Received 12 October 2020; revised 9 November 2020; accepted 11 November 2020. Date of publication 17 November 2020;
date of current version 10 December 2020.

Digital Object Identifier 10.1109/OJCOMS.2020.3038590

Fast Big Data Analytics for Smart Meter Data

MORTEZA MOHAJERI[1], ABOLFAZL GHASSEMI[2] (Senior Member, IEEE),
AND T. AARON GULLIVER[2] (Senior Member, IEEE)

[1]School of Electrical and Computer Engineering, University of Tehran, Tehran 1417466191, Iran

[2]Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8W 2Y2, Canada

CORRESPONDING AUTHOR: A. GHASSEMI (e-mail: aghassem@uvic.ca)

ABSTRACT A polar projection-based algorithm is proposed to reduce the computational complexity associated with dimension reduction in unsupervised learning. This algorithm employs K-means clustering. A new distance metric is developed to account for peak consumption in cluster consumer load profiles. It is used to cluster the load profiles according to both total and peak consumption. To accelerate the clustering process, a stochastic-based approach is developed to reduce the search space to find the cluster centers. Numerical results are presented which show a significant reduction in computational complexity using both polar-based and stochastic-based clustering compared to conventional approaches. Further, the estimation error is low.

INDEX TERMS Clustering, computational complexity, dimension reduction, smart grid.

I. INTRODUCTION

ADVANCED metering infrastructure (AMI), known as smart metering, is a key component of the smart grid that intelligently connects utility operators to the consumer and distribution domains. AMI typically provides two major types of services [1]–[3]. The first provides consumers with consumption data, critical peak pricing, outage reports, and remote meter management. Further, it enables utilities to record operational and security events as well as collect power quality and load profile (data over a particular time period). The second service type is to provide information such as alerts and control settings for actuators and sensors located within power distribution networks to support advanced distribution automation.

The main AMI components are smart meters installed on the consumer side and the metering data management system (MDMS) located in the utility operations center. A smart meter provides energy measurement, monitoring and control capabilities. It compiles real-time consumption data and transmits this to the MDMS. As a data repository, the MDMS manages and processes energy consumption, fault detection, and power restoration data. It uses tools such as *big data analytics* to facilitate the interaction with other operation and management systems.

The large number of smart meters deployed can generate a tremendous amount of data [1]. For example, if a smart meter records data every half hour, the number of records generated by twenty million smart meters is 3.5×10^{11} over just one year [5]. Further, the diversity and nature of smart meter data (e.g., consumption data, fault and outage information, and price information), results in heterogeneous multi-source data [4]. Thus, smart meter data has the characteristics of big data such as large volume and high heterogeneity. In addition, smart meter data must be quickly processed and analyzed to facilitate real-time analysis. This is a significant challenge which necessitates the use of data dimensionality reduction algorithms. The goal of these algorithms is to represent the smart meter data in a lower dimensional space while retaining the essential data properties. This can speed up computations and the creation of clusters to classify smart meter data into groups with common characteristics. For example, clusters of consumers with similar consumption patterns can be used for tariff design, expansion planning, and other operational functionalities.

The dimensionality reduction algorithms employed to reduce the computational complexity associated with smart meter data can be computationally intensive for practical applications. This is the motivation behind the low complexity solutions proposed in the literature [6]–[9]. A kernel-based method was presented in [6] to reduce the dimensionality of smart meter data. The statistical properties of the data were investigated to determine consumer behavior in terms of energy consumption. This method employs principal component analysis (PCA) as well as non-parametric clustering to classify consumers and allocate the power load efficiently. However, the computational complexity of this method is high when there is a large volume of smart meter data, specifically in the dimension reduction process. Further, this method distributes the consumers using homogenous consumption distances. This can cause improper classification as peak consumption is not used to determine the clusters. The clustering method proposed in [7] is based on the K-means algorithm with the initial points selected either randomly or deterministically. This can increase the search space for locating similar load profile data within a cluster as it necessitates spanning the entire dataset, which is typically large.

The method proposed in [8] considers load profiles to analyze consumer behavior and obtain consumption patterns. These results can be used to forecast energy and capacity market prices. Two steps are used to extract data for analyzing customer consumption behavior dynamics. First, a symbolic aggregate approximation is employed to reduce the dimensionality of the dataset. Then, a Markov model is used to model the consumption behavior. However, using two steps can slow the data dimensionality reduction process. A distributed classification algorithm was employed in [9] to distribute the processing load over several servers. Although, this can speed up computations, it can also degrade classification performance.

The above results indicate that previous approaches for clustering can be slow to cluster load profile data. This motivates the development of a fast, low complexity algorithm for analyzing load profile data. First, a polar projection method is introduced which significantly reduces the computational complexity of dimension reduction. To the best of our knowledge, this is the first time this approach has been adopted for data dimensionality reduction in big data analytics. Second, heterogeneous clusters are considered where the thresholds for determining the cluster sizes are based on peak consumption. This can improve the performance of load profile clustering while reducing the computational burden. Finally, a stochastic-based algorithm is proposed to reduce the search space for polar-based clustering. A multistage algorithm is employed where small subspace searches are used to find the cluster centers. Then, load profile data are allocated using these centers. This multistage approach reduces the computational complexity associated with a large search space and so speeds up clustering, which is particularity important for real-time data analysis.

The rest of this article is organized as follows. Section II presents the system model and assumptions as well as a review of the related work. The proposed low complexity algorithm which employs a polar projection for dimension reduction is given in Section III. Performance results and a comparison with previous techniques are presented in Section V. Finally, some conclusions are given in Section VI.

II. BACKGROUND, DEFINITIONS, AND ASSUMPTIONS

This section first provides the background for data dimensionality reduction. Then the principles, definitions, and assumptions used throughout the paper are discussed. The K-means clustering algorithm is described and previous approaches to clustering and dimensionality reduction for smart meter data are reviewed.

A. DATA DIMENSIONALITY REDUCTION

Dimensionality reduction is used to convert data into an appropriate representation with reduced dimensionality. There are numerous practical applications such as image segmentation [10], text classification [11], face recognition [12], and pattern recognition, where the data dimensionality is very high. To reduce the dimensionality of data without losing a significant amount of information, many methods have been proposed, most of which use machine learning algorithms. These algorithms can be classified as supervised, unsupervised, or semi-supervised.

With supervised methods, a training set from known data is labeled. This is used to find a good predictor for the data classes. Supervised algorithms include linear discriminant analysis (LDA) [13], support vector machine (SVM) [14], independent component analysis (ICA) [15], and kernel principal component analysis (PCA) [6]. Kernel PCA employs a kernel to create a nonlinear version of PCA as linear projections can significantly degrade PCA performance. Unsupervised methods discover structures from unlabeled data. Unsupervised algorithms used for dimensionality reduction include PCA, singular value decomposition (SVD) [16] and ICA [15]. Typically, the amount of labeled data is limited due to the high cost of labeling data while unlabeled data is relatively easy to obtain. Semi-supervised methods are used to effectively exploit both labeled and unlabeled data [17].

In this article, an unsupervised dimension reduction algorithm is proposed which is based on characteristics of the load profiles. As discussed above, previous techniques can be slow and have high computational complexity if they are employed for dimension reduction of large datasets. A major contribution of this article is the use of a polar projection method to obtain a low complexity solution for data dimension reduction.

B. CLUSTERING

Clustering is the ordering or separation of data into different categories. One of the most commonly employed clustering algorithms is K-means which classifies a dataset into K clusters [18]. The following two steps are executed iteratively: 1) assign the data points and 2) update the centroids. In the first step, every point is assigned to the cluster whose centroid is nearest to it. The second step updates the centroid locations according to the mean (center) of the points assigned to a cluster. This process is iterated until the centroids do not move, or equivalently when the points stop switching clusters. In the remainder of the paper, we employ the notation presented above and use the K-means clustering algorithm.

C. DISTANCE METRICS

The distance metric is a key component of a clustering algorithm. The choice of metric depends on the characteristics of the dataset as well as the dataset operations. For example, the *Euclidean* distance is commonly employed as it is not affected by rotation, reflection or translation operations. Several metrics have been proposed in the literature [19]. In [20], the sum of the differences between features was considered as the distance between data points. This method is used in the *Fuzzy-ART* algorithm due to its low computational complexity, and it has also been employed to create rectangular clusters in [21]. The *Sup* distance proposed in [22] uses the greatest difference between features. In this case, features that have a large variance dominate other features, so they have a greater effect on clustering. These metrics are all special cases of the distance metric proposed in [23]. In this article, a new distance metric is introduced considering the characteristics of load profile data. These characteristic are related to the total as well as peak consumption.

III. POLAR PROJECTION FOR LOW COMPLEXITY DIMENSION REDUCTION

Assume a data set is represented by an $N \times M$ matrix X with M dimensions consisting of N data vectors X_n, $n \in \{1, 2, \ldots, N\}$. If this dataset has intrinsic dimensionality M' where $M' < M$, then the data points $x_{n,m}$ within dataset X are located on or near a manifold with dimensionality M' that is embedded in the M-dimensional space. Dimensionality reduction is used to transform X with dimensionality M into a new dataset Y. The goal is to retain as much of the corresponding geometry as possible. Throughout this article, the low-dimensional counterpart of X_n is denoted by Y_n, where Y_n is the nth row of Y. Then, the matrix X with row vectors X_n can be expressed as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} & \cdots & x_{2,M} \\ \vdots & \vdots & & & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} & \cdots & x_{n,M} \\ \vdots & \vdots & & & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,m} & \cdots & x_{N,M} \end{pmatrix}. \tag{1}$$

Define the projection vector as

$$P = \left[p_1, p_2, \ldots, p_m, \ldots, p_M\right]^T.$$

This is used to transform X into a matrix Y

$$Y = XP, \tag{2}$$

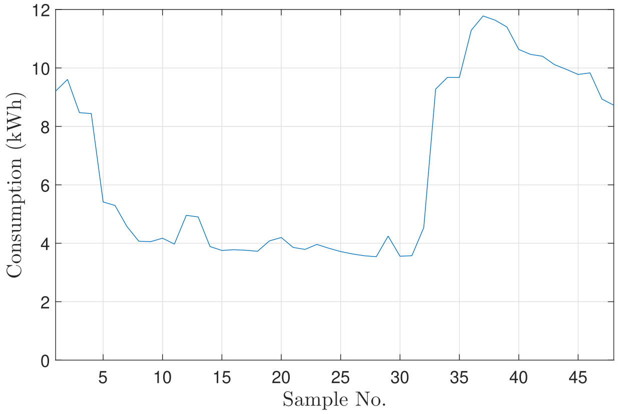

FIGURE 1. **A typical customer consumption load profile over one day with a 30 minute sampling period.**

and expanding Y gives

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} X_1 P \\ X_2 P \\ \vdots \\ X_n P \\ \vdots \\ X_N P \end{pmatrix} = \begin{pmatrix} \sum_{m=1}^{M} x_{1,m} p_m \\ \sum_{m=1}^{M} x_{2,m} p_m \\ \vdots \\ \sum_{m=1}^{M} x_{n,m} p_m \\ \vdots \\ \sum_{m=1}^{M} x_{N,m} p_m \end{pmatrix}. \tag{3}$$

Therefore, the elements of the projection vector P convert a basis of the X space into a basis of the Y space. In this case, the columns of Y are linear combinations of the columns of X mapped via P.

Two key features related to customer load profiles are *total consumption* and *peak consumption*. Dimensionality reduction is employed to select and extract these features. Feature selection selects a subset of the original feature set while feature extraction builds a new set of features from the original feature set. The proposed low complexity solution using a *polar projection* performs both feature selection and feature extraction for customer load profiles. Let element $x_{n,m}$ in X be the power consumption of consumer n measured during a particular sampling period m, i.e., X_n describes the load profile collected by a *smart meter* for customer n, $1 \leq m \leq M$, and M is the number of samples in a day. The total consumption of customer n during the day is then

$$C_{n,total} = \sum_{m=1}^{M} x_{n,m}. \tag{4}$$

A typical customer load profile is illustrated in Fig. 1. This shows the consumption over one day with a 30 min sampling period.

The proposed technique for data dimensionality reduction employs the projection vector

$$p_m = e^{\alpha \theta_m}, \tag{5}$$

where

$$\theta_m = \frac{\pi}{2} \frac{(m-1)}{M-1}, \tag{6}$$

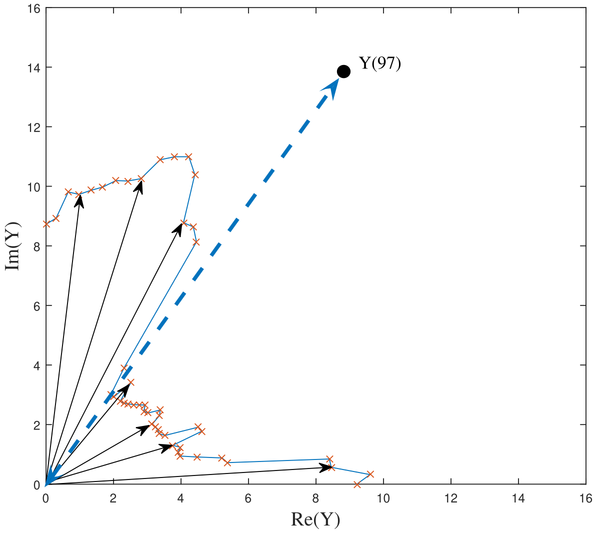

FIGURE 2. The polar projection of the load profile in Fig. 1.

and $\alpha = \sqrt{-1}$. θ_m lies between 0 and $\frac{\pi}{2}$, so the product $x_{n,m}p_m$ approximates $C_{n,total}$ for the reduced dimension vectors. Further, feature extraction is performed to obtain the peak consumption. This is estimated as the angle of the projected data point Y_n.

From (3) and (5), the projected matrix is

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \\ \vdots \\ Y_N \end{pmatrix} = \begin{bmatrix} \sum_{m=1}^{M} x_{1,m}e^{\alpha\theta_m} \\ \sum_{m=1}^{M} x_{2,m}e^{\alpha\theta_m} \\ \vdots \\ \sum_{m=1}^{M} x_{n,m}e^{\alpha\theta_m} \\ \vdots \\ \sum_{m=1}^{M} x_{N,m}e^{\alpha\theta_m} \end{bmatrix}. \tag{7}$$

The vectors Y_n represent the consumer load profiles as complex variables. These values can be expressed using magnitude (r) and angle (θ) in a polar coordinate system which gives the matrix

$$B = \begin{pmatrix} B_1 \\ B2 \\ \vdots \\ B_n \\ \vdots \\ B_N \end{pmatrix} = \begin{pmatrix} |Y_1| & \angle Y_1 \\ |Y_2| & \angle Y_2 \\ \vdots & \vdots \\ |Y_n| & \angle Y_n \\ \vdots & \vdots \\ |Y_N| & \angle Y_N \end{pmatrix} = \begin{pmatrix} r_1 & \theta_1 \\ r_2 & \theta_2 \\ \vdots & \vdots \\ r_n & \theta_n \\ \vdots & \vdots \\ r_N & \theta_N \end{pmatrix} \tag{8}$$

Fig. 2 gives the projection of the load profile in Fig. 1. The number of calculations required with the proposed method varies linearly with the number of samples N and dimension M of the data, so the computational complexity is $O(4NM)$. This is much less than the complexity of previous dimension reduction techniques, e.g., PCA which has complexity $O(NM^2 + M^3)$. The difference will be significant with a large dataset, i.e., when N and M are large. The proposed method also has the advantage over previous techniques that dimension reduction can be done on each load profile individually, so that the smart meters can perform this procedure and send the result to the MDMS. This leads to a significant reduction in bandwidth for data transmission, memory for data storage, and time for data processing.

A. LOAD PROFILE CLUSTERING

We introduce a distance metric based on features extracted from a dataset corresponding to consumer load profiles. The clusters are obtained using the K-means algorithm. They represent consumers whose load profiles are similar and the cluster sizes can vary. The K-means algorithm is used to obtain K clusters from the projected consumer load profiles B. Let $\epsilon_k = [r_k, \theta_k]$ be the center of cluster k in the polar coordinate system. Then, the new distance metric is defined as

$$D(\epsilon_k, B_n) = \sqrt{\left(\frac{r_n - r_k}{r_k}\right)^2 + (\theta_n - \theta_k)^2}, \tag{9}$$

where $B_n = [r_n, \theta_n]$ is an element of B. This distance is normalized by r_k which is the magnitude of the cluster center, so it is adjusted according to the total consumption of consumers within the cluster. Further, as the peak time of load profile n is denoted by the angle θ_n, $D(\epsilon_k, B_n)$ includes the difference in peak times between load profile n and the cluster center k.

An important task in clustering is to identify the appropriate number of clusters K. A large number of clusters increases the complexity of the clustering algorithm while a small number can cause information loss. Several techniques have been proposed to determine the number of clusters for the K-means algorithm. The sum of squared errors (SSE) is the most widely used technique for this purpose and so is employed here. It is given by

$$\text{SSE} = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N} D(\epsilon_k, B_n)}{N}, \tag{10}$$

where $D(\epsilon_k, B_n) = 0$ if B_n is not in cluster k.

B. STOCHASTIC BASED FAST CLUSTERING

In this section, a multi-stage algorithm is proposed to speed up the clustering process. Geometry-based clustering algorithms such as K-means and Gaussian mixture models cluster data according to geometric (Euclidean) distances. This motivates the use of a multi-stage algorithm in which a subset of the data is randomly chosen to represent the dataset, and clustering is performed on this subset. Since this method uses a subset of the data, it can reduce the computational complexity related to a large search space and thus speed up the clustering process. The remainder of the data is then assigned to clusters using the chosen distance metric. To determine the effect of using a subset of data on clustering, we show that the mean for a subset is linearly related to the mean of the entire dataset.

Let $[B_1, B_2, \ldots, B_l, \ldots, B_L]^T$ be a subset of the mapped dataset B, $L < N$, with the elements B_l randomly chosen

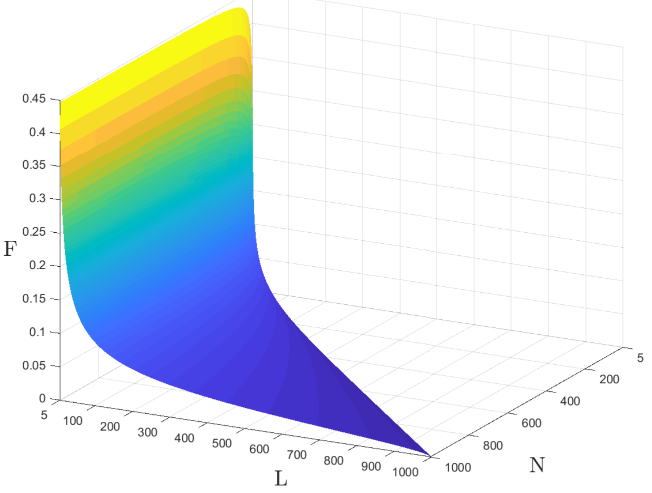

FIGURE 3. The value of *F* versus subset size *L* and dataset size *N*.

Algorithm 1 Polar-Based Projection Data Dimensionality Reduction for Load Profile Clustering

```
Input: Data points $X$, number of clusters $K$, number of subsets $L$, projection vector $P$
Step 1: Select $L$ data points from $X$ randomly
Step 2: Obtain dataset $Y$ by projecting $X$ using $P (Y = XP)$
Step 3: Convert $Y$ to $B$ by decomposing the complex elements into their magnitudes $r$ and angles $\theta$
Step 4: Initialize the $K$ cluster centers randomly
Step 5: Allocate the $L$ data points to the $K$ clusters using the distance metric
Step 6: Update the positions of the centers and $K$ such that the required SSE is met
Step 7: Repeat steps 5 and 6 until there are no more in the data points in the clusters
Step 8: Associate the data points other than the $L$ data points considered in clustering with the nearest centers
```

from $B_1, B_2, \ldots, B_N$. The subset mean is $\overline{B}_l = \frac{1}{L}\sum_{l=1}^{L} B_l$ and the mean of B is $\overline{B}$. The covariance matrix is

$$\begin{aligned}
&\mathbb{E}\Big[\big(\overline{B}_l - \overline{B}\big)\big(\overline{B}_l - \overline{B}\big)^T\Big]\\
&\quad = \mathbb{E}\Big[\big(\overline{B}_l\big)\big(\overline{B}_l\big)^T\Big] + \mathbb{E}\Big[\overline{B}\overline{B}^T - \overline{B}\big(\overline{B}_l\big)^T - \big(\overline{B}_l\big)\overline{B}^T\Big]\\
&\quad = \mathbb{E}\Bigg[\big(\overline{B}_l\big)\,\big(\overline{B}_l\big)^T\Big) - \overline{B\overline{B}^T}\Bigg]\\
&\quad = \frac{1}{L^2} \times \frac{\binom{N-1}{L-1}\sum_{i=1}^{N} B_i B_i^T}{\binom{N}{L}} + \frac{2}{L^2} \times \frac{\binom{N-2}{L-2}\sum_{j=1}^{N}\sum_{i=1}^{N} B_i B_j^T}{\binom{N}{L}}\\
&\qquad - \frac{\sum_{i=1}^{N} B_i B_i^T}{N^2} - \frac{2\sum_{j=1}^{N}\sum_{i=1}^{N} B_i B_j^T}{N^2}\\
&\quad = \frac{\sum_{i=1}^{N} B_i B_i^T}{LN} + \frac{2(L-1)\sum_{j=1}^{N}\sum_{i=1}^{N} B_i B_j^T}{LN(N-1)} - \frac{\sum_{i=1}^{N} B_i B_j^T}{N^2}\\
&\qquad - \frac{2\sum_{j=1}^{N}\sum_{i=1}^{N} B_i B_j^T}{N^2}\\
&\quad = \frac{(N-L)(N-1)\sum_{i=1}^{N} B_i B_j^T}{LN^2(N-1)} - \frac{(N-L)\sum_{j=1}^{N}\sum_{i=1}^{N} B_i B_j^T}{kN^2(N-1)}\\
&\quad = \frac{N-L}{L(N-1)}\left(\frac{\sum_{i=1}^{N} B_i B_j^T (N-1)}{N^2} - \frac{2\sum_{j=1}^{N}\sum_{i=1}^{N} B_i B_j^T}{N^2}\right)\\
&\quad = \frac{N-L}{L(N-1)}\mathbb{E}\Big[\big(B - \overline{B}\big)\big(B - \overline{B}\big)^T\Big]. \qquad (11)
\end{aligned}$$

where $\mathbb{E}$ denotes expectation. This shows that the covariance matrix of the random variables $\overline{B}_l$ is related to the covariance matrix of the original dataset B by a factor

$$F = \sqrt{\frac{N-L}{L(N-1)}}. \qquad (12)$$

This factor can be used to determine L based on an acceptable degradation in clustering performance. Fig. 3 gives values of F for different values of L and N. This shows that F decreases with increasing N and L.

Let ϵ'_k and ϵ_k denote the cluster centers in subset B_l and the original dataset B, respectively. The estimation error is defined as

$$E = \sum_{k=1}^{K} \frac{|\epsilon'_k - \epsilon_k|}{|\epsilon_k|}, \qquad (13)$$

where K is the number of clusters. For example, if $N = 1000$, $E \approx 0.12$ when $L = 400$. The proposed stochastic method using polar-based projection is summarized in Algorithm 1. This method first reduces the dimension of the dataset using a projection vector P. Then, the projected dataset is clustered using the stochastic-based method with the new distance metric, and the ratio L/N is set according to the required SSE from (10).

As previously mentioned, the computational complexity of polar-based clustering is $O(4NM)$, so the stochastic-based algorithm reduce this to $O(4LM)$, i.e., L load profiles are used for clustering instead of all N load profiles.

IV. NUMERICAL RESULTS

In this section, numerical results are presented to illustrate the performance of the proposed polar-based clustering with the new distance metric and multi-stage stochastic clustering. Real consumption data from [25] is used with $m = 48$ samples per day (30 min intervals). First, the polar-based projection is used to reduce the dimensionality of the dataset to 5466×2. Then, the performance of K-means clustering using the proposed and Euclidean metrics is evaluated. The number of clusters required to achieve a reasonable SSE is $K = 4$ and the initial cluster centers are chosen randomly. The performance and computational complexity of stochastic-based clustering and conventional K-means clustering are then evaluated.

A. DISTANCE METRIC PERFORMANCE

Fig. 4 shows the clusters obtained using the K-means algorithm with the Euclidean distance. In this case, only the magnitude r_n is used to assign consumer load profiles to

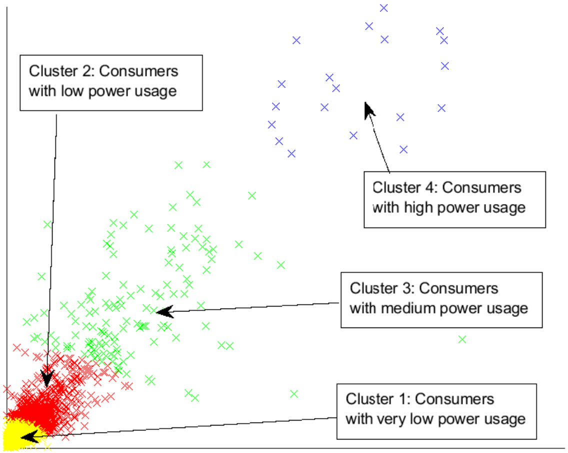

FIGURE 4. ***K*-means clustering with *K* = 4 and the Euclidean distance where consumers are classified according to total consumption.**

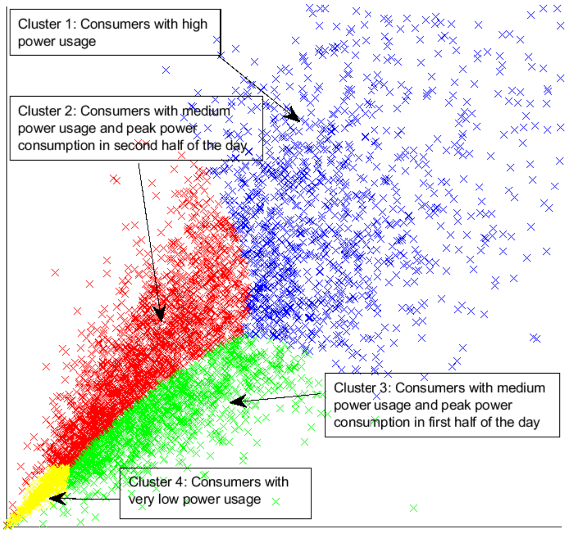

FIGURE 5. ***K*-means clustering with *K* = 4 and the new distance metric where consumers are classified according to total and peak consumption.**

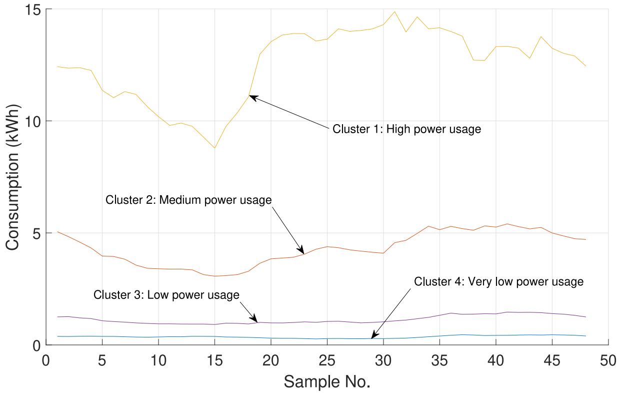

FIGURE 6. Average consumption within each cluster with *K* = 4 versus sample number where the clusters are classified as high, medium and low power usage.

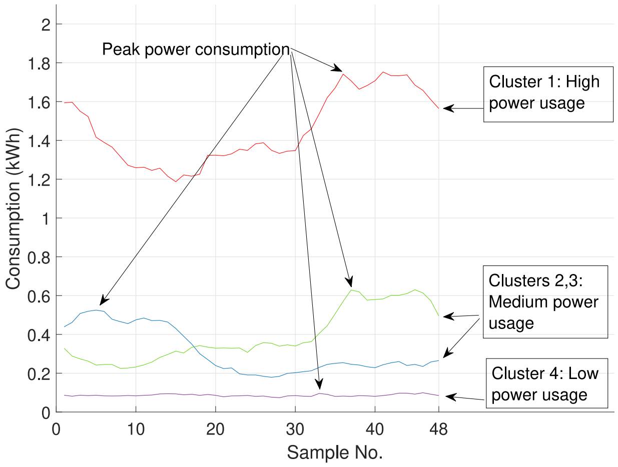

FIGURE 7. Average consumptions within each cluster with *K* = 4 versus the sample number where the clusters are classified as high, medium and low power usage. The corresponding peak consumption is also indicated.

clusters. Thus, the clusters represent consumers with similar total power consumption as discussed in Section III-A. However, the peak consumption is an important feature for classifying load profiles. The proposed distance metric incorporates the angle θ which represents the peak time. Therefore each cluster represents both similar total consumption and similar peak consumption. The corresponding clusters are shown in Fig. 5. This indicates that the new distance metric can take into account the peak consumption which is an important feature to classify consumer load profiles.

Fig. 6 presents the clustering results versus the sample number during a day. This shows typical consumer load profiles where the consumers are classified into clusters with high, medium and low power usage. In this case, the peak consumption was not taken into account. Fig. 7 presents the results for the same scenario when the clustering is performed according to peak consumption. This clearly illustrates the impact of peak consumption on the clustering using the new distance metric, i.e., the clustering is improved. For example, the consumers within clusters 2 and 3 in Fig. 7 are classified more accurately via their peak consumption compared to the consumers located in the same clusters in Fig. 6.

B. NUMBER OF CLUSTERS

The SSE is used to determine the appropriate number of clusters. Fig. 8 presents the SSE for polar-based K-means clustering using the new distance metric with $K = 2$ to 7. This shows that the SSE decreases significantly when the number of cluster is increased from 3 to 4. However, this decrease in minimal with $K > 4$. This shows that the new distance metric is suitable for a large number of clusters which is desirable for load profile clustering.

C. ESTIMATION ERROR WITH STOCHASTIC-BASED CLUSTERING

As indicated previously, stochastic-based clustering uses a subset of the dataset to speed up the clustering process.

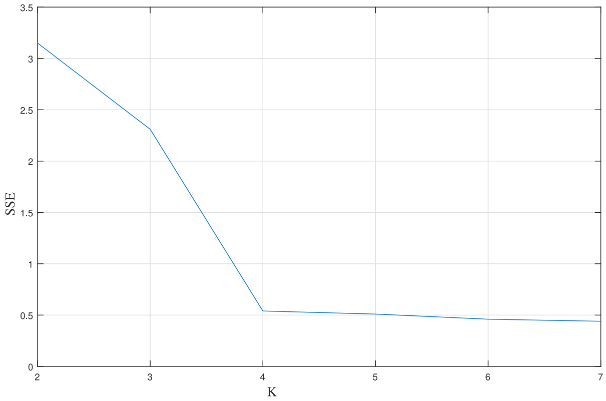

FIGURE 8. SSE versus the number of clusters K for polar-based K-means clustering using the new distance metric.

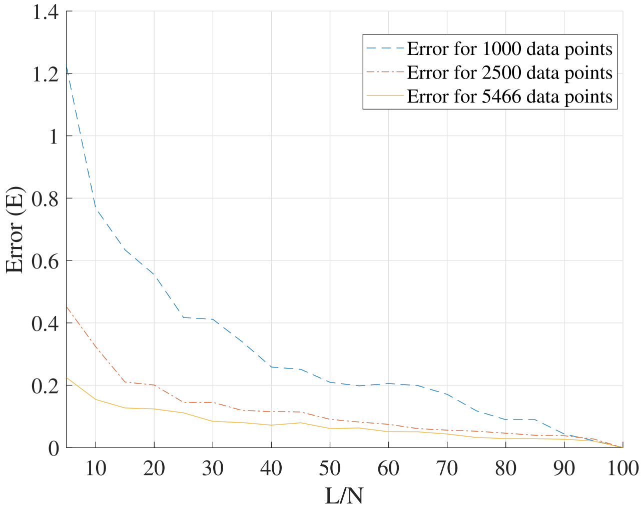

FIGURE 9. The estimation error E versus L/N for multi-stage stochastic clustering.

Fig. 9 presents the error averaged over 1000 trials for the conventional and stochastic-based clustering algorithms versus L/N for $N = 1000$, 2500 and 5466. This shows that the error is significant for $L/N < 0.5$ and decreases quickly when $L/N > 0.5$, particularly for $N = 5466$. This is important as the dataset is typically large. In this case, polar-based dimensionality reduction can significantly reduce the computational complexly of clustering, and this can be reduced further with stochastic-based clustering.

D. COMPUTATIONAL COMPLEXITY

In this section, the computation time of polar-based K-means clustering with and without stochastic clustering is presented. The execution time reduction ratio is defined as

$$R = 1 - \left(t_p/t_s\right), \tag{14}$$

where t_p is the execution time of polar-based stochastic-clustering using subset size L and t_s is the execution time of polar-based clustering using dataset size N. Table 1 gives the execution time reduction ratio R and the estimation error E versus the ratio L/N for $N = 5466$. This indicates that the reduction in computation time with stochastic clustering is 43% for $L/N = 0.60$ and the corresponding estimation error is only $E = 3.12\%$.

TABLE 1. Computational complexity reduction R and estimation error E for the polar-based K-means algorithm with and without stochastic clustering for various ratios L/N.

L/N	R (%)	E (%)
.30	73	8.42
.40	59	7.22
.50	51	6.15
.60	43	3.12
.70	32	2.39

V. CONCLUSION

In this article, a polar-based dimensionality reduction technique was proposed to reduce the complexity associated with clustering large smart meter datasets. This approach employs K-means clustering. A new distance metric was proposed based on total consumption and peak consumption from consumer load profiles. The minimum number of clusters was determined based on the sum of squared errors (SSE). A stochastic-based algorithm was presented to reduce the search space for clustering. Numerical results were presented to illustrate the reduction in computational complexity. These results indicate that the proposed polar-based technique greatly reduces the complexity over the conventional K-means algorithm. Further, stochastic clustering can further decrease the computation complexity while maintaining a reasonable estimation error.

REFERENCES

[1] A. Ghassemi, P. Goudarzi, M. R. Mirsarraf, and T. A. Gulliver, "A stochastic approach to energy cost minimization in smart-grid-enabled data center network," *J. Comput. Netw. Commun.*, vol. 2019, Mar. 2019, Art. no. 4390917. [Online]. Available: https://www.hindawi.com/journals/jcnc/2019/4390917/

[2] P. Goudarzi, A. Ghassemi, M. R. Mirsarraf, and R. Buyya, "Joint energy-QoE efficient content delivery networks using real-time energy management," *IEEE Syst. J.*, vol. 14, no. 1, pp. 927–938, Mar. 2020.

[3] D. Alahakoon and X. Yu, "Smart electricity meter data intelligence for future energy systems: A survey," *IEEE Trans. Ind. Informat.*, vol. 12, no. 1, pp. 425–436, Feb. 2016.

[4] M. Ghorbanian, S. H. Dolatabadi, and P. Siano, "Big data issues in smart grids: A survey," *IEEE Syst. J.*, vol. 13, no. 4, pp. 4158–4168, Dec. 2019.

[5] S. Singh, A. Yassine, and S. Shirmohammadi, "Incremental mining of frequent power consumption patterns from smart meters big data," in *Proc. IEEE Electr. Power Energy Conf.*, Ottawa, ON, Canada, Oct. 2016, pp. 1–6.

[6] E. Pan, H. Li, L. Song, and Z. Han, "Kernel-based non-parametric clustering for load profiling of big smart meter data," in *Proc. IEEE Wireless Commun. Netw. Conf.*, New Orleans, LA, USA, Mar. 2015, pp. 2251–2255.

[7] S. A. Azad, A. B. M. Shawkat Ali, and P. Wolfs, "Identification of typical load profiles using K-means clustering algorithm," in *Proc. Asia Pac. World Congr. Comput. Sci. Eng.*, Nadi, Fiji, Nov. 2014, pp. 1–6.

[8] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2437–2447, Sep. 2016.

[9] H. Wang, Y. Gao, Y. Shi, and H. Wang, "A fast distributed classification algorithm for large-scale imbalanced data," in *Proc. IEEE Int. Conf. Data Min.*, Barcelona, Spain, Dec. 2016, pp. 1251–1256.

[10] G. Bilgin, S. Ertürk, and T. Yildirim, "Nonlinear dimension reduction methods and segmentation of hyperspectral images," in *Proc. IEEE Signal Process. Commun. Appl. Conf.*, Aydin, Turkey, Apr. 2008, pp. 1–4.

[11] T. Gasanova, R. Sergienko, E. Semenkin, and W. Minker, "Dimension reduction with coevolutionary genetic algorithm for text classification," in *Proc. Int. Conf. Inform. Control Autom. Robot.*, Vienna, Austria, Sep. 2014, pp. 215–222.

[12] J. Wang, "Efficient face recognition research based on random projection dimension reduction," in *Proc. Int. Conf. Meas. Technol. Mechatronics Autom.*, Nanchang, China, Jun. 2015, pp. 415–418.

[13] E. I. G. Nassara, E. Grall-Maës, and M. Kharouf, "Linear discriminant analysis for large-scale data: Application on text and image data," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, Anaheim, CA, USA, Dec. 2016, pp. 961–964.

[14] J. H. Ang, E. J. Teoh, C. H. Tan, K. C. Goh, and K. C. Tan, "Dimension reduction using evolutionary support vector machines," in *Proc. IEEE Congr. Evol. Comput.*, Hong Kong, Jun. 2008, pp. 3634–3641.

[15] 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, Feb. 2007.

[16] P. Song *et al.*, "Accelerated singular value-based ultrasound blood flow clutter filtering with randomized singular value decomposition and randomized spatial downsampling," *IEEE Trans. Ultrason., Ferroelect., Freq. Control*, vol. 64, no. 4, pp. 706–716, Apr. 2017.

[17] 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.

[18] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 136–144, Jan. 2016.

[19] X. Rui and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[20] C. H. Chen, *Handbook of Pattern Recognition and Computer Vision*, 5th ed. Singapore: World Sci., 2016.

[21] V. G. Kaburlasos and X. Gerhard, *Computational Intelligence Based on Lattice Theory*. Berlin, Germany: Springer, 2007.

[22] L. Bobrowski and J. C. Bezdek, "c-means clustering with the l_1 and l_∞ norms," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 545–554, May/Jun. 1991.

[23] A. Okabe, B. Boots, and K. Sugihara, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. New York, NY, USA: Wiley, 1992.

[24] J. Kleinberg, "An impossibility theorem for clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2002, pp. 463–470.

[25] Irish Social Science Data Archive. *Commission for Energy Regulation Smart Metering Project—Electricity Customer Behaviour Trial, 2009-2010 [dataset]*. Accessed: 2018. [Online]. Available: http://www.ucd.ie/issda/cer-electricity

MORTEZA MOHAJERI received the B.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2010, and the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran in 2013. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Tehran. His research interests include smart grid, big data analytics, and intelligent information processing technology and its application in power system.

ABOLFAZL GHASSEMI (Senior Member, IEEE) received the M.A.Sc. and Ph.D. degrees in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2003 and 2009, respectively. He was a Postdoctoral Fellow with the University of British Columbia from 2009 to 2010. From 2010 to 2012, he was a Postdoctoral Fellow with Stanford University in the Space, Telecommunications, and Radio Science Laboratory. He is currently working with the Department of Electrical and Computer Engineering, University of Victoria. His research interests include big data analytics, Internet of Things, smart grid, and wireless communications. He received the Best paper Award at the IEEE Communication Networks and Services Research conference in 2009 and was awarded a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship in 2009. He is the Founder of Nika Smart Tech, Vancouver, BC, Canada.

T. AARON GULLIVER (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 1989. From 1989 to 1991, he was employed with Defence Research Establishment Ottawa, Ottawa, ON, Canada. He has held academic appointments with Carleton University, Ottawa, ON, Canada, and the University of Canterbury, Christchurch, New Zealand. He joined the University of Victoria in 1999, where he is a Professor with the Department of Electrical and Computer Engineering. His research interests include information theory and communication theory, algebraic coding theory, multicarrier systems, smart grid, relay, and cooperative communication systems and security. In 2002, he became a Fellow of the Engineering Institute of Canada. In 2012, he was elected as a Fellow of the Canadian Academy of Engineering.