC2FBA on Indian residential smart electricity meter datasets are explain in this section.
This research discovered the relationship between different parameters and used cluster analysis to determine the parameters that are the underlying causes of an outlier in a data sequence.
With the help of the clusters formed an exploratory data analysis as shown in figure 5, we can analyze a few insightful details:

- All the clusters formed in the incremental phase are appended to previously formed clusters in the basic phase, with most of the reading belonging to cluster 1.
- The graphs represent the range of electricity con-

Algorithm 2 DIC2FBA on AWS

1) Connect S3 bucket to Closeness Factor Based Algorithm(CFBA) code
   a) Create an S3 bucket on AWS
   b) Install boto3 library at your local machine to connect to AWS CLI
   c) Load the boto3 library to connect to your AWS Account using the access key and access secret passcode
   d) Pass the necessary intermediate results to the boto3 client API $client_s3 = boto3.client('s3', aws_access_key_id = access_key, aws_secret_access_key = access_secret)$ where s3 is the AWS cloud storage service we intend to use, $aws_access_key_id$is the access key of the AWS console and $aws_secret_access_key$ is the secret passcode of the AWS console.
2) The intermediate results are successfully uploaded to the AWS cloud storage and thus can be accessed from different sites to perform clustering
3) Perform distributed clustering using EC2 virtual instance
   a) Launch EC2 Instance A (Site 1) and transfer the files using Winscp for conducting basic clustering operations on one chunk of the dataset.
   b) Now, after completion of initial clustering, upload the clustered data into the S3 bucket so that data from other sites can cluster based on formed clusters. With this, we completed Iteration 1.
   c) Launch EC2 Instance B (Site 2) and transfer the files using Winscp for conducting Iteration 2 with the second chunk of the dataset.
   d) After running the file, the new dataset from site 2 has appended to the initially clustered data from site 1.

sumption, here for example, we can divide the consumption range into 4 categories based on the reading range.

- After acquiring knowledge about the electricity consumption categories, we can identify the value of electricity used based on the voltage and watt information of each cluster. Since the data is vast, an in-depth analysis is required to determine values of high and low electricity consumption.

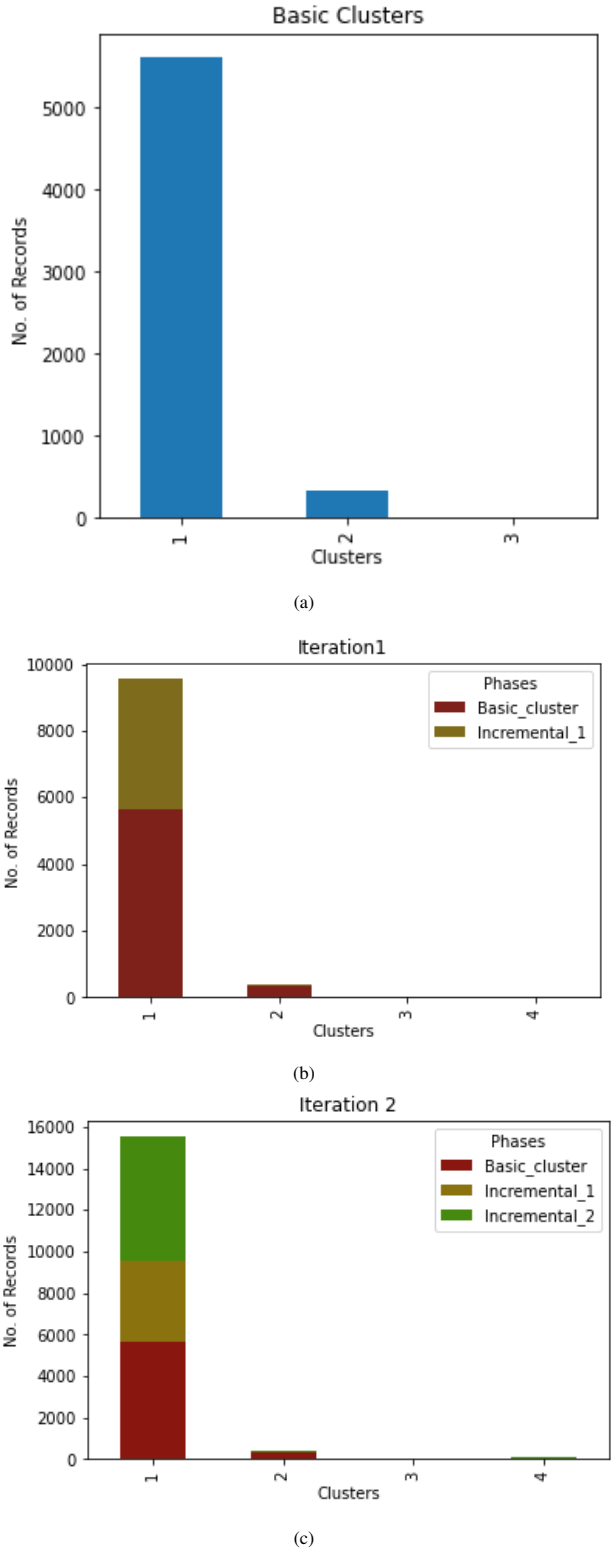

Figure 5. DIC2FBA Clustering Result: a :Describes the basic clusters; b :Formed new cluster; c :Update the existing clusters

C. Result Analysis and Performance Evaluation

The cluster-wise consumption patterns of the residential customer are depicted in figure 6. It indicates that cluster 4 contains the consumers who have consumed the highest electricity.

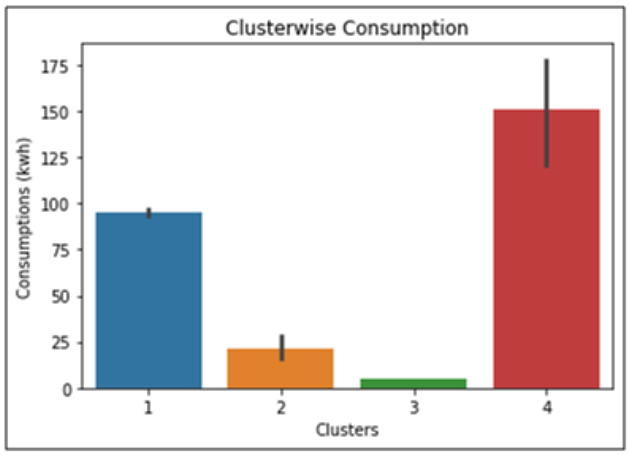

Figure 6. Cluster wise Consumption

As shown in Table II we can see that, from a total of 15986K smart meter data, 5966K were used for basic clustering and 3978K for iteration 1, and reaming were used for iteration 2.

TABLE II. Cluster Label for Electricity Smart Meter

	Basic Clusters	Iteration1	Iteration2
Frequency	5966K	3978K	6042K
Percentage of data	37.3	24.9	37.8
Valid percent of data	37.3	24.9	37.8
Cumulative percent of data	37.3	62.2	100.0

1) Cluster Validation

Parameters such as the no. of features and execution time also play a vital role in deciding which algorithm is much more beneficial. The prerequisite for true comparison is to make sure that every algorithm with the same data is evaluated in the same way, which means that the comparison metric should be similar for all the considered algorithms.

We have used the validation technique of the Dunn validity index (DVI) and Silhouette Criterion (SC) [14]. It's used to figure out how far apart the resulting clusters are. If the score is close to 0, the sample is close to the decision boundary between two neighbouring clusters, and if the score is negative, the samples are allocated to the incorrect clusters. Furthermore, DVI and SC score maximum indicates that the clusters are highly dense and do not overlap. Table III shows the SC and DVI scores of IK means [8] and NFICA [9], and the proposed DIC2FBA algorithm when performed on smart meter datasets.

TABLE III. Cluster Performance comparisons with IKmeans, NFICA and DIC2FBA

Algorithm	SC^+	DVI^+
DIC2FBA	0.912350252853	0.1090
IKmean	0.672232701595	0.0996
NFICA	0.32654879454	0.0998

+: maximum is the best

It is now clear that DIC2FBA outperforms IK means algorithm as DIC2FBA incremental behavior allows us to append the new records into the initially formed clusters without clustering from scratch as in the case of IK means. On the contrary, DIC2FBA does not have any additional dependencies or parameters which hamper the accuracy of clusters. It is also evident that the deployment of Distributed incremental DIC2FBA on Amazon Web Services EC2 Instance and S3 bucket and Microsoft Azure HDInsight service will speed up the executions.

5. CONCLUSION AND FUTURE WORK

Accurate and incremental electricity consumption patterns of residential ESM consumption users are calculated by using the proposed Distributed Incremental Clustering with Closeness Factor Based Algorithm (DIC2FBA). The use of AWS EC2 Instance and S3 bucket to ensure that our model can handle data from a variety of sources and create clusters incrementally.The incremental learning of the proposed DIC2FBA system will benefit:

- For starters, the incrementally shaped clusters on the smart meter electricity dataset would aid in the continuous detection of household electricity consumption.
- This will aid the concerned authorities in determining the maximum and minimum electricity usage and the time and location of the region associated with the above observation. This will allow the electricity department to take appropriate steps in certain circumstances, such as determining when and where to switch off electricity for a specific period in order to ensure its long-term use, and so on.

The researchers would like to improve the DIC2FBA for solving image clustering problems in the future.

References

[1] P. Mulay and P. A. Kulkarni, "Knowledge augmentation via incremental clustering: new technology for effective knowledge management," *International Journal Business Information Systems*, vol. 12, no. 1, pp. 68–87, 2013.

[2] S. Kuralkar, P. Mulay, and A. Chaudhari, "Smart energy meter: applications, bibliometric reviews and future research directions," *Science & Technology Libraries*, vol. 39, no. 2, pp. 165–188, 2020.

[3] A. Chaudhari and P. Mulay, "A bibliometric survey on incremental clustering algorithm for electricity smart meter data analysis," *Iran Journal of Computer Science*, vol. 2, no. 4, pp. 197–206, Dec 2019. [Online]. Available: https://doi.org/10.1007/s42044-019-00043-0

[4] L. A. Elrefaei, A. Bajaber, S. Natheir, N. AbuSanab, and M. Bazi, "Automatic electricity meter reading based on image processing," in *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*. IEEE, 2015, pp. 1–5.

[5] C. Ibrahim, I. Mougharbel, N. Abou Daher, H. Y. Kanaan, M. Saad, and S. Georges, "An optimal approach for offering multiple demand response programs over a power distribution network," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2018, pp. 95–102.

[6] R. Laroca, V. Barroso, M. A. Diniz, G. R. Gonçalves, W. R. Schwartz, and D. Menotti, "Convolutional neural networks for automatic meter reading," *Journal of Electronic Imaging*, vol. 28, no. 1, pp. 013 023–013 023, 2019.

[7] P. M. Mammen, H. Kumar, K. Ramamritham, and H. Rashid, "Want to reduce energy consumption, whom should we call?" in *Proceedings of the Ninth International Conference on Future Energy Systems*, 2018, pp. 12–20.

[8] R. K. Prasad, R. Sarmah, and S. Chakraborty, "Incremental k-means method," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2019, pp. 38–46.

[9] A. Y. Chaudhari and P. Mulay, "Cloud4nfica-nearness factor-based incremental clustering algorithm using microsoft azure for the analysis of intelligent meter data," in *Research Anthology on Smart Grid and Microgrid Development*. IGI Global, 2022, pp. 423–442.

[10] A. Cominola, E. S. Spang, M. Giuliani, A. Castelletti, J. R. Lund, and F. J. Loge, "Segmentation analysis of residential water-electricity demand for customized demand-side management programs," *Journal of cleaner production*, vol. 172, pp. 1607–1619, 2018.

[11] S. Chakraborty and N. Nagwani, "Analysis and study of incremental k-means clustering algorithm," in *International Conference on High Performance Architecture and Grid Computing*. Springer, 2011, pp. 338–341.

[12] H. S. Boudet, J. A. Flora, and K. C. Armel, "Clustering household energy-saving behaviours by behavioural attribute," *Energy Policy*, vol. 92, pp. 444–454, 2016.

[13] S. Ryu, H. Kim, D. Oh, and J. No, "Customer load pattern analysis using clustering techniques," *KEPCO Journal on Electric Power and Energy*, vol. 2, no. 1, pp. 61–69, 2016.

[14] A. Rajabi, M. Eskandari, M. J. Ghadi, S. Ghavidel, L. Li, J. Zhang, and P. Siano, "A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications," *Energy and Buildings*, vol. 203, p. 109455, 2019.

[15] J. Li and A. Nehorai, "Gaussian mixture learning via adaptive hierarchical clustering," *Signal Processing*, vol. 150, pp. 116–121, 2018.

[16] F. N. Melzi, A. Same, M. H. Zayani, and L. Oukhellou, "A dedicated mixture model for clustering smart meter data: identification and analysis of electricity consumption behaviors," *Energies*, vol. 10, no. 10, p. 1446, 2017.

[17] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 420–430, 2014.

[18] W. Labeeuw, J. Stragier, and G. Deconinck, "Potential of active demand reduction with residential wet appliances: A case study for belgium," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 315–323, 2014.

[19] DR-BOB, "The DR-BOB collaborative project," Web Page, 25.

[20] S. Zhong and K.-S. Tam, "Hierarchical classification of load profiles based on their characteristic attributes in frequency domain," *IEEE Transactions on Power Systems*, vol. 30, no. 5, pp. 2434–2441, 2014.

[21] R. Mehra, N. Bhatt, F. Kazi, and N. Singh, "Analysis of pca based compression and denoising of smart grid data under normal and fault conditions," in *2013 IEEE International Conference on Electronics, Computing and Communication Technologies*. IEEE, 2013, pp. 1–6.

[22] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: A new data clustering algorithm and its applications," *Data mining and knowledge discovery*, vol. 1, pp. 141–182, 1997.

[23] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artificial intelligence*, vol. 40, no. 1-3, pp. 11–61, 1989.

[24] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[25] P. A. Kulkarni and P. Mulay, "Evolve systems using incremental clustering approach," *Evolving Systems*, vol. 4, pp. 71–85, 2013.

[26] P. Mulay, "Threshold computation to discover cluster structure, a new approach." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 6, no. 1, 2016.

[27] P. Mulay, R. Joshi, and A. Chaudhari, "Mapping of six sigma to threshold based incremental clustering algorithm," in *2018 IEEE Punecon*. IEEE, 2018, pp. 1–8.

[28] K. Shinde and P. Mulay, "Cbica: Correlation based incremental clustering algorithm, a new approach," in *2017 2nd international conference for convergence in technology (I2CT)*. IEEE, 2017, pp. 291–296.

[29] A. A. Chaudhari and P. Mulay, "Scsi: real-time data analysis with cassandra and spark," *Big Data Processing Using Spark in Cloud*, pp. 237–264, 2019.

[30] A. Chaudhari and P. Mulay, "Algorithmic analysis of intelligent electricity meter data for reduction of energy consumption and

carbon emission," *The Electricity Journal*, vol. 32, no. 10, p. 106674, 2019.

[31] A. Chaudhari, R. R. Joshi, P. Mulay, K. Kotecha, and P. Kulkarni, "Bibliometric survey on incremental clustering algorithms," *Library Philosophy and Practice*, vol. 2019, pp. 1–24, 2019.

[32] S. Kuralkar and A. Mulay, Preeti andChaudhari, "Mobile phone charging: Power statistics and energy consumption pattern analysis using developed powerstats android application," *International Journal of Modern Agriculture*, vol. 9, pp. 1682 – 1710, 2020.

Dr. Archana Chaudhari is currently working in Dr. D. Y. Patil Institute of Technology, Pimpri, Pune. She received PhD degree in Engineering from Symbiosis International (Deemed University), Pune, India. She has obtained her B.E. degree in Computer Engineering from North Maharashtra University, Jalgoan, India, and the M.E. degree in Computer Engineering from Savitribai Phule Pune University, Pune, India. Her current research interest smart meter data analysis, outage management, machine learning, incremental learning, data analysis, Energy. She received grants from "Sakal India Foundation's" and Microsoft Azure for "AI for Earth" project. She Got the first prize in Best Poster Award at KPIT-IISER, Pune Energy and Mobility PhD Conference in 2022. She Published research papers in Elsevier, Springer, InderScience, Emerald and IEEE Xplore. Also she has published an Indian patent and an International patent. She has invited and participated in Microsoft AI for Earth Education Summit and Hackathon held at Microsoft, Redmond, Washington, USA during 14th May 2019 to 16th May 2019 sponsored by Microsoft.

Dr. Preeti Mulay is currently associated with the Symbiosis International (Deemed University), Pune, India. She has authored or coauthored more than 50 technical papers in journal and conference. Her areas of interest include machine learning, data mining, software engineering and knowledge augmentation. Dr. Preeti Mulay recognized as IEEE reviewer, Certified Springer Reviewer. She is reviewer of International Journals and conference. She received Microsoft Azure Research Grant for the integrated research related to Machine Learning and Diabetes study, in Feb, 2018, in the areas of Data Analytics. Extended the Microsoft Azure Research Grant for the integrated research related to Machine Learning and Smart-meter data analysis study, from May 2018-May 2020, in the areas of AI for Earth. She Published research papers in ACM, Elsevier, Springer, InderScience and IEEE Xplore.

Ms. Ayushi Agarwal is software developer for Tibco in India for the last one year. In addition, she has enrol in University College Dublin's Master of Science in Computer Science programme this September. She completed her internship with Cummins India Limited as a Software Developer while earning her Bachelor of Science in Computer Science from Symbiosis Institute of Technology, Pune, India.

Ms. krithika Iyer currently pursuing masters degree in computer science at Northeastern University, Boston, US. She is also working on full time internship as a Software Developer/Data Analyst at Ellington Management Group LLC, old Greenwich, CT,USA. Prior to this she has completed undergraduate in Bachelors of Technology Computer Science from Symbiosis Institute of Technology Pune, India. During her undergraduate course she also had an opportunity to work as a full time mobile application developer for seven months at Bajaj Finserv Health, Pune, India.

Ms. Saloni Sarbhai is currently working as a Business Analyst at miniOrange Inc India. Prior to this she has completed undergraduate in Bachelors of Technology Information Technology from Symbiosis Institute of Technology Pune, India. During her undergraduate course she also had an opportunity to work as a full time Research and Development intern for six months at Indian Institute of Technology, Bombay (IIT) and an Android development intern at Ketiot Pvt Ltd.