Advanced Metering Infrastructure Data Driven Phase Identification in Smart Grid

Wenyu Wang, Nanpeng Yu, Zhouyu Lu

Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, California 92521
Email: wwang032@ucr.edu, nyu@ece.ucr.edu, zlu044@ucr.edu

***Abstract*—Many important distribution network applications, such as load balancing, state-estimation, and network reconfiguration, depend on accurate phase connectivity information. The existing data-driven phase identification algorithms have a few drawbacks. First, the existing algorithms require the number of phase connections as an input. Second, they can not provide accurate results when there is a mix of phase-to-neutral and phase-to-phase connected smart meters, or when the distribution circuit is less unbalanced. This paper develops an advanced metering infrastructure (AMI) data driven phase identification algorithm that addresses the drawbacks of the existing solutions in two ways. First, it leverages a nonlinear dimensionality reduction technique to extract key features from the voltage time series. Second, a constraint-driven hybrid clustering (CHC) algorithm is developed to dynamically create smart meter clusters with arbitrary shapes. The field validation results show that the proposed algorithm outperforms the existing ones. The improvement in the phase identification accuracy is more pronounced for distribution feeders that are less unbalanced. In addition, this paper discovers that more granular voltage time series leads to higher phase identification accuracy.**

***Keywords*—AMI; density-based clustering; phase identification; smart grid; t-SNE.**

I. Introduction

It is estimated that electric utilities around the world will spend $10.1 billion on advanced metering infrastructure (AMI) data analytics solutions through 2021 [1]. The boom in the development and implementation of AMI data analytics is driven by two trends. First, electric utilities which have already deployed or plan to deploy the AMI are looking for new value streams to justify the business case of the AMI projects. Second, the advent of distributed energy resources (DERs) on the edge of the distribution grid is creating significant challenges and opportunities for the electric utilities and third-party aggregators.

The phase identification problem is defined as identifying the phase connectivity of each smart meter and structure in the power distribution network [2]. It is a critical AMI data analytics application due to two reasons. First, the rise of DERs requires the distribution system operators to actively manage the distribution grid to coordinate the operations of the DERs. However, most electric utilities in the world do not have accurate records of the phase connectivity of their distribution networks to enable advanced control strategies. Second, it is labor and capital intensive to perform phase identification using field validation tools. Therefore, conducting phase identification with AMI data driven analytics can provide another useful justification for the deployment of AMI projects.

In this paper, an AMI data driven machine learning algorithm is developed to solve the phase identification problem. The proposed algorithm leverages voltage magnitude data recorded by the AMI to identify the phase connection of each smart meter and structure. A nonlinear dimensionality reduction technique is first used to extract key features from the voltage time series. A constraint-driven hybrid clustering (CHC) algorithm is then developed to separate smart meters/structures into various clusters. Finally, the phase connection of each cluster can be identified by performing field validations on the phase connections of very few smart meters. Comprehensive case studies are conducted on 5 distribution circuits, which went through detailed field validations. The AMI data driven machine learning algorithm has yielded high accuracies on all circuits. In addition, this paper discovers that more granular voltage readings will lead to even more accurate phase identification results.

Compared to the existing data-driven phase identification algorithms, the proposed method has the following advantages:

1) The proposed algorithm does not require prior knowledge about the number of phase connections in the distribution system. Most of the existing AMI data driven methods need the number of phase connections as an input parameter.
2) The proposed algorithm works well with distribution feeders that have both phase-to-neutral and phase-to-phase connections. Most of the existing techniques are only capable of identifying the phase connections in distribution feeders with only phase-to-neutral connections or phase-to-phase connections.
3) The accuracy of the proposed phase identification algorithm is not very sensitive to the level of unbalance in a distribution feeder.

Currently, most electric utilities conduct phase identification using special phase meters [3][4]. Typically, two phase meters/units are used. One unit is located at the substation to serve as the reference. The other is called the field unit and is located at the smart meter/structure of interest in the distribution feeder. The working mechanism of these special phase meters is

 ISBN: 978-1-61208-588-3

very similar to that of the phasor measurement units except that the phase meters are mobile. With GPS time, the phase angle difference between the reference point and the field structure can be accurately measured, which then determines the phase connectivity of the field structure. Although phase meters provide highly accurate phase identification results, this solution is very time consuming and labor intensive, which make it unsuitable for large-scale deployment.

The existing data-driven algorithms leverage electric load and voltage magnitude measurements from the AMI to identify the phase connections of the smart meters and structures in the distribution network. These data-driven algorithms include supply and consumption balancing [5][6], linear regressions and correlation analysis [7][8], and constrained k-means clustering algorithm (CK-Means) [2]. However, the existing data-driven algorithms have three drawbacks. First, all of these methods assume that the number of phase connections are known. Second, the existing methods can not provide accurate phase identification results when there is a mix of phase-to-neutral and phase-to-phase connected smart meters and structures. Third, the existing methods are quite sensitive to the level of unbalance in a distribution feeder. The proposed AMI data driven phase identification algorithm addresses these drawbacks by leveraging a nonlinear dimensionality reduction technique to extract hidden features from voltage time series and using the CHC algorithm to dynamically create smart meter clusters with arbitrary shapes. The field validation results show that the proposed algorithm outperforms the existing methods in all of the 5 distribution feeders.

The rest of this paper is organized as follows. Section II studies the drawbacks of the existing data-driven phase identification algorithms. Section III describes the proposed phase identification algorithm in detail. Section IV presents the case studies on multiple distribution feeders to validate the proposed phase identification algorithm. Section V provides the conclusions.

II. Drawbacks of the Existing Data-driven Phase Identification Methods

Three main drawbacks of the existing phase identification methods are studied in detail below. As the CK-Means method is the most promising algorithm among the existing data-driven phase identification methods, it will be used as an example in the performance evaluation. A comprehensive study is conducted on 5 distribution feeders and 18 data sets to analyze the impact of unbalance level and the mix of phase connection types on the phase identification accuracy for the CK-Means method.

The general descriptions of the 5 distribution feeders and 18 data sets are shown in Table I. The feeder and smart meter data is provided by the Pacific Gas & Electric Company and Southern California Edison. The number of customers, feeder voltage level, proportion of the major phase connection types, and feeder peak load are listed in the second column of the table. A distribution feeder can have 3 possible phase-to-neutral connections, AN, BN, and CN, and/or 3 possible phase-to-phase connections, AB, BC, and CA, where A, B, C, and N denote the three phases' wires and the neutral wire. 2 months of smart meters' voltage data with 5-minute granularity is gathered from feeder 1, 2, and 3. 6 months of smart meters' voltage data with hourly granularity is gathered from feeder 4 and 5.

TABLE I. Descriptions of the Distribution Feeders

Feeder	Number of Customers, Feeder Voltage, and Peak Load	Month	Data Set
1	3200 customers (99.8% phase-to-neutral), 12.47 kV, 4.4 MW.	Nov 2016	s_1
		Dec 2016	s_2
2	4800 customers (98.8% phase-to-neutral), 12.47 kV, 8.3 MW.	Nov 2016	s_3
		Dec 2016	s_4
3	4000 customers (97% phase-to-neutral), 12.47 kV, 6.4 MW.	Nov 2016	s_5
		Dec 2016	s_6
4	1500 customers (100% phase-to-phase), 12.47 kV, 5.2 MW.	Aug 2015	s_7
		Sep 2015	s_8
		Oct 2015	s_9
		Nov 2015	s_{10}
		Dec 2015	s_{11}
		Jan 2016	s_{12}
5	2400 customers (84% phase-to-phase), 12.47 kV, 8.5 MW.	Aug 2015	s_{13}
		Sep 2015	s_{14}
		Oct 2015	s_{15}
		Nov 2015	s_{16}
		Dec 2015	s_{17}
		Jan 2016	s_{18}

In feeder 1, 2, and 3, some meters have missing voltage readings at different time intervals, making up 9%, 21%, and 18% of the total customer population respectively. The missing readings are filled in using the k-nearest neighbor (k-NN) imputation method. A meter's missing readings are imputed using the average values of the five nearest neighbor meters' corresponding readings. The distance between meters are measured by the Euclidean distance of the voltage time series of the corresponding meters.

To make the results comparable, the hourly average voltage magnitudes are calculated for feeder 1, 2, and 3. The hourly average voltage magnitudes are used as inputs in this section. Each of the 18 data sets includes one month of voltage magnitude data from a feeder. The drawbacks of the existing data-driven phase identification algorithms are explored in the next three subsections.

A. Number of Phase Connections

In order to solve the phase identification problem, the supply and consumption balancing approach [5][6] requires the number of phase connections in the distribution feeder as an input. In fact, the problem formulation in [5][6] only allows the identification of phase-to-neutral connections where the number of phase connections is 3. In the linear regression and correlation analysis [7][8], the number of phase connections in the feeder is also a mandatory input. In fact, both linear regression and correlation analysis work well when there are only three phase-to-neutral connections. The k-means clustering algorithm is used in the CK-Means method [2], where

 ISBN: 978-1-61208-588-3

the number of phase connections/clusters needs to be known as prior knowledge. When applying the CK-Means method to identify the phase connections of the 5 distribution feeders, the number of clusters is set to be 3 for feeders 1 to 4, given that over 97% of the smart meters in these feeders only have 3 connection types. The number of clusters is set to be 6 for feeder 5.

B. Impact of Unbalance Level on the Phase Identification Accuracy

This subsection evaluates the impact of the distribution feeder's unbalance level on the phase identification accuracy of the CK-Means algorithm. The CK-Means algorithm works as follows: The voltage magnitude measurements are first standardized. Linear features are then extracted by using principal component analysis (PCA) and the top d components are selected. To provide a fair comparison with the proposed phase identification algorithm in Section IV, the number of principal components is set to 30. Next, the data points in the low-dimensional space are clustered by using a constrained k-means clustering algorithm. Must-link constraints are derived from the distribution feeders' connectivity information, which is typically available from the Geographical Information System (GIS). The must-link constraints state that if some smart meters are connected to the same lateral or transformer, then they must be linked together and grouped into the same cluster. To identify the phase of each cluster, field validations are performed on a must-link group of at least 20 smart meters that has the least mean squared distance to the cluster center.

The CK-Means algorithm is applied on the 18 voltage time series from the 5 distribution feeders. The phase identification accuracy is calculated based on independent field validations conducted by the electric utility companies. To measure the level of unbalance of a distribution feeder, define $u(t)$ as the level of unbalance of a feeder at time interval t:

$$u(t) = \frac{|I_A(t) - I_m(t)| + |I_B(t) - I_m(t)| + |I_C(t) - I_m(t)|}{3I_m(t)} \quad (1)$$

where $I_m(t) = \frac{1}{3}(I_A(t) + I_B(t) + I_C(t))$ is the mean of the distribution substation line currents of the three phases. $u(t)$ can be interpreted as the ratio of the average three-phase current deviation to the mean. The average level of unbalance for a distribution feeder $u(t)$ over a month is calculated for each data set.

Figure 1 plots the phase identification accuracy against the level of unbalance. It shows that the CK-Means algorithm is very accurate for the highly unbalanced data sets. As the level of unbalance decreases, the phase identification accuracy drops quickly. This result is very intuitive. Imagine there is a perfectly balanced distribution feeder whose three phase wires have the same load distribution all the time. In this case, the level of unbalance should be zero. Therefore, it is impossible to distinguish the phase connections of the smart meters on the three phases with unsynchronized voltage magnitude measurements.

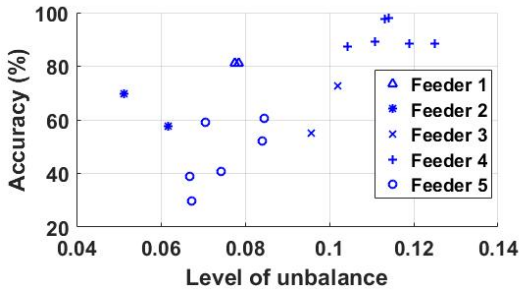

Figure 1. The phase identification accuracy of the CK-Means method under different levels of unbalance.

C. A Mix of Phase-to-Neutral and Phase-to-Phase Connections

In general, the existing data-driven phase identification algorithms do not perform well for the distribution feeders with a mix of phase-to-neutral and phase-to-phase connections. For example, Figure 1 shows that the phase identification accuracy is the lowest for feeder 5. This is because feeder 5 not only has a lower degree of unbalance, but also has all 6 possible phase connections types, AN, BN, CN, AB, BC, and CA. In this case, the default phase identification accuracy is only 16.7% instead of 33.3% for the distribution feeders with only three possible phase connections.

III. TECHNICAL METHODS

The overall framework of the proposed phase identification algorithm is illustrated in Figure 2. The phase identification methodology involves three stages. In stage 1, voltage magnitude measurements are collected from the smart meters. Each smart meter's readings are centered and normalized by their standard deviation. Key features are then extracted from the preprocessed voltage time series with a nonlinear dimensionality reduction method. In stage 2, the CHC algorithm is leveraged to cluster the low-dimensional data points generated in stage 1. In stage 3, the phase connection of each cluster

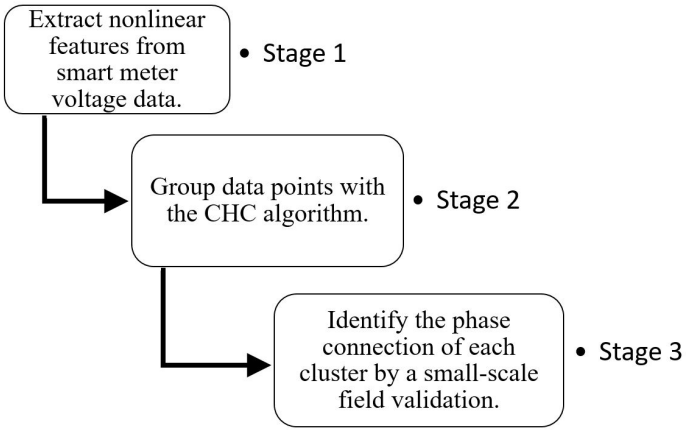

Figure 2. The overall framework of the proposed phase identification algorithm.

 ISBN: 978-1-61208-588-3

is identified by performing field validations on a very small number of smart meters. The three stages are explained in detail below.

A. Stage 1: Feature Extraction from Voltage Time Series

It is undesirable to directly work with raw voltage readings, which are high-dimensional and noisy. Therefore, in the first stage, dimensionality reduction techniques will be applied to extract key features from the raw voltage time series. The extracted features will then be fed into the CHC algorithm in stage 2.

Dimensionality reduction techniques can be divided into two categories, linear dimensionality reduction methods and nonlinear ones. Linear dimensionality reduction techniques, such as PCA, are restricted to learning only linear manifolds. However, high-dimensional data typically lies on or near a low-dimensional, nonlinear manifold [9]. Furthermore, it is very difficult for linear mappings to keep the low-dimensional representations of very similar points close together. This explains the lower accuracy of the phase identification algorithm using linear features for less unbalanced feeders. To address this problem, we turn to nonlinear dimensionality reduction methods. Many nonlinear dimensionality reduction techniques have been proposed, e.g., Sammon mapping [10], curvilinear components analysis (CCA) [11], Isomap [12], and t-distributed stochastic neighbor embedding (t-SNE) [9]. This paper adopts t-SNE, because it has been shown to work well with a wide range of data sets and captures both local and global data structures. t-SNE improves upon SNE [13] by 1) simplifying the gradient calculation with a symmetrized version of the SNE cost function and 2) adopting a Student's t-distribution rather than a Gaussian distribution to compute the similarity between two points in the low-dimensional space [9].

The basic idea of t-SNE is to convert the high-dimensional Euclidean distances between data points into joint probabilities and represent the data points in a low-dimensional space, so that similar joint probabilities are preserved. Suppose we need to map a high-dimensional data set $X = \{x_1, x_2, ..., x_n\}$ to a low-dimensional data set $Y = \{y_1, y_2, ..., y_n\}$. Define p_{ji} as a joint probability of X. p_{ji} is a symmetric approximation of the conditional probability that x_i would pick x_j as its neighbor. The neighbors are picked in proportion to their probability density under a Gaussian distribution centered at x_i with a variance σ_i. p_{ji} is calculated as $p_{ji} = p_{ij} = (p_{j|i} + p_{i|j})/2n$, where $p_{j|i}$ is calculated as:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{l \neq i} \exp(-\|x_i - x_l\|^2/2\sigma_i^2)} \tag{2}$$

In the same way, define q_{ji} as a joint probability in Y, but under a Student's t-distribution with one degree of freedom. Then q_{ji} can be calculated as:

$$q_{ji} = q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{l \neq m}(1 + \|y_l - y_m\|^2)^{-1}} \tag{3}$$

Then given X, the mapping Y is found by minimizing the Kullback-Leibler divergence between joint probability distribution P, in the high-dimensional space, and the joint probability distribution Q, in the low-dimensional space:

$$C = D_{\mathrm{KL}}(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{4}$$

The t-SNE algorithm requires three input parameters: 1) the output dimension d_{out} (typically selected to be either 2 or 3); 2) the initial dimension d_{in}, which is the dimension that the original data set is reduced to by PCA before performing t-SNE; 3) perplexity p, which is a measure of effective number of neighbors and controls σ_i. Since the objective function (4) is minimized using a gradient descent optimization that is initiated randomly, each run of t-SNE produces a slightly different mapping result. In practice, it is recommended to run t-SNE multiple times and select the result with the lowest cost function value in (4). More details of the t-SNE algorithm can be found in [9].

B. Stage 2: Group Data Points with the CHC Algorithm

After the preprocessed voltage time series are mapped to a 2-dimensional or 3-dimensional feature space through t-SNE, they need to be grouped into clusters. Three features of the phase identification problem need to be considered when designing the clustering algorithm. First, many electric utility companies do not know the number of phase connections for each of their distribution feeders. Second, the customers with the same phase connection in the low-dimensional feature space do not necessarily form a convex-shape cluster, which is very common in t-SNE applications [9][14][15]. Third, valuable distribution network connectivity information which defines the mapping between smart meters and laterals/transformers should be incorporated into the clustering algorithm.

In order to leverage the features of the phase identification problem, the CHC algorithm is developed and applied to solve the smart meter clustering problem. The proposed CHC framework synergistically combines the merits of an unsupervised density-based clustering algorithm and a supervised classification algorithm. This paper selects the density-based spatial clustering of applications with noise (DBSCAN) [16] as the unsupervised clustering algorithm in the CHC framework, because it naturally incorporates the first two features of the phase identification problem. Unlike centroid-based or medoid-based methods, DBSCAN does not need the number of clusters as an input parameter. In addition, DBSCAN is capable of discovering clusters with arbitrary shapes.

DBSCAN separates data points into different clusters and noise/outliers. The noise/outliers do not belong to any cluster. However, in the phase identification application, all smart meters must have a particular phase connection. To mitigate this drawback, k-nearest neighbor (k-NN) classification is adopted as the supervised machine learning algorithm in the

 ISBN: 978-1-61208-588-3

CHC framework to assign these outliers and points in the low-density region into one of the existing output clusters from DBSCAN. At last, the must-link constraints defined by the feeder connectivity model will be considered in reassigning smart meters connected to the same lateral/transformer to the same cluster.

1) Review of DBSCAN: A brief review of DBSCAN is provided here. DBSCAN is one prominent example of density-based clustering approach with high computational efficiency. The good efficiency of DBSCAN is crucial for deploying phase identification algorithms in electric utilities with t-housands of distribution feeders. The DBSCAN algorithm defines clusters and outliers based on four key concepts: ϵ neighborhood of a point, directly density-reachable, density-reachable, and density-connected. The algorithm requires two parameters: ϵ, the radius of neighborhood, and *MinPts*, the minimum number of data points in an ϵ neighborhood. The ϵ neighborhood of a point p is defined as the set of points in the data set with a distance to p less than ϵ. A point p is a core point if it has at least *MinPts* neighbors within the radius ϵ. These neighbors are directly density-reachable from p. A point q is density-reachable from p if there is a path $p, p_1, p_2, ..., p_m, q$ such that each point is directly reachable from the previous point. Two points are considered density-connected if they have a distance of less than ϵ. These four definitions allow us to define the transitive hull of density-connected points, forming density-based clusters. The points on the border of the clusters are called border points. Any point(s) not reachable from a core point is counted as an outlier or noise.

2) The CHC Algorithm: The framework of the CHC algorithm is shown in Algorithm 1. It requires four input parameters, α, k, ϵ, and *MinPts*. α is a threshold parameter used to filter out very small clusters. k is the parameter in the k-NN algorithm representing the number of nearest neighbors.

1: Run the DBSCAN algorithm on a preprocessed data set D with n data points with parameters ϵ and *MinPts*.
2: Define a threshold coefficient $\alpha \in (0, 1)$. Given the output of step 1, keep the data points from the clusters of size greater than or equal to αn as the training data set. Suppose there are c clusters kept. All the data points outside these clusters are "un-clustered" data points.
3: Assign all un-clustered data points to one of the c clusters with the k-NN algorithm.
4: With must-link constraints, the data set D can be divided into N groups $D_1, ..., D_N$. If a data point has no links to others, it forms a group itself. In each group D_i, the data points may have been assigned to different clusters. To enforce the constraints, assign all data points in group D_i to the cluster that contains the largest number of data points in D_i.
5: Return the final clustering result.

Figure 3. Algorithm 1: the CHC algorithm

The CHC algorithm has 5 steps. Step 1 runs the DBSCAN algorithm on features extracted by the t-SNE algorithm. Depending on the distribution of data points in the low-dimensional feature space, the DBSCAN output may include large clusters, small clusters, and noise/outliers. Step 2 filters out the points in the small clusters and noise/outliers and only keeps the large clusters as the training data set for the next step. Step 3 classifies the points from small clusters and noise/outliers with k-NN algorithm using the training data points from the large clusters. Step 4 enforces the must-link constraints by assigning all smart meters connected to the same lateral/transformer to the same cluster. The final clustering results will be returned in step 5.

Note that researchers have proposed alternative approaches, such as C-DBSCAN [17] to integrate constraints into density-based clustering algorithms. In the C-DBSCAN algorithm, the data points from different clusters involved in a must-link constraint are simply forced to merge together. However, when the preprocessed voltage time series are mapped to the low-dimensional space, we often encounter cases where a very small number of meters connected to one phase are spread over two clusters representing two phases. To address this issue, in step 4 of the proposed CHC algorithm, we only reassign all the data points connected by a must-link constraint to the same cluster without affecting the grouping of other data points.

C. Stage 3: Phase Identification for Clustered Customers

The final stage identifies the phase connection of the clusters determined in stage 2. This can be accomplished by performing field validations on a small number of samples of smart meters with phase measurement tools [3][4]. The cost associated with the field validation is minimal as the number of customers that require phase measurement is quite small. To achieve the highest accuracy, the small sample of customers should be chosen as close to the clusters' centers as possible. Depending on the availability of must-link constraints, two sampling strategies can be implemented:

1) If there are no must-link constraints, then in each cluster choose m smart meters that are closest to the cluster center. Field validations can then be performed on these m smart meters. The most frequent phase connection of these m meters is selected as the phase connection of all the customers in the cluster.
2) If must-link constraints are available, then in each cluster choose the group D_g that contains at least w customers and has the least mean squared distance to the cluster center. Field validations will be performed on any of the smart meters in group D_g. The phase connection of the group is selected as the phase connection of all the customers in the cluster.

 ISBN: 978-1-61208-588-3

IV. CASE STUDIES

A. Experimental Design

Two types of experiments are designed below to 1) examine the performance of the proposed phase identification algorithm and 2) explore the impact of smart meter data granularity on the phase identification accuracy.

The first set of experiments compare the performance of the constrained k-means clustering algorithm with linear dimensionality reduction [2] and the CHC algorithm with nonlinear dimensionality reduction proposed in this paper. The constrained k-means clustering algorithm with linear dimensionality reduction is referred to as "CK-Means" method. Both methods are evaluated over 18 hourly voltage time series gathered from 5 distribution feeders as described in Table I.

The second set of experiments evaluate the impact of smart meter sampling frequency on the accuracy of the proposed phase identification algorithm. The experiments are conducted over 6 voltage time series gathered from 3 distribution feeders. The smart meters on distribution feeder 1-3 were configured to record voltage magnitudes every 5 minutes. The average voltage magnitudes with hourly, 15-minute, and 5-minute granularity are used as inputs.

B. Parameter Selection

A few parameters need to be set up in order to run the proposed phase identification algorithm. In the feature extraction stage, three parameters from the t-SNE algorithm need to be selected. The dimensionality of the PCA output and t-SNE input d_{in} is set to be 30. The perplexity p is set to be 100. Note that these two parameters can be tuned by running the optimization several times on a data set and picking the parameters that yield the best map [9]. The dimensionality of the t-SNE output d_{out} is typically set to be 2 or 3. For better visualization, we set d_{out} to 2. In fact, the case study results with $d_{\text{out}} = 2$ and $d_{\text{out}} = 3$ are very similar.

In the proposed CHC algorithm, three key parameters *MinPts*, ϵ, and α need to be tuned first. The typical ranges for the three parameters are 8 to 20 for *MinPts*, 1 to 3 for ϵ, and 0.005 to 0.01 for α. When tuning these parameters, the aim is to see the data points in the t-SNE space being clustered appropriately. For example, assume we select some initial settings for *MinPts*, ϵ, and α, and get the clustering results as shown in Figure 5. Intuitively, cluster 11 and 15 should be two separate clusters. If the initial parameter setting merges these two clusters, then the parameters need to be tuned so that they are separated in the clustering results. In this particular case, we should decrease ϵ and/or increase *MinPts* to separate cluster 11 and 15. Note that ϵ is the radius of neighborhood and *MinPts* is the threshold for determining if a point p is a core point or a border point in a cluster. The parameter α controls the number of output clusters. If the value of α is too large, then the phase identification accuracy will be lower. However, if the value of α is too small, then a large number of meters need to be field validated, which increases implementation costs. k, the parameter of the k-NN, can be selected to be equal to *MinPts*. At last, in the field validation, choose the must-link group with at least $w = 20$ customers.

C. Performance of the Proposed Phase Identification Algorithm

The phase identification accuracies of the CK-Means method and the proposed phase identification algorithm are calculated based on field validation results. For the proposed algorithm, 30 runs of t-SNE are conducted. 10 runs with the lowest cost function values are kept. The average accuracy over the 10 runs are reported in Table II and Figure 4. As shown in the table, the proposed phase identification algorithm significantly outperforms the CK-Means method with all the data sets in terms of accuracy. On average, the proposed phase identification algorithm improves the identification accuracy by 19.81% over the CK-Means algorithm. Figure 4 shows that the improvement in phase identification accuracy varies by the unbalance level of the distribution circuit. The improvement is more significant for periods when the distribution feeder is less unbalanced.

The combinations of phase connections in the 5 testing feeders include 3 phase-to-neutral connections, 3 phase-to-phase connections, and a mix of all 6 possible connections. The accuracy of the proposed phase identification algorithm is very high under most cases. s_{13}, s_{14}, and s_{15} have relatively lower accuracy, because they have lower levels of unbalance and they have all 6 possible connections, which is more difficult to identify than other feeders. When the level of unbalance is higher, the accuracy is greatly improved in s_{16}, s_{17}, and s_{18}, whose accuracies are very decent for a feeder with all the 6 possible phase connections. Figure 5 illustrates the clustering result of data set s_{18} in the 2-dimensional t-SNE map, using the proposed phase identification algorithm. In the figure, each dot represents a smart meter. Figure 6 depicts the actual phase connection of each smart meter. By comparing

TABLE II. PHASE IDENTIFICATION ACCURACIES

Feeder	Data Set	Level of Unbalance	CK-Means Accuracy (%)	Proposed Algorithm Accuracy (%)
1	s_1	0.0785	81.21	93.06
	s_2	0.0776	81.18	93.62
2	s_3	0.0514	69.67	87.55
	s_4	0.0617	57.51	87.79
3	s_5	0.0956	54.91	83.94
	s_6	0.1019	72.78	82.83
4	s_7	0.1109	89.29	98.60
	s_8	0.1141	97.82	98.94
	s_9	0.1131	97.79	99.63
	s_{10}	0.1190	88.42	99.66
	s_{11}	0.1043	87.49	99.88
	s_{12}	0.1250	88.34	99.65
5	s_{13}	0.0673	29.80	73.18
	s_{14}	0.0668	38.80	73.32
	s_{15}	0.0705	59.07	67.01
	s_{16}	0.0742	40.56	88.19
	s_{17}	0.0846	60.49	87.11
	s_{18}	0.0842	52.02	89.84

 ISBN: 978-1-61208-588-3

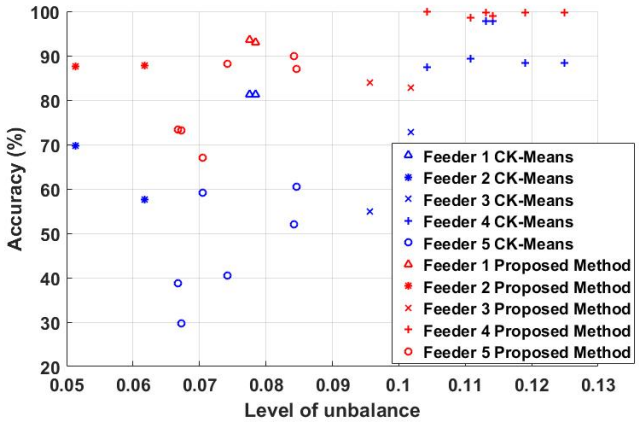

Figure 4. The phase identification accuracy with CK-Means and proposed algorithm.

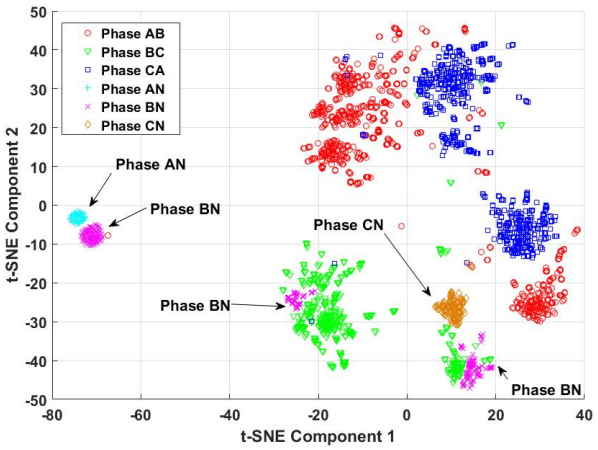

Figure 6. Field validated phase connections of data set s_{18} in the 2-dimensional t-SNE map.

Figure 5 and Figure 6, it is shown that the proposed phase identification algorithm not only groups phase-to-phase meters accurately, but also groups phase-to-neutral meters with a high accuracy. Cluster 2, 11, 12, 13, and 15 each represents one of the phase-to-neutral connections AN, BN, and CN, as indicated by the arrows in Figure 5 and Figure 6.

As a comparison, Figure 7 shows the distribution of smart meters from data set s_{18} in the 2-dimensional PCA map. The data points are not well separated according to phase connection. From Figure 7 and Figure 6, it is clear that the nonlinear dimensionality reduction technique, t-SNE, does a much better job in extracting hidden features from the voltage time series during a less unbalanced period for the feeders.

As shown in Figure 5, the clusters are in different sizes and shapes. Some of the clusters are non-convex. The proposed CHC algorithm has a great advantage in identifying clusters with such data point distributions. Figure 5 also shows how the must-link constraints could improve the phase identification accuracy. In the top right cluster 8, a few data points are linked together. Although a small number of the data points are located in cluster 14, they will eventually be assigned to cluster 8 due to the must-link constraint. From Figure 6, these data points should belong to cluster 8, which is connected to phase CA instead of phase AB.

D. Impact of the Smart Meter Sampling Frequency on the Phase Identification Accuracy

The phase identification accuracies of the proposed algorithm under 3 different meter reading granularity levels are calculated and summarized in Table III. It shows that as the granularity of meter readings increases from hourly to every 15 minutes and then 5 minutes, the phase identification accuracy increases. The average increase in the phase identification accuracy over the 3 distribution circuits is 3.36% when the meter reading granularity increases from hourly to 5 minutes. More granular voltage readings allow extractions of features/patterns

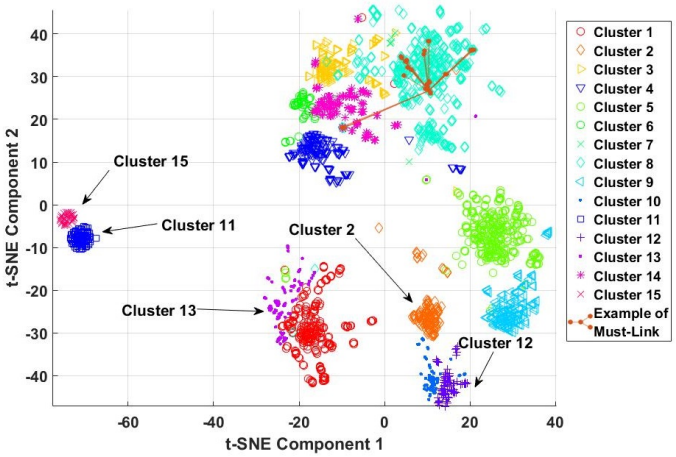

Figure 5. The clustering result in the 2-dimensional t-SNE map on data set s_{18}.

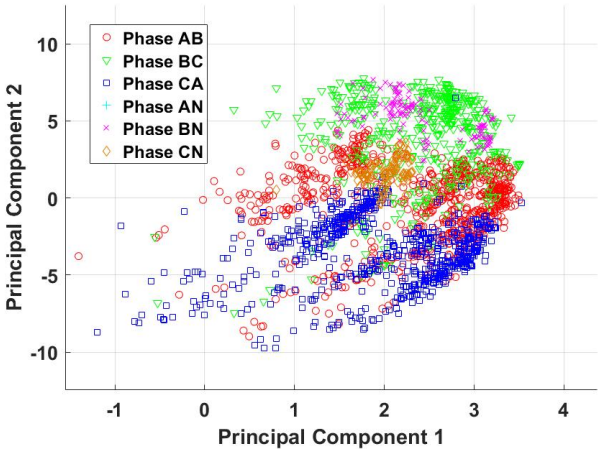

Figure 7. Field validated phase connections of data set s_{18} in the 2-dimensional PCA map.

 ISBN: 978-1-61208-588-3

TABLE III. IMPACT OF SAMPLING FREQUENCY ON THE PHASE IDENTIFICATION ACCURACY

Feeder	Data Set	Granularity of Meter Readings		
		1 hour	15-minute	5-minute
1	s_1	93.06%	93.93%	93.88%
	s_2	93.62%	94.32%	94.40%
2	s_3	87.55%	88.86%	92.03%
	s_4	87.79%	90.47%	89.93%
3	s_5	83.94%	90.02%	91.56%
	s_6	82.83%	84.51%	87.16%

that may not be present in coarse data sets. However, it should be noted that there are additional costs associated with gathering more granular smart meter data. Note that the phase identification accuracy decreases slightly for data set s_1 and s_4 when the sampling frequency increases from 15-minute to 5-minute. This is partly due to the randomness of the t-SNE mapping.

V. CONCLUSION

This paper develops an AMI data driven phase identification algorithm that addresses the drawbacks of the existing solutions. Compared to the existing solutions, the proposed algorithm has three main advantages. First, the proposed algorithm does not require prior knowledge about the number of phase connections in the distribution system. Second, the proposed algorithm works well with distribution feeders that have both phase-to-neutral and phase-to-phase connections. Third, the accuracy of the proposed phase identification algorithm is not very sensitive to the level of unbalance in a distribution feeder. Comprehensive field testing results on 5 distribution feeders show that the proposed algorithm significantly outperforms the existing methods. In addition, we discover that more granular voltage time series leads to higher phase identification accuracy.

In the proposed CHC algorithm, a few parameters need to be tuned manually. To implement the proposed AMI data driven phase identification algorithm on thousands of distribution feeders, we plan to develop an automatic parameter tuning algorithm.

ACKNOWLEDGMENT

The authors would like to thank Austen D'Lima, Joshua Davis, and Tom Martin from Southern California Edison and the Pacific Gas and Electric Company for fruitful discussions and supplying AMI, network connectivity, and field validation data.

REFERENCES

[1] A. Gupta and B. Kellison, "Utility AMI analytics at the grid edge: Strategies, markets and forecasts," GTM Research, Tech. Rep., 2016.

[2] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, "Phase identification in electric power distribution systems by clustering of smart meter data," in *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE, Dec. 2016, pp. 259–265.

[3] L. A. Pomatto, "Apparatus and method for identifying the phase of a three phase power line at a remote location," Apr. 23 1996, US Patent 5,510,700.

[4] W. S. Bierer, "Long range phasing voltmeter," Oct. 5 2010, US Patent 7,808,228.

[5] M. Dilek, "Integrated design of electrical distribution systems: Phase balancing and phase prediction case studies," Ph.D. dissertation, Virginia Polytechnic Institute and State University, 2001.

[6] V. Arya *et al.*, "Phase identification in smart grids," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*. IEEE, Oct. 2011, pp. 25–30.

[7] T. A. Short, "Advanced metering for phase identification, transformer identification, and secondary modeling," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 651–658, Jun. 2013.

[8] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart meter data analytics for distribution network connectivity verification," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1964–1971, Jul. 2015.

[9] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.

[10] J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 100, no. 5, pp. 401–409, May 1969.

[11] P. Demartines and J. Hérault, "Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 148–154, Jan. 1997.

[12] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[13] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, 2003, pp. 857–864.

[14] E. Z. Macosko *et al.*, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, May 2015.

[15] A. Frome *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, 2013, pp. 2121–2129.

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

[17] C. Ruiz, M. Spiliopoulou, and E. Menasalvas, "C-DBSCAN: Density-based clustering with constraints," in *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*. Springer, May 2007, pp. 216–223.

 ISBN: 978-1-61208-588-3