An Integrated Data Mining Framework for Analysis and Prediction of Battery Characteristics

Marjan Momtazpour
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA

Ratnesh Sharma
NEC Laboratories America, Inc.
Cupertino, CA, USA

Naren Ramakrishnan
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA

***Abstract*—Batteries play an important role in modern sustainable energy systems. However, batteries are expensive and have a limited life time. Having a deep understanding of how batteries operate in working situations is crucial to designing advanced control mechanisms. Battery performance and life time is highly dependent on how it is used and also on environmental working conditions. While batteries have been extensively studied through model-based approaches, there is no previous work about modeling behavior based on data analytic methods. In this paper, we propose an integrated data-driven framework to study the behavior of battery systems in a grid, based on data mining techniques. The proposed method provides a high level characterization of battery behavior and online parameter estimation using supervised and unsupervised learning methods. This work can be used in intelligent control systems and would help administrators to know what is happening inside a battery system.**

I. Introduction

Energy storage systems like batteries provide flexibility in use of generation resources and management of demand, e.g., in demand shifting, peak shaving and efficient operation of energy resources [1]. However, efficient operation of batteries to maximize lifetime and optimize efficiency requires a good understanding of battery charging models, data from modern measurement tools, and capturing the effects of battery usage and environmental conditions on performance.

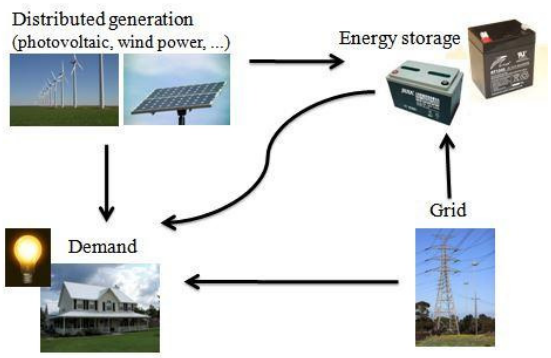

Fig. 1. Structure of micro-grid

Here we propose a data mining approach to model and optimize stationary batteries based on available information (e.g. current, voltage, and depth of discharge of batteries). The operating characteristics of the batteries can be understood by defining the states of our system and drawing state transition diagrams. Results would be valuable for storage administrator regarding performance and optimization of system. We use the available data taken from measurement units to estimate the remaining life of the battery in terms of efficiency and capacity using regression methods. Furthermore, the pattern of battery usage is discovered which results in a more accurate estimation battery efficiency. The ultimate goal is to use these results to develop more intelligent control strategies and improve the efficiency of whole battery energy storage system.

II. Previous works

Previous studies on battery systems can be roughly categorized in two groups: estimating state of charge (SOC) and estimating state of health of the battery (SOH).

SOC estimation is a challenging task in batteries. SOC describes battery's remaining capacity which is an important parameter for control strategy [2]. A good estimation of SOC can protect the battery, prevent overdischarge, and also improve the battery life and moreover, allows application to make rational control strategies to save energy [3]. SOC estimation methods fall into two categories: direct computational methods and intelligent ones. In direct computational methods, SOC is directly calculated based on the relationship between battery parameters such as open-circuit voltage or internal impedance. Alternatively, approaches such as [4] use Kalman filters to estimate SOC of lithium-ion batteries in electric vehicles with the use of an equivalent circuit model of the battery.

Overcharging, overdepleting, and other reasons can damage the battery. Also, battery operation is dynamic and its performance varies with its age. State of health of battery (SOH) captures the ability of the battery to store and deliver energy. Two common methods for calculating SOH involve battery impendance or battery power, and

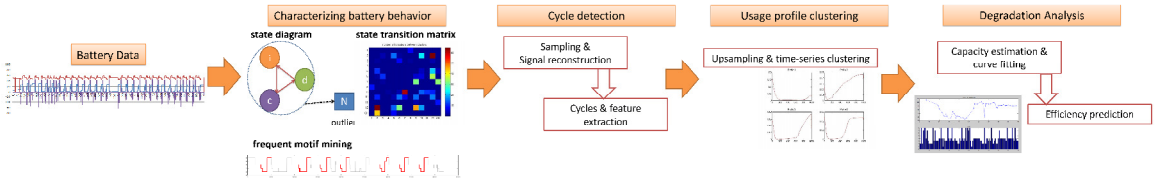

Fig. 2. The overall framework

using battery capacity [5]. Typical SOH methods characterize battery power or energy. In [5], new SOH estimation methods are proposed based on battery energy represented by Ampere-hour throughput (Ah).

III. Dataset and System Model

In a typical micro-grid infrastructure, the battery operates in order to minimize the costs (minimize energy from utility) in cooperation with generators such as solar PV and diesel generators [6]. In this study we use a dataset which contains information gathered from a microgrid for over a year (provided by NEC Labs). The database has information about the battery for that year (from February-2012 to April-2013) with the sampling rate of one sample per 20 seconds. Several parameters were logged such as terminal voltage, current, and Depth of Discharge (DoD) of the valve-regulated lead-acid battery (VRLA). DoD determines the level of charge of the battery (DoD of 100% means it is empty).

In order to identify the states of the battery, we calculated two other features: Ah_{in} and Ah_{out}. Ah_{in} indicates how much of energy is saved in the battery during a specific charging period while Ah_{out} is an indicator for the amount of energy which is discharged from the battery. To calculate Ah_{in} and Ah_{out}, the sample time immediately after termination of a full charge is considered as a reference point. Reference point is the time that we start our calculations. First of all, in complete charging events, the value for Ah_{in} and Ah_{out} would be set to zero as a reference point. Starting from the reference point, if battery is getting partially charged ($DoD \neq 0$), Ah_{in} would increase and if battery is idle or discharged, Ah_{out} would increase. Ah_{in} is calculated based on the following equations:

$$Ah_{in} = \int_{t \in PartialCharge} I \, dt \qquad (1)$$

Ah_{out} is calculated similar to Eq. 1 except that integral is taken over discharge period.

Due to the nature of devices, sometimes measurement tools cannot log data in a regular sampling rate. Hence, we might miss some valuable data or we may encounter variations in sample rate or different sampling rates in different devices. However, in order to have a higher accuracy and consistency in our experiments, we have to make sure of having a consistent sampling rate. Therefore, measurements are pre-processed to identify the missing values and also alleviate the sampling rate to a static degree.

Our proposed framework is illustrated in Figure 2. At the first step, the behavior of the battery is characterized in terms of state diagrams and matrix transitions. States of the battery are derived using clustering methods. After that, in this phase, sets of frequent sequence of state transitions are detected based on frequent episode mining algorithm [7]. At the second phase, after data alignment step, the cycle of charging followed by discharging events are detected and the relations between different parameters of state are illustrated. At the next step, a time-series clustering algorithm is applied on all cycle profiles to group similar profiles in terms of usage. At the end, capacity and efficiency estimation are compared with and without the help of profile clusters.

IV. Battery Characterization

A. State detection

It is well-known that batteries have three major states: charging, discharging, and being idle. However, experiments show that several states can be recognized based on the history of battery (amount of energy remained in battery) and its current operating status. As an example, battery can be in bulk phase charging (charging with constant current) while it was completely depleted beforehand. Hence, different states can be identified such as charging while it was depleted, idle while it was full, discharging while it is almost empty, and so on. Identifying these states helps us to understand the behavior of battery more precisely, which in turn, helps us to design more accurate control strategies. In this section, we provide a framework for battery state detection based on unsupervised learning methods.

In order to detect states of the battery we applied clustering methods on the dataset without considering the time-dependencies between data points. Here, we use K-means and density-based clustering algorithm (DBSCAN) on two sets of features. At first, DBSCAN is applied on DoD, current, and voltage of battery. Clustering results show that by DBSCAN algorithm, data points are categorized into 8 different clusters. Furthermore, K-means (with K=3) is applied on the difference between the normalized values of Ah_{in} and Ah_{out}. Prototypes of each cluster is illustrated in Figures 3 and 4.

Each data point belongs to one cluster of DBSCAN and one cluster of k-means. Combination of these two clustering approaches can be considered as the final state of each data point. Experiments show that while we have 24 possible combinations of clusters (3 K-means clusters

TABLE I
STATES OF BATTERY WITH THEIR AVERAGE STAY TIME AND STATE PROBABILITY.

State ID	DBSCAN label	k-means label	Average stay time(h)	State probability
1	1	1	4.34	0.1741
2	1	2	2.77	0.2316
3	1	3	0.09	0.4112
4	2	1	1.51	0.0582
5	2	2	0.69	0.0344
6	2	3	0.76	0.0421
7	3	2	0.07	0.0253
8	4	1	0.01	0.0074
9	4	2	0.07	0.0001
10	4	3	0.01	0.0019
11	5	2	0.95	0.0097
12	6	2	0.64	0.0006
13	7	2	2.99	0.0028
14	8	2	0.54	0.0005

and 8 DBSCAN ones), for the studied battery, only 14 combinations occur. Each of these 14 combinations is considered as a state of battery and are shown in Table I. The last two columns of this table show the average stay time and the steady state probability of each state, respectively. The whole state-diagram of the battery is illustrated in Figure 5. In this figure, arrows show the available transitions from each state and the value on each arrow represents the probability of that transition.

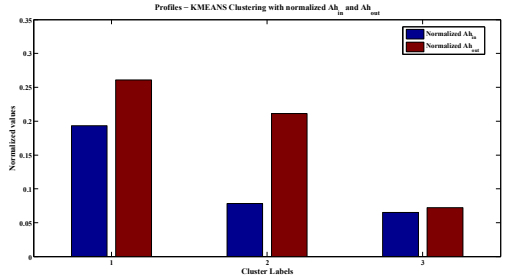

Fig. 3. Profiles resulted from k-means clustering with difference of normalized Ah_{in} and Ah_{out}.

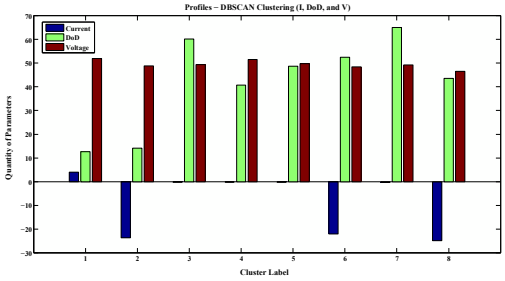

Fig. 4. Profiles resulted from DBSCAN clustering with current, DoD, and voltage.

B. Frequent motif mining

In the last subsection, we determined battery states and state-diagram. We use state IDs of Table I to encode states as symbols for further analysis. Hence, we encode a multivariate series of states as a stream of symbols. In this stream, times where state transitions occur are considered as points of interest. We use motif mining algorithms to extract frequent patterns of states.

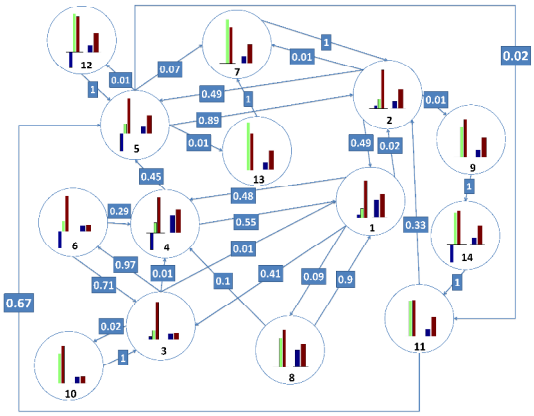

Fig. 5. State diagram of the battery. Numbers are the probability of transition from one state to another.

The idea of frequent motif mining is to identify the contiguous sub-sequence of patterns which are repeated several times in the time series. In our work, we look for a sequence of state transitions and we aim to know what the frequent sequences are. In this part, state transitions are considered as an event in a timely manner. This process is the Apriori-like algorithm that iteratively generates candidates and counts their frequencies. This algorithm can accommodate "don't care" states and this is beneficial in discovering of hidden patterns. The algorithm has been described in details in [7]. The output of the algorithm is a set of frequent episodes for a given frequency threshold.

For the studied battery we use states of Table I to build event sequences. We set inter-event time constraint to 24 hours (one day) and the minimum frequency threshold to 35. Two sets of frequent motifs are shown in Figure 6. Here, the first sequence contains state transitions from 1 to 4 and subsequenntly from 4 to 1. The second motif shows transitions $3 \rightarrow 6$, $6 \rightarrow 4$, $4 \rightarrow 5$, $2 \rightarrow 1$, and $1 \rightarrow 3$. The first sequence spans 1.84h and the second one spans for 11h. The frequency of first sequence is 39 and for the second sequence this frequency is 36.

V. LIFE OF BATTERY

In this section, we study how data mining can help us to identify the changes in capacity and degradation in efficiency. In order to study the life parameters of the battery including the capacity and efficiency of battery, we need to expand our view to cycles rather than individual time-steps. In the following subsections, we first define a cycle and provide an algorithm for cycle detection. In the next step, we use clustering to categorize battery cycles based on DoD. Then, we study efficiency degradation for each cycle based on the information we extracted from the clustering results.

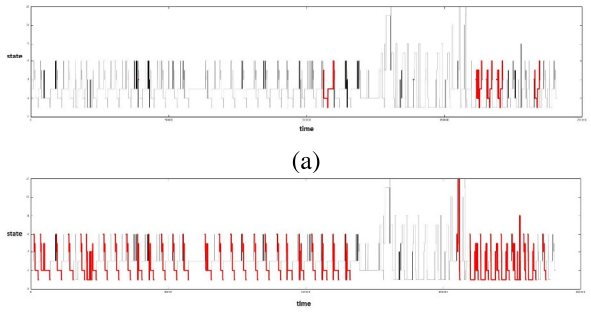

Fig. 6. Examples of frequent motifs: (a) sequence of 1-4,4-1. (b) sequence of 3-6,6-4,4-5,2-1,1-3.

A. Cycle detection

In order to study the parameters of battery, the ideal way is to measure the battery in open circuit model. Since removing battery from the system and measuring its voltage is not always applicable, especially when battery is working in a micro-grid and system relies on that, we need to look at available cycles of charge and discharge. Since, the battery is not fully charged after each charging period and it is not fully discharged in corresponding discharge periods, we are not able to estimate the true capacity of the battery for all the cycles. However, if we estimate the capacity for those cycles that have incomplete charge/discharge periods, the value can give us an idea of how capacity was degraded over time. The important thing that should be taken into consideration is that if we want to compare the resulted values for different cycles, DoD values before and after these events should be similar. The parameters of different cycles (with different DoDs) can be converted to a reasonable ratio for further comparison. We define a cycle as a charging event including a fully charged battery (DoD=0) followed by combination of discharge and idle events (current<0).

Cycle detection is performed with Algorithm 1. This algorithm, at first finds the full charging points in the dataset. Full charging points are the times when battery is in idle condition (no charging and no discharging occure) and DoD is zero. For each of these points it expands its range from both directions in time backward and forward. At first, it determines the starting point where charging event starts (t_s) and then it looks for the ending point when the next charging event starts (t_e). Due to the nature of the data, sometimes we encounter a missing value which may alter the final results. Hence, before this step, signal reconstruction step is done in which it construct the missing value based on linear interpolation of data. In order to compare time-series with different length, a time alignment step is done to make their sampling rate even (to 1 sample per minute). At the end of this algorithm, a refining step prunes the irregular events (e.g. pruning a cycle of complete charging event followed by a long idle event).

Algorithm 1: Cycle detection algorithm

Input: Current(I), Depth of Discharge (DoD) time series, Thresholds Ie and De, Time step Δ_t.
Output: Set of complete cycles (C)

Reconstruction and Interpolation for noise removal
$FullChargeSet \longleftarrow$Find set of full charge points $(I < Ie \, and \, DoD < De)$
$C = \varnothing$
while $FullChargeSet \neq \varnothing$ **do**
  $t \longleftarrow FullChargeSet(1)$
  $ts \longleftarrow t$
  while $I(ts) > 0$ **do**
    $ts \longleftarrow ts - \Delta_t$
  end
  $te \longleftarrow t$
  while $I(te) < IeDO$ **do**
    $te \longleftarrow te + \Delta_t$
  end
  $C \longleftarrow C \cup \{(ts, te)\}$
  for $t \in FullChargeSet$ **do**
    if $t \in (ts, te)$ **then**
      $FullChargeSet \longleftarrow FullChargeSet - t$
    end
  end
end
Pruning incomplete cycles
return C

After finding cycles, several features are extracted which specify the characteristics of each cycle. These features include duration of absorption phase, maximum current at the time of charging, cut-off current, real DoD before and after charging, and charging and discharging capacity.

Cut-off current is current of battery right before charging event stops. It is also worth mentioning that charging event contains two consecutive phases: bulk phase and absorption phase. In bulk phase charging, current is constant while in absorption phase charging, voltage remains constant until it shuts off after a minimum current (cut-off current) is reached.

Here, we assume that when battery is idle, terminal voltage measurements of battery is almost similar to the value of open-circuit voltage (OCV). Based on this, DoD at time t, DoD_t can be derived as the complement of state of charge (SOC), SOC_t.

For each time, when voltage is v_t, SOC_t is calculated based on linear interpolation on the values in Table II.Values of V and SOC in Table II are derived from factory datasheets of battery.

Capacity at charging Cap_{ch} periods is derived as follows:

$$Cap_{ch} = \int_{t \in I^-} I * dt \tag{2}$$

where I^- is the set of times where current is negative. Capacity at discharging period (Cap_{disch}) is determined similarly when current is positive.

In order to compare capacities, the DoD before charging, after charging, and after discharging must be scaled to one value. Here, we convert all parameters to have DoD of 40% ($DoD_r = 40$). For example for charging capacity, the converted capacity is as follows:

TABLE II
SOC AND OCV OF VRLA RECHARGEABLE BATTERY (246AH)

SOC	V	SOC	V
100	13.180	40	12.289
90	12.932	30	12.150
80	12.808	20	12.003
70	12.682	10	11.842
60	12.554	0	11.620
50	12.423	-	-

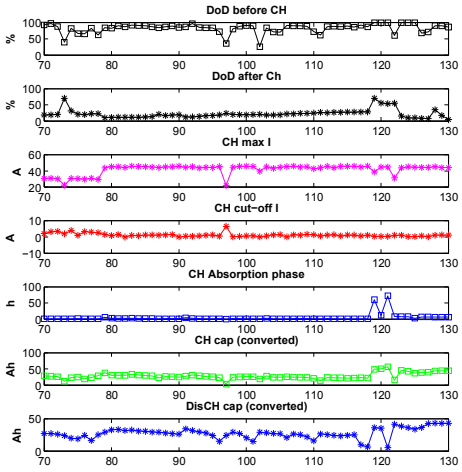

Fig. 7. Different features of cycles through time. From top to the bottom: DoD at one step before charging starts, DoD after charging is complete, maximum current during charging, cut-off current in charging period, duration of absorption phase in charging (hour), charging capacity, discharging capacity

$$ChargingCap_r = Cap_{ch} * \frac{DoD_r}{DoD} \tag{3}$$

Figure 7 shows how the extracted features affect on the charging and discharging capacity. For example, near cycle 80 (in day 80), the charging capacity jumps up which is due to the increase in maximum current of charging. Also, in 120th cycle, the absorption phase increases and lead to a high jump in capacity of battery. It is obvious that the way we use our battery has a great effect on the capacity and life of the battery. Hence, the profile of usages must be specified to have a better understanding of these effects.

In addition to the above analysis, a polynomial curve fitting ($p(x) = p_1x^2 + p_2x + p_3$) is applied on capacity of battery (discharge) for different time ranges. To see the slope of degradation, we need to study those cycles that have been derived under the same condition. Hence, these cycles are divided into two based on Figure 7. Figure 9 shows that the degradation is faster when absorption phase was lower. This re-emphasizes an essential characteristic of VRLA (Valve-Regulated Lead Acid) battery.

B. Clustering profile of usage

For the purpose of characterizing the profile of usage on battery, we deployed a time-series based clustering algorithm (k spectral clustering) [8]. The distance metric in K-SC clustering algorithm makes the results invariant to scale and translation (shift):

$$\hat{d}(x, y) = \min_{\alpha,q} \frac{||x - \alpha y_{(q)}||}{||x||} \tag{4}$$

where $y_{(q)}$ is a shifted time series y (by q time unit) and α is scaling coefficient.

K-SC algorithm uses k-means algorithm with the above distance metric. Details of this algorithm is described in [8]. A large scale version of this algorithm, called incremental K-SC is used in our paper which utilizes the benefit of discrete Haar wavelet transform.

Figure 8 shows the resulting prototypes of clustering algorithm. In our method, the DoD values are considered as a metric to specify the usage profiles. As this figure depicts, the usage patterns vary within the time. The first group consists of cycles where after charging, battery remains idle and then discharged a little bit. Second group are the profiles where discharging event occurs smoothly. In group 3 and 4, idle event and discharging event occurs where discharging rate is different between these two groups. Since not all the cycles have a same duration, a down-sampling method is applied to scale all cycles to the same duration.

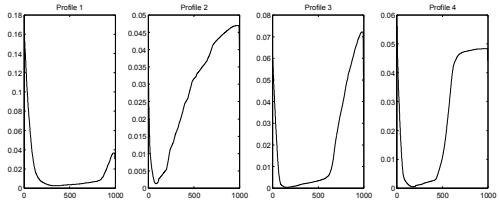

Fig. 8. prototypes of usage profile.

C. Degradation analysis

In battery life cycle, battery degradation plays an important role and before making any decision about the battery control strategy, battery degradation should be addressed precisely. In this section, we use regression analysis to study degradation in terms of battery efficiency.

Here, we applied regression to predict the efficiency of our battery. Efficiency can be calculated for each cycle as follows:

$$\eta = \frac{E_{out}}{E_{in}} = \frac{\int_{t \in Discharge} I.V\, dt}{\int_{t \in Charge} I.V dt} \tag{5}$$

Before calculating η, values of E_{in} and E_{out} are multiplied by $\frac{DoD_r}{DoD}$ where DoD_r is a reference value.

Let us assume that we have a time series of battery efficiencies, measured based on Eq. 5. In order to estimate the efficiency at time t, $\eta(t)$, we define a window of size m that covers m consequent efficiencies, $\eta(t-m), ..., \eta(t-1)$. Then based on the following regression method we perform the estimation:

$$\eta(t) = a_0 + \sum_{1}^{m} a_i \eta(t - i) \tag{6}$$

where, a_is are constant coefficients. In the experiments we use the window size of $m = 2$. Using training data, we are able to estimate a_is.

In addition to the above regression task, we studied the effect of cluster profiles from previous subsection on the accuracy of our regression results. In the second regression problem, we estimate efficiency based on efficiency at earlier sample times and K-SC profiles based on the following equation:

$$\eta(t) = C + \sum_{1}^{m} a_i \eta(t-i) + b_i P(t-i) \tag{7}$$

where, $\eta(t)$ is efficiency at time t, $P(t)$ is K-SC profile at time t, and C, a_i, b_is are constant coefficients determined through the training process.

After training the regression algorithms, we can use resulted coefficients to predict efficiency of battery in the future (or on test dataset). To measure the accuracy of our methods, the following relative error metric is used:

$$error = \frac{|\eta(t) - \hat{\eta}(t)|}{\eta(t)} \tag{8}$$

where, $\eta(t)$ is efficiency at time t and $\hat{\eta}(t)$ is the predicted efficiency at that time.

The experiment results are shown in Table III. As it is obvious from this table, by using the usage profiles, the accuracy of efficiency estimation is improved by more than %20 (Figure 10).

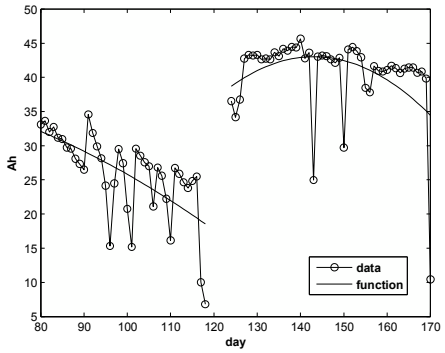

Fig. 9. Polynomial Curve fitting in discharging capacity

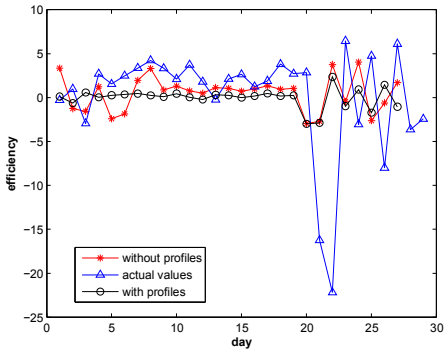

Fig. 10. Regression result on test set

TABLE III
RELATIVE ERROR OF REGRESSION

	Without profiles	with profiles
Relative Error	36%	9%

VI. CONCLUSION

Energy Storage systems like batteries are key to grid modernization and implementation of smart grid technologies. Performance of batteries highly depends on their working conditions. However, existing model based methods for battery behavior study, do not consider details of battery usage profile. In this paper, we proposed a data-driven approach to study the behavior of batteries while they are in the circuit. This approach helps us to model the battery based systems more precisely. We developed an integrated solution to identify similar usage patterns and deploy these information in efficiency estimation of batteries. Experiments with real datasets show that the proposed method effectively improves the accuracy of the predictions. This work can be used in intelligent control systems and would help the administrator to know what is happening in battery. However, applied techniques are generic (technology-agnostic) and can be used for non-battery technologies too. As a future work, the whole battery storage system including power electronics can be considered for efficiency calculations. Developing more intelligent control strategies and considering all tools in micro-grid as an integrated unit for further analysis can be studies further.

ACKNOWLEDGMENT

We would like to thank Babak Asghari and Anupama Keely for providing the data and their valuable comments and feedback.

REFERENCES

[1] W. Y. Chang, "State of charge estimation for lifepo4 battery using artificial neural network," *International Review of Electrical Engineering*, vol. 7, no. 5, pp. 5874–5800, 2012.

[2] H. W. He, R. Xiong, and H. Q. Guo, "Online estimation of model parameters and state-of-charge of lifepo4 batteries in electric vehicles," *Applied Energy*, vol. 89, no. 1, pp. 413–420, 2012.

[3] W.-Y. Chang, "The state of charge estimating methods for battery: A review," *ISRN Applied Mathematics*, vol. 2013, 2013.

[4] C. Zhang, J. Jiang, W. Zhang, and S. M. Sharkh, "Estimation of state of charge of lithium-ion batteries used in hev using robust extended kalman filtering," *Energies*, vol. 5, no. 4, pp. 1996–1073, 2012.

[5] D. Le and X. Tang, "Lithium-ion Battery State of Health Estimation Using Ah-V Characterization," in *Annual Conference of the Prognostics and Health Management Society*, 2011.

[6] A. Hooshmand, B. Asghari, and R. Sharma, "A Novel Cost-Aware Multi-Objective Energy Management Method for Microgrids," in *2013 IEEE PES Innovative Smart Grid Technologies*, 2013.

[7] D. Patnaik, M. Marwah, R. K. Sharma, and N. Ramakrishnan, "Temporal data mining approaches for sustainable chiller management in data centers," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 4, pp. 34:1–34:29, Jul. 2011.

[8] J. Yang and J. Leskovec, "Patterns of Temporal Variation in Online Media," in *ACM WSDM'11*, 2011, pp. 177–186.