Machine Learning for Inferring Phase Connectivity in Distribution Networks

Sambaran Bandyopadhyay, Ramachandra Kota,
Rajendu Mitra, Vijay Arya
IBM Research
{sambandy, rama.chandra, rajendum, vijay.arya}@in.ibm.com

Brian Sullivan, Richard Mueller,
Heather Storey, Gerard Labut
DTE Energy, USA
{sullivanbj, muellerrj, storeyh, labutg}@dteenergy.com

Abstract—**The connectivity model of a power distribution network can easily become outdated due to system changes occurring in the field. Maintaining and sustaining an accurate connectivity model is a key challenge for distribution utilities worldwide. This work focuses on inferring customer to phase connectivity using machine learning techniques. Using voltage time series measurements collected from customer smart meters as the feature set for training classifiers, we study the performance of supervised, semi-supervised and unsupervised techniques. We report analysis and field validation results based on real smart meter measurements collected from three feeder circuits of a large distribution network in North America.**

I. Introduction

The connectivity model (CM) of the physical power distribution network specifies how the devices, assets, and customers are interconnected downstream of a distribution substation. For example, which customer is powered by which distribution transformer, which customer is powered by which *phase* of the feeder, and so on. CM essentially provides a coarse-grained view of the network topology. A common problem faced by distribution utilities worldwide is an inaccurate or unknown CM of their network when compared to the actual connectivity that exists on the field. The CM may not always be updated or tracked based on changes made by field crews and its accuracy deteriorates over time due to maintenance, repairs, and restoration activities following faults or outages. Moreover during large scale outages, there is often a trade-off between expediting restoration versus tracking changes being made to the distribution network.

While CM is foundational to planning, operations, and maintenance of distribution networks, the key factors driving utilities to improve its accuracy are effectively faster restoration and the ability to accurately communicate with impacted customers during outages. The annual cost of power interruptions in US is estimated to be $79B with 106 ± 54 outage minutes per customer. Interruptions in electric service occur from time to time due to a number of reasons including storms, aging assets, excess loading from heat waves, and other system disturbances. Any analysis following a fault in the distribution system uses the CM to identify root causes and determine the appropriate course of action. An accurate CM minimizes diagnostic time and the crew time in the field, leading to reduced outage minutes and improved system availability [1].

During outages, utilities seek to inform customers about the status of restoration and the expected downtimes. The CM is required to localise customers downstream of a faulted device and to map each fault with the right set of outaged customers for communication. An inaccurate CM increases the risk of erroneous communication and limits a utility's ability to provide customized and timely information to their customers, which may negatively impact customer relations. Additionally, an accurate CM is required to localise losses, estimate loading at unmetered points such as distribution transformers, and ensure a balanced infusion of energy back into the distribution grid when customers have behind-the-meter resources such as distributed generation and electric vehicles/storage [2, 3, 4].

Existing techniques to maintain an accurate CM include the use of manual field inspections and power line communications (PLC). Manual inspections are expensive and unsustainable as field configurations change over time and therefore these need to be repeated periodically. The PLC approach requires meters to be able to read and write signals onto the power-line and is capital intensive. Additionally, challenges arise as signals may not propagate over long distances or across assets.

Leading utilities are undertaking initiatives to modernize their information management and data analytics capabilities in order to realize the benefits of smart grid deployments. This work focuses on the use of data already available from a distribution network to infer its *phase connectivity model* i.e. which customer is powered by which phase of a given feeder. Prior work has proposed analytics techniques that require voltage measurements from both customer smart meters as well as feeder meters to infer phase [5]. In practice however, not all feeders are instrumented with meters, which limits the applicability of such techniques. Instead, our approach relies solely on voltage measurements collected from smart meters in combination with any existing but partially accurate customer to phase mapping. Customer phase is inferred using learning techniques that can either predict phase based on a training set or update an inaccurate customer to phase mapping. We consider supervised, semi-supervised, and unsupervised learning techniques and report experimental as well as field validation results based on analysis of data collected from customers of a large distribution network in North America. The data includes average RMS values of voltage recorded once every 5 minutes from more than 5K customers across 3 feeders for about 2 months. The performance of techniques is estimated by comparing the computed phase mapping with the ground truth obtained through manual field inspections and also the prior customer phase as per the utility's database.

Use of machine learning techniques to infer customer phase has a number of advantages:
1. During restoration efforts following storms and outages, the crew may alter the phase in certain segments of the feeder to restore power to customers. Therefore voltage data from customers under *unaffected* segments of the feeder may be

used to train a classifier that can predict the new phases of affected customers.
2. By training a classifier using an existing yet partially accurate customer to phase mapping, learning techniques may be used to identify inconsistencies and correct the existing mapping. This is useful in operational settings as each restoration affects a limited set of customers. Therefore the existing mapping may be automatically updated with the help of subsequent voltage measurements.

We view the main contributions of this work as follows:
1. A novel approach is proposed that infers customer to phase mapping using machine learning techniques. The algorithms infer customer phase solely from voltage measurements collected from customer smart meters in combination with existing but partially accurate customer to phase mapping.

2. We compare the performance of different learning approaches: support-vector machines (SVM), label propagation, and K-MEANS algorithm initialized using existing customer to phase mapping. We show that these techniques may be used to both predict the unknown phase of customers using a small training set or update an existing but inaccurate phase connectivity model. We observe that while SVMs yield the highest accuracy, the K-MEANS algorithm is more robust to errors in the existing connectivity.

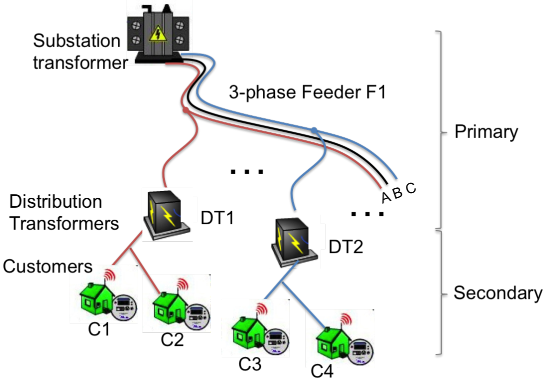
3. The methods are outlined along with empirical and field validation results based on real voltage measurements collected from customers of a large distribution network in North America. Our results indicate that despite inaccuracies in the existing customer to phase mapping, the algorithms generally yield an updated mapping of higher accuracy.

The rest of the paper is organized as follows. Section II explains the topology of a distribution network and its connectivity model. Section III describes our approach to infer phase using SVM, label propagation, and K-MEANS respectively. Section IV and V present empirical results while Section VI presents results from field validation. Finally, Section VII describes prior work before we conclude in Section VIII.

II. Distribution Network & Phase Connectivity

Electric power is generated in large power plants as 3-phase AC voltage and reaches distribution via a transmission system. The distribution system starts from the distribution substation and consists of primary and secondary networks. The primary network consists of 3-phase *feeders* which carry power at medium voltage from the substation transformers to the *distribution transformers* (DTs). A 3-phase feeder consists of three transmission lines, usually labeled as A, B and C, which carry AC power with their voltage waveforms shifted by 120^o. A DT receives power by tapping onto one of the 3 phases of a feeder and is generally single-phase (1-phase). On average a DT might serve about $8-10$ single-phase residential customers. A few DTs that serve larger loads such as super-markets or office buildings may be 2 or 3-phase. [4].

The CM of a distribution network specifies the mapping between various assets and customers downstream of a substation. In particular, which customer is powered by which DT, which DT is powered by which feeder and phase, and so on. For example, the CM of the distribution system in Fig. 1



Customer	Transformer	Phase
$C1, C2$	$DT1$	A
$C3, C4$	$DT2$	C

Fig. 1. Simplified view of a distribution network and customer to phase mapping

records that customers $C1, C2$ are powered by distribution transformer $DT1$, while $C3, C4$ are powered by $DT2$. $DT1$ is powered by phase A of feeder $F1$, while $DT2$ is powered by phase C of $F1$. In the above example, since the distribution transformers are single-phase (1-phase), customer phase is same as the transformer phase.

Phase Connectivity Errors. The most common errors in the CM are related to the mapping between customers and phases of the feeder. For example, the phase of a customer or distribution transformer may be recorded incorrectly or missing in the database. This work focusses on identifying and correcting these errors using voltage data from smart meters.

A. Data Sources

Power measurements in a distribution network are generally available from customer smart meters and on-grid sensors. Customer smart meters generally record periodic measurements of load (kWh) and voltage (RMS value) over small time intervals of 5 to 30 min as setup by the utility. Instrumentation of on-grid sensors or feeder meters (SCADA devices) depends on the level of monitoring enabled in the distribution network. These are not always available and not all feeders may have been metered.

In this work, we consider voltage measurements from customers belonging to 3 anonymized feeder circuits from two substations of a North American distribution network. Each circuit is fed at 13.2kV and serves more than 2K residential, commercial and industrial customers. For each feeder circuit, our dataset comprises of the following:

Smart meter data: We have two months of voltage measurements from customers which are powered by 1-phase overhead (poletop) distribution transformers. These are average RMS values of voltage in the 120V range, which are recorded once every 5 minutes up to a precision of one decimal place. The measurements also have missing values.

Customer to Phase Connectivity: The phase of overhead 1-phase distribution transformers is available before and after manual field inspections which we denote as the *database solution* (possibly inaccurate) and the *ground truth* (100% accurate) respectively. Customer phase is same as transformer phase in case of 1-phase transformers. In Sections IV & V,

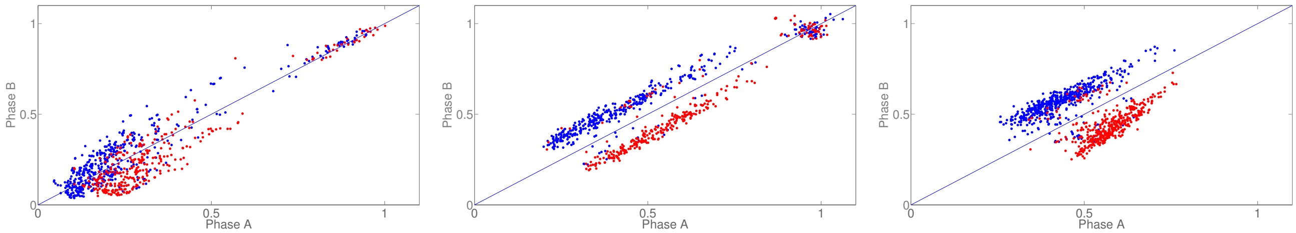


Fig. 2. Cluster scatter plots showing the partition of customers based on phase. (left): *Raw* dataset, (center): *Delta* dataset and (right): *Binary* dataset

we test our results against the ground truth. In Section VI, we use both the database solution as well as the ground truth.

III. Machine Learning Approaches

Our approach to infer phase connectivity leverages the fact that the voltage variations observed by customers powered by the same phase tend to be more "similar". However, there are different ways of comparing voltage measurements. In the following, we compare 3 data transformations and then describe the different machine learning approaches. Our goal is to essentially classify customers into three groups and label these groups as A, B, and C corresponding to three phases, using customer voltage measurements as the feature sets.

A. Voltage Data Transformations

Let $v_{i,t}$ denote the raw RMS voltage measurements from customer i in the t^{th} time step, $t \in \{1, \ldots, m\}$. We compute the continuous voltage fluctuations as $\delta_{i,t} = v_{i,t} - v_{i,t-1}$. We also discretize these to obtain binary fluctuations, which we denote by $b_{i,t}$. Let $\mathbf{V}_i = [v_{i,t}]_{m\times 1}$ denote the m-dimensional *observation vector* that holds the time series of voltage measurements obtained from customer i. Similarly, let Δ_i and $\mathbf{B}_i$ denote the delta and binary observation vectors respectively. We assume that voltage measurements are approximately synchronised across customers.

Thus for each customer i, we have three datasets: (i) *Raw*: original data V_i, (ii) *Delta*: continuous fluctuations Δ_i, and (iii) *Binary*: discretized fluctuations $\mathbf{B}_i$. In order to identify which among the three forms of data is best suited to be used as the feature set, we consider the simple problem of partitioning customers belonging to two phases (A and B) of a circuit into 2 groups. Towards this, we apply the K-MEANS algorithm [6,7] on each dataset using correlation distance as the measure of similarity between the customers. Fig. 2 shows the benchmark cluster scatter plots for two phases A and B, which compare the computed clusters with the ground truth. The plots show higher accuracy for *Binary* when compared to *Raw* or *Delta* datasets. Thus, in the rest of the work, we use the *Binary* dataset as the feature set for classification.

B. Support Vector Machines (SVM)

Support Vector Machine is considered to be one of the best supervised learning models available in state-of-the-art machine learning literature [8,9]. Given a labeled training set $\{\mathbf{x_i}, y_i\}$, $i = 1, 2, \cdots, n$, $\mathbf{x_i} \in \mathbb{R}^d$, and $y_i \in \{-1, +1\}$, SVM learns an optimum hyperplane that separates the positive from the negative samples and maximizes the margin between the two classes. Given a hyperplane characterized by its direction $\mathbf{w}$ and position b, SVM learns these parameters by solving the following optimization problem

$$\min \frac{||\mathbf{w}^2||}{2} + C\sum_i \xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i \qquad (1)$$

where $\xi_i, \forall i = 1, 2, \cdots n$ are slack variables introduced to relax the optimization as the dataset may not always be linearly separable. Once the training is complete and the parameters $\mathbf{w}$ and b are learnt, it is possible to classify any unlabeled point $\mathbf{y_i}$ as $+1$ if $\mathbf{w}^T\mathbf{y_i} \geq +1$; and -1 otherwise. The above approach is extended for multi-class classification i.e. when there are more than two labels in the data set, by "one-against-one" or "one-against-all" [10] methods. For our experiments, we use the LIBSVM package [11].

C. Label Propagation

Semi-supervised learning approaches are those in which the learning mechanism takes advantage of both the labeled and the unlabeled data for classification problems. Label propagation is a well known semi-supervised learning technique. In this work, we use the algorithm proposed by Raghavan et. al. [12] which uses a graph-structure based approach for label propagation. In addition to being a simple algorithm that is easy to visualise, its main advantages over other label propagation approaches include the fact that it only uses the network structure for guidance and does not require optimisation or an objective function. The algorithm begins by assigning a label to each nodes initially. At every iteration, every node takes up that label which occurs the most among its neighbours, thereby using the network structure to identify labeling of the nodes. The iterations continue until every node has the same label as the majority of its neighbours.

We used the R implementation which allowed the provision to initialize and fix the labels of some of the nodes (equivalent to providing a training set). In our setting, the nodes represent customers while edges represents similarity between them. As described before, we use the binary correlations of the voltage data as the similarity distance. However, using the correlation matrix directly to generate the graph would result in a completely connected graph between the nodes, thus being of no help to the label propagation algorithm. Instead, we used a threshold value to generate the unweighted graph. Between every two nodes $i, j \in C$ where C is the set of customers, undirected edge E_{ij} exists *iff* $1 + corr_{ij} \geq \eta$, where $corr_{ij}$ denotes the binary correlation of i & j as described above and η is a numerical threshold for the circuit.

The next requirement is to determine η. Empirical observation has shown that a higher threshold produces better accuracy for the learning approach. A benchmark plot is shown in Fig. 3. Here, we used 25% of data as the training set (fixed initial labels) for label propagation over each of the

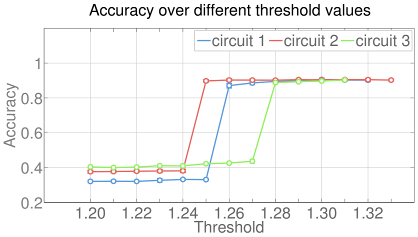


Fig. 3. Accuracy of label propagation over increasing threshold values

three circuits. The accuracy obtained for the different ranges of the threshold is plotted for the three circuits. We see that accuracy improves with increasing threshold. However, the graph becomes disconnected at higher thresholds (e.g. 1.33 for circuit 1) and label propagation cannot be used then. Therefore, the threshold has to be determined so as to avoid making the graph disconnected. We use a minimum spanning tree based method to set the threshold that achieves this balance. Our method of using label propagation is as follows:

1) Generate a completely connected, undirected graph G with the set of customers as nodes N and and for every pair of nodes i,j, edge E_{ij} has weight $w_{ij} = 1 + corr_{ij}$.
2) Find the minimum spanning tree MST of graph G.
3) Assign threshold $\eta = max(w_{ij}), E_{ij} \in MST$.
4) $\forall i, j \in N$, set $w_{ij} = 0$ if $w_{ij} < \eta$ or set $w_{ij} = 1$ otherwise.
5) Run the label propagation algorithm on the resultant undirected, unweighted graph G.

D. K-MEANS

To represent unsupervised learning approach, we used the widely-used and robust clustering algorithm K-MEANS [6] K-MEANS algorithm finds a solution that minimizes the within-cluster distances:

$$\underset{\ell_1,\dots,\ell_k}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \ell_i} d(\mathbf{x}_j - \mathbf{C}_i) \tag{2}$$

where $\mathbf{C}_i$ is the centroid of observation vectors in cluster ℓ_i.

In all our experiments, the major goal is to infer the phase for the customer meters. As K-MEANS in general is an unsupervised technique, it would not be able to provide phase labels to customers directly. To address this problem, we use the phase connectivity present in the potentially inaccurate data to initialize the centroids of K-MEANS. More precisely, for each phase, we calculate the mean of the voltage vectors corresponding to that phase from the inaccurate connectivity present in the utility's database, and initialize the centroid (corresponding to that phase) by the mean vector.

IV. Training on Partial Connectivity

In this section, we discuss the experiments which simulate scenarios wherein accurate phase connectivity is available for a subset of customers within a feeder circuit. As described in Section I, such a situation arises when maintenance activities in a part of the circuit may affect customer to phase mapping, while the rest of the feeder is unaffected.

For each circuit, we conduct Monte Carlo experiments over varying training size. In each run, a subset of the customers are randomly selected and used as the training set, while the rest of the customers form the test set. While this division is self-explanatory for SVM, in the case of label propagation, this means that the nodes (i.e. customers) belonging to the training set are assigned fixed labels while those in the test set are not given any labels. With both methods, the accuracy denotes the proportion of customers in the test set that are assigned correct labels (as compared with their true labels)[1].

Fig. 4 plots the accuracy achieved by the two methods with training size varying from 5% to 85% of the overall dataset. The error bars show the standard deviation of results. We see that both the methods perform very well across the range of training sizes. Particularly, SVM method achieves 90% accuracy for 5% training size and that increases to almost 100% for 85% training size for two circuits. However, label propagation performs close to 90% accuracy for 5% with only marginal improvement with increasing training size.

The above experiments were conducted using the whole dataset spanning two months. In the next set of experiments, we study the sufficiency of data for such performance. For this, we divide the dataset into batches spanning a few days and conduct experiments for each batch just as before, and average the results. For ease of presentation, Fig. 5 plots the results for the two methods for the two extreme settings of training sizes – 5% and 85%. We see that for all cases, and all circuits, there is only a marginal improvement in performance beyond a batch size of 4 days. Another finding is that label propagation shows higher variance in its performance across batches. Thus it is more sensitive to the size of the dataset (due to missing data, some batches have lower number of data-points than others for a given batchsize).

V. Improving Inaccurate Connectivity

In this section, our experiments simulate the second major use case – the existing phase connectivity in the utility's database is inaccurate due to undocumented changes, however, it is not known as to which customers have inaccurate phase labels. To recreate this situation, we introduce errors in the accurate customer to phase mapping by randomly switching the phase of a subset of customers. This partly accurate customer to phase mapping covering all customers is then used to train the 3 classifiers – SVM, Label propagation and K-MEANS. Moreover, all customers are then considered in the test set as well. Accuracy is computed based on the proportion of customers assigned the correct phase when compared with the true phase labels. For these experiments, we vary the proportion of customers with false labels from 10% to 60%.

Fig. 6 shows the results of Monte Carlo experiments for the three circuits and the three methods. The plots show that K-MEANS is very robust and performs at the same level despite increasing inaccuracy in the dataset. Label propagation also demonstrates similar robustness but only up to 50% inaccurate connectivity. Beyond that, its performance deteriorates rapidly and it also shows higher variance. On the other hand, the performance of SVM is strongly correlated with inaccuracy and results in a gradual decline. It is interesting to see that the three learning methods produce the same accuracy when 40% of labels are inaccurate for all the three circuits.

[1] K-MEANS being an unsupervised approach, was not applied for this completely supervised task.

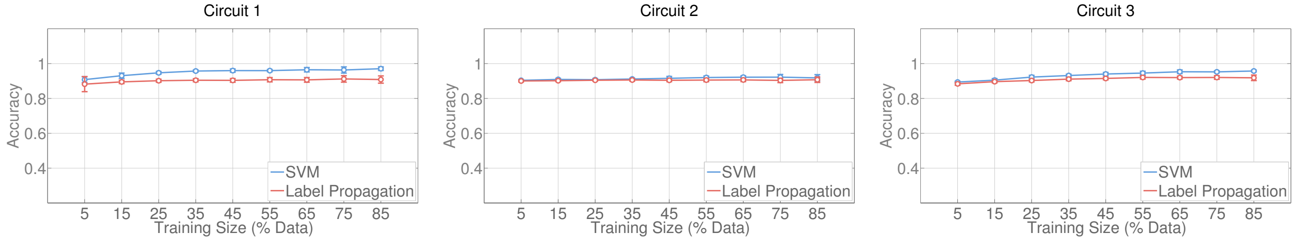


Fig. 4. Accuracy of SVM and Label Propagation for different training set sizes for the three circuits

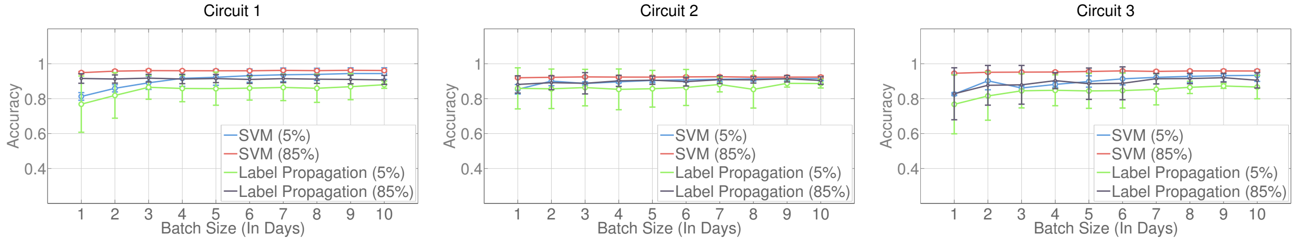


Fig. 5. Accuracy of SVM and Label Propagation over varying batch sizes of the data for the three circuits

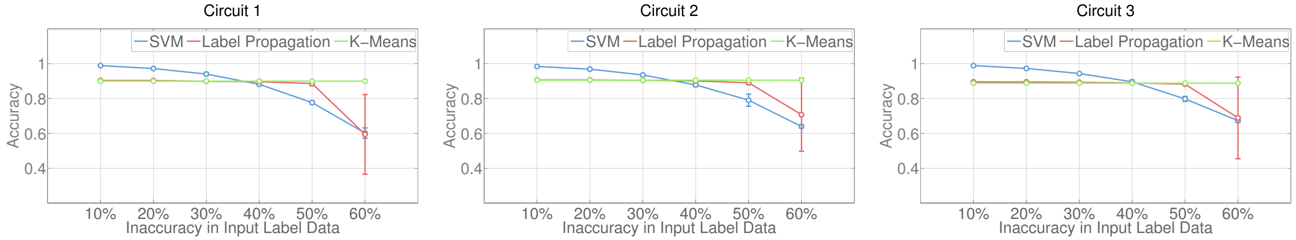


Fig. 6. Accuracy of rectifying a given labelling over increasing inaccuracy of the input phase labelling

Circuits	**DB Accuracy w.r.t GT**	**SVM**		**Label Propagation**		**K-MEANS**	
		Accuracy w.r.t GT	**Accuracy w.r.t DB**	**Accuracy w.r.t GT**	**Accuracy w.r.t DB**	**Accuracy w.r.t GT**	**Accuracy w.r.t DB**
Circuit 1	86.33%	89.29%	88.72%	88.93%	87.45%	89.98%	88.27%
Circuit 2	59.61%	61.71%	70.79%	59.77%	72.92%	29.11%	55.10%
Circuit 3	72.89%	82.71%	80.13%	87.68%	76.12%	88.81%	74.19%

TABLE I. FIELD VALIDATION RESULTS – WITH RESPECT TO GROUND TRUTH (GT) AND DATABASE SOLUTION (DB)

VI. FIELD VALIDATION

In the previous section, we introduced random errors into the accurate (field validated) phase connectivity to simulate erroneous scenarios. In this section, we conduct experiments using the actual inaccurate customer to phase mapping obtained from the utility's database before field inspections. Therefore, this setting gives further indication of the performance of the methods in a realistic setting and demonstrates the ability of the methods to correct an inaccurate customer to phase mapping.

As described in Section II-A, we use the *database solution* that denotes the old customer to phase mapping available in the utility's database, to train the 3 classifiers – SVM, label propagation and K-MEANS. We then compare the labels generated by these methods against the field validated *ground truth*. We conduct experiments using the dataset of discretized binary voltage measurements as before. The computed phase labels are then compared against the ground truth to determine accuracy, as shown in table I. The computed labels are also compared against the old database solution (DB). The accuracy of the database solution with respect to ground truth is also shown for reference.

It can be observed that for circuits 1 and 3, the accuracy of label propagation and K-MEANS w.r.t ground truth are similar to what was seen in Section V. However, the performance of SVM varies significantly. Moreover, for circuit 2, all of the three methods perform poorly when compared to their respective simulation counterparts in Section V.

The discrepancy between the results of simulated inaccurate input and real inaccurate input arise due to the kind of inaccuracies present. In Section V, the noise or error in the input labeling is uniform and independent across the phases (the connectivity was made inaccurate by randomly switching phases). However, in the database, the inaccuracies were generated because the wrong phase was reported at the transformer level, so all the customers connected to the transformer were assigned that wrong phase. Hence, the labeling noise present in the database is biased and correlated, thus different to simulated noise. The accuracies for circuits 1 and 3 is still as good as simulated noise in case of label propagation and K-MEANS because these two methods are robust towards noisy input. SVM is less robust to biased noise because of its high dependency on support vectors during the training phase [13]. Furthermore, the performance for circuit 2 is poor across the three methods. On inspection, we found that in the database solution, the class size ratios were highly skewed i.e., almost all transformers (hence customers) connected to one of the phases (as per ground truth) were being reported as connected to another phase in the database. As there were very few samples for one of the phases, none of the methods were able to improve the connectivity.

Table I also shows the comparison of the labeling of the methods with the old database solution (DB). It can be observed that for all of the methods, the accuracy values with respect to the database solution are consistent with the accuracy of the database w.r.t ground truth. Therefore, the methods can be used to provide a quality check of a circuit's existing customer to phase mapping. That is, given an existing mapping for each circuit, the methods can compare and rank circuits on the basis of the accuracy of their existing mapping.

VII. Related Work

This section summarizes prior work on inferring customer phase. Caird [14] discloses a system and method for phase identification with suitably enhanced meters that can detect phase based upon a unique signal injected into the phase line. The disadvantage of signal injection methods or those that rely on power-line communication (PLC) is that they require enhanced hardware to transmit and receive signals at different points of the grid, increasing capital and maintenance costs. Our approach on the other hand simply relies on voltage measurements already available from customers smart meters, and requires no specialised hardware like repeaters.

Prior work [15], [16] has proposed optimization techniques that can infer customer phase using a time series of customer and feeder load (kWh) measurements. Voltage-based techniques have two advantages over load-based techniques. First, load-based optimization techniques are based on the principle of conservation and therefore their accuracy may reduce in the presence of non-AMI or unmetered loads. The voltage-based approaches however are insensitive to unmetered loads in the system. Second, the load-based optimization approaches require both customer and feeder measurements while voltage-based techniques proposed in this work need only customer smart meter measurements.

In [17], authors analyse measurements from a low voltage network consisting of a 3-phase distribution transformer and its customers. Voltage measurements of individual customers are matched with the per-phase measurements taken at the distribution transformer to assign phase. Our prior work [18] analyses voltage measurements from a microgrid to infer different connectivity relationships. In [5], we show that low voltage measurements from customer meters as well as medium voltage measurements from feeders may be simultaneously correlated to infer phase. On the other hand, our current work relies solely on low voltage measurements from smart meters and studies different machine learning methods to infer phase.

VIII. Conclusions

An accurate connectivity model is required in the planning, operations, and maintenance of distribution networks. It enables faster restoration, accurate and timely communication with impacted customers during outages, and is also needed to efficiently integrate distributed generation and behind the meter resources. Automated inference of customer to phase mapping using data already available from smart meters allows utilities to infer, validate and maintain their phase connectivity without resorting to expensive manual field inspections.

In this work, we study the performance of machine learning approaches for this problem through support vector machines, label propagation and k-means as representatives of supervised, semi-supervised and unsupervised techniques. We present experimental and field validation results based on real smart meter measurements from a large North American distribution network. Our analyses shows that SVM provides the highest accuracy for settings where accurate phase connectivity is known beforehand for a subset of customers, whereas label propagation and k-means may be more robust to noisy labels. All methods improved the accuracy of the existing customer to phase mapping for two of the three feeder circuits analysed. Future work will focus on studying machine learning techniques particularly suitable for noisy labels for the setting of inaccurate connectivity. Similarly, we also seek to focus on developing supervised and semi-supervised techniques for inferring customer-transformer connectivity and study the impact of distributed generation on customer voltage measurements.

References

[1] G. Clark, "A changing map: Four decades of service restoration at alabama power," *Power and Energy Magazine, IEEE*, vol. 12, pp. 64–69, Jan 2014.

[2] J. Bouford and C. Warren, "Many states of distribution," *Power and Energy Magazine, IEEE*, vol. 5, no. 4, pp. 24–32, 2007.

[3] J. Fan and S. Borlase, "The evolution of distribution," *Power and Energy Magazine, IEEE*, vol. 7, no. 2, pp. 63–68, 2009.

[4] J. D. D. Glover and M. S. Sarma, *Power System Analysis and Design*. Pacific Grove, CA, USA: Brooks/Cole Publishing Co., 3rd ed., 2001.

[5] R. Mitra, R. Kota, S. Bandyopadhyay, V. Arya, B. Sullivan, R. Mueller, H. Storey, and G. Labut, "Voltage correlations in smart meter data," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015 (to appear).

[6] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations.." Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, pp. 281-297, 1967.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[9] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.

[10] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[12] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E*, vol. 76, p. 036106, 2007.

[13] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[14] K. Caird, "Meter Phase Identification." US Patent Application 20100164473, January 2010. Patent No. 12/345702.

[15] M. Dilek, *Integrated Design of Electrical Distribution Systems: Phase Balancing and Phase Prediction Case Studies*. PhD thesis, Virginia Polytechnic Institute and State University, 2001.

[16] V. Arya, T. S. Jayram, S. Pal, and S. Kalyanaraman, "Inferring connectivity model from meter measurements in distribution networks," in *ACM e-Energy*, pp. 173–182, 2013.

[17] H. Pezeshki and P. Wolfs, "Correlation based method for phase identification in a three phase lv distribution network," in *Universities Power Engineering Conference (AUPEC), 22nd Australasian*, pp. 1–7, 2012.

[18] V. Arya and R. Mitra, "Voltage-based clustering to identify connectivity relationships in distribution networks," in *SmartGridComm*, pp. 7–12, 2013.