Appliance signature identification solution using K-means clustering

Post-print deposited in the CityU Institutional Repository, City University of Hong Kong.

Citation:

© 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Appliance Signature Identification Solution using K-means Clustering

Department of Electronic Engineering
City University of Hong Kong
Hong Kong, China

Abstract—Sustainability, energy conservation and demand response have become an inevitable concern around the world. In the light of electricity companies’ demand about what electric appliances that the end-users are switching on, appliance signature is suggested to increase the performance of demand response. The main idea behind appliance signature is that it utilizes the characteristic that same types of electric appliances should have similar features like current, power and harmonic distortion. Utility can get not only the energy profile of households with current metering system but also acquires evidence in energy management according to the energy usage pattern. In this paper, K-means clustering is used for the classification of eight types of common household electric appliances which is an appliance signature identification solution for appliance signature. A digital Butterworth filter has been firstly introduced to remove noisy data before data analyzing by K-means clustering. The performance is evaluated by 10-fold cross validation. Three indexes, CH index, DB index and SH index have been calculated to determine the optimal number of clusters used in K-means clustering. These indexes achieve accuracy of 55.5%, 42.1% and 67.7% respectively.

Keywords—Appliance Signature; K-means Clustering; Demand Response; Classification

I. INTRODUCTION

With the increasing number of world population and utilizing of electric goods, the annual electricity consumption has been increased to 19,090 billion kWh in 2012. The average growth rate is 4.2% from 2003 [1]. The amounts of pollutants generated by production of electricity can wreak havoc on environment and human beings. Besides energy experts, environmental groups and researchers, countries like United Kingdom, Canada, Germany, USA and Singapore, have aspired to find any possible means to tackle with this urgent issue. Lots of researchers have been discussed the reliability, security, communication of advanced metering infrastructure (AMI) for smart grid as an infrastructure for reduction of electricity consumption as a result of energy audit [2]-[5]. For energy audit, some researchers focus on specific electric appliances, lighting devices [6] and air conditioners [7]. Reference [8] discusses ZigBee communication protocol in wireless sensor network for energy audit. The third solution is appliance signature. Appliance signature is a new research area in recent decade which focuses on classification of electric equipment for demand response (DR) and energy management by analyzing electricity consumption depleted by specific types of appliances. DR consists of two parts, home area network and utility. In home area network, the major components are in-home display, sockets, appliances and gateway. For the utility, it consists of a smart meter, power line network and power plant. The total energy consumption of appliances is collected to smart meter through home gateway. The reading is sent to the utility. After analyzing, if heavy loading is concluded by utility, a control signal will be sent back to the end-users in order to switch off the appliance. Basically, an agreement should be signed between utility and end-users. Appliance signature can help to identify which appliances the users are using so that utility will not switch off the unwanted appliances. Appliance signature is similar to pattern recognition. However, pattern recognition is usually referred to macroscopic term which can include various targets. For instance, human writing, transaction data, climate and energy consumption are all belonged to pattern recognition but not appliance signature. Researchers have proposed various methods for appliance signature like threshold detection [9], event-window [10], selective acquisition [11], state-transition [12] and K-means clustering [13]. In this paper, an appliance signature identification solution (ASIS) using K-means clustering is proposed for appliance signature. The reason for choosing unsupervised K-means clustering is that it cannot be guaranteed that same types of appliance has very close features. Active power, reactive power, total harmonic distortion and maximum transient current (rms) have been used as feature vector. Also, it provides additional digital filter, Butterworth filter for filtering before clustering. This paper is organized as follows: Section II describes the theory of ASIS. The feature vector extraction is discussed in section III. Performance evaluation of ASIS is presented in section IV. Lastly, conclusion is made.
II. THEORY OF ASIS

A. K-means clustering

In clustering algorithm, centre-based is an efficient type algorithm for large number and multi-dimension data sets. K-means clustering is one of the most popular algorithms in statistical clustering attributable to its simple and time efficiency characteristics. It is classified as non-hierarchical or partitioned clustering. The number of clusters k has to be specified at the beginning of the algorithm. The basic concept of K-means algorithm is for any given value of k, the remaining data is allocated to the nearest clusters and this step repeats according to the error function until the error function does not varying much. Note that if the value of k equals to one or exactly the same as number of data points in the data set, K-means algorithm is meaningless. For k equals to 1, there should be no cluster. For k equals to the number of points in data set, it is trivial solution by using K-means because every data point can form a single cluster by itself. Let D be the data set with n data points, i.e. \( D = \{x_1, x_2, \ldots, x_n\} \), and assume \( C = \{C_1, C_2, \ldots, C_k\} \) be the collection of all k non-overlapping clusters. The conventional K-means algorithm by Hartigan in 1975 [14] and the detail are illustrated as follows: The error function of K-means algorithm is defined as

\[
E = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu(C_i))
\]

(1)

where \( \mu(C_i) \) is the centroid of cluster \( C_i \) and \( d(x, \mu(C_i)) \) is the distance between x and centroid of cluster \( C_i \). In normal practice, the default distance is chosen to be Euclidean distance as

\[
d_{\text{euclidean}}(x, \mu(C)) = \left[ \sum_{i=1}^{d} (x_i - \mu(C)_i) \right]^{1/2}
\]

(2)

for d-dimensional of data point.

By [15], eight steps have been summarized for the K-means algorithm as follows:

(i) The initial partition of data set \( D = \{C_1, C_2, \ldots, C_k\} \) for \( k > 1 \) and iteration phase;
(ii) Repeat;
(iii) Let \( d_{ij} \) be the distance between cluster j and case i;
(iv) Let \( n_j = \arg \min_{i \in j \leq k} d_{ij} \);
(v) The case i is assigned to cluster \( n_j \);
(vi) The mean of cluster is recomputed. There may be some changes of clusters in above steps;
(vii) A complete iteration is obtained until no further changes of cluster membership;
(viii) K-means algorithm is finished and result is obtained.

K-means algorithm can be adopted as an optimization problem which minimizes the objective function. The detail is explained as follows:

Firstly, the data set D is defined as \( D = \{x_i, i = 1,2, \ldots, n\} \) with n data points and integer k. Then, the objective function can be expressed as

\[
P(W, Q) = \sum_{i=1}^{k} \sum_{j=1}^{n} w_{il} d_{\text{euclidean}}(x_i, q_l)
\]

(3)

where \( Q = \{q_l, l = 1,2, \ldots, k\} \) is a set contains k objects, and \( W \) is an \( n \times k \) matrix that fulfils two criteria:

(i) \( w_{il} \in \{0,1\} \) for \( i = 1,2, \ldots, n, l = 1,2, \ldots, k \);
(ii) \( \sum_{i=1}^{k} w_{il} = 1 \) for \( i = 1,2, \ldots, n \).

By [16], the optimization problem is solved by (4) and (5).

\[
P(W, Q) \text{ is minimized}
\]

\[
\iff \ w_{il} = \begin{cases} 1, & d_{\text{euclidean}}(x_i, q_l) = \min_{l \in \{1,2, \ldots, k\}} d_{\text{euclidean}}(x_i, q_l) \\ 0, & \text{otherwise} \end{cases}
\]

(4)

where \( i = 1,2, \ldots, n \) and \( l = 1,2, \ldots, k \).

\[P(W, Q) \text{ is minimized}
\]

\[
\iff q_l = \frac{\sum_{i=1}^{n} x_i w_{il}}{\sum_{i=1}^{n} w_{il}}
\]

(5)

where \( l = 1,2, \ldots, k \) and \( j = 1,2, \ldots, d \).

B. CH index, DB index and SH index

Since the optimal number of clusters k has to be specified for K-means clustering, CH index, DB index and SH index are calculated [13]. Therefore, the performance of K-means algorithm composes of three numbers based on different indexes.

CH index is calculated by

\[
CH = \frac{\text{SSB}/(M-1)}{\text{SSW}/(N-M)}
\]

(6)

where SSB is the sum square between cluster variance and SSW is the sum square within cluster variance. Also, N is 164 which is the total number of data sets and M is number of clusters.

DB index is obtained by

\[
R_{ij} = \frac{S_i + S_j}{d_{ij}}, i \neq j
\]

(7)
where \( d_{ij} = \|C_i - C_j\|^2 \), \( S_i = \frac{1}{n_j} \sum_{j=1}^{n_j} \|x_j - C_i\|^2 \), and

\[
R_i = \max_{j=1,\ldots,M} R_{ij}, \quad i = 1, \ldots, M
\]  \hspace{1cm} (8)

Finally,

\[
DB = \frac{1}{M} \sum_{i=1}^{M} R_i
\]  \hspace{1cm} (9)

SH index is found by

\[
a(x_j) = \frac{1}{n_m} - \sum_{j=1}^{n_m} \|x_i - x_j\|_2^2 \quad \forall i, j \in C_m
\]  \hspace{1cm} (10)

\[
b(x_j) = \min_{i} \left\{ \frac{1}{n_m} \sum_{j=1}^{n_m} \|x_i - x_j\|_2^2 \right\} \quad \forall i, j \in C_m
\]  \hspace{1cm} (11)

and

\[
s(x_j) = \frac{b(x_j) - a(x_j)}{\max(a(x_j), b(x_j))}
\]  \hspace{1cm} (12)

Therefore,

\[
SH = \frac{1}{N} \sum_{i=1}^{N} s(x_i)
\]  \hspace{1cm} (13)

III. FEATURE VECTOR EXTRACTION

From Fig. 1, both current and voltage waveforms are measured and collected with a floppy disk via cathode-ray oscilloscope (CRO). The data is passed into computer for further analysis. The data is first performs noise filtering. Then, four parameters, active power, reactive power, total harmonic distortion and maximum transient current (rms) are calculated mathematically which form the feature vector of K-means clustering. The number of clusters for K-means clustering is based on CH index, DB index and SH index. All these parameters are used for training data set of ASIS. Finally, unknown new data set can be classified with the trained ASIS algorithm. Current transformer (CT) with ratio 15/5 is used to measure the instantaneous current waveform of electric appliance. Simultaneously, the instantaneous supply voltage of 220Vrms is measured with differential probe using CRO (TEK TDS3012). Noted that both current and voltage waveforms must be captured simultaneously. Otherwise, significant error arises for active and reactive power. For current waveform, transient and steady state waveforms are measured. The transient current is defined as the first 0.2s once the electric appliance is turned on [13]. The waveforms are saved with floppy disk. The resistor \( R_1 \) is chosen to be large value of 10MΩ because it has negligible effect to the circuit. The resistor \( R_2 \) is connected in series with the secondary winding of CT to avoid open circuit. The circuit diagram of measurement of two waveforms is shown in Fig. 2.

There are eight types of electric equipment commonly used in household being recorded for analysis of ASIS. The types and number of samples are listed in Table I. They are mobile charger, notebook, fan, iron brush, electric stove, hair dryer, light bulb and vacuum cleaner. 24 samples are taken for hair dryer whereas 20 samples are taken for remaining seven types of appliances. Thus, the training data set of the ASIS is 164.
TABLE I. TRAINING DATA SET OF ASIS.

<table>
<thead>
<tr>
<th>Name of electric appliance</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile charger</td>
<td>20</td>
</tr>
<tr>
<td>Notebook</td>
<td>20</td>
</tr>
<tr>
<td>Fan</td>
<td>20</td>
</tr>
<tr>
<td>Iron brush</td>
<td>20</td>
</tr>
<tr>
<td>Electric stove</td>
<td>20</td>
</tr>
<tr>
<td>Hair dryer</td>
<td>24</td>
</tr>
<tr>
<td>Light bulb</td>
<td>20</td>
</tr>
<tr>
<td>Vacuum cleaner</td>
<td>20</td>
</tr>
</tbody>
</table>

The collected data is filtered with 6th order Butterworth filter. Fig. 3 shows the current waveform of fan and hair dryer.

After filtering the noise of current and voltage waveforms by Butterworth filter, the computation of feature vector of ASIS is needed. The feature vector composes of active power, reactive power, maximum transient current (rms) and total harmonic distortion. The formula of active power is given by:

\[ P_{\text{active}} = V_{\text{rms}} \times I_{\text{rms}} \cos \phi \]  \hspace{1cm} (14)

Also, the equation of reactive power is governed by

\[ P_{\text{reactive}} = \sqrt{(V_{\text{rms}} \times I_{\text{rms}})^2 - (V_{\text{rms}} \times I_{\text{rms}} \cos \phi)^2} \]  \hspace{1cm} (15)

Maximum transient current (rms) as defined in [13] is the maximum value of current (rms) in the first 10 signal periods after the appliance is turned on. It is computed as

\[ \text{Current}_{\text{transient (rms)}} = \max_{15 \leq i \leq 10} \{I_{\text{rms}}(i)\} \]  \hspace{1cm} (16)

Total harmonic distortion (THD) is measurement of the harmonic distortion as the ratio of sum of all powers in all harmonic components (except fundamental harmonic component) to the fundamental power of the first harmonic. It can be calculated as

\[ \text{THD} = \frac{P_2 + P_3 + \ldots + P_n}{P_1} \]  \hspace{1cm} (17)

Therefore, the four-dimensional feature vectors of ASIS are obtained as shown in Fig. 4. For visualization of the feature vector, multi-dimension scaling [15] is utilized which is shown in Fig. 5. It visualizes as three-dimension for better illustration.

![Fig. 4. Feature vector of ASIS.](image)

![Fig. 5. Three-dimension feature space visualization.](image)

IV. PERFORMANCE EVALUATION

Performance evaluation of proposed ASIS is carried out using 10-fold cross validation. 10-fold cross validation is used to verify the accuracy of proposed ASIS. The idea of cross validation of 164 data sets is as follows:

(i) Select 147 data sets as training of ASIS and the remaining 17 data sets as testing data set. The testing data sets are composed with 2 data sets from each type of appliances except hair dryer (3 data sets instead).

(ii) An accuracy of testing data set is compared to the ideal class of the appliance.

(iii) One of the fold has completed.

(iv) Repeat (i)-(iii) four times but once the data sets have used as testing data sets, it cannot be used again for testing data sets in the remaining fold of cross validation.

(v) Select 148 data sets as training of ASIS and the remaining 16 data sets as testing data set. The testing data sets are composed with 2 data sets from each type of appliances.

(vi) An accuracy of testing data set is compared to the ideal class of the appliance.

(vii) One of the fold has finished.

(viii) Repeat (V)-(vii) six times but once the data sets have used as testing data sets, it cannot be used again for
testing data sets in the remaining fold of cross validation

(ix) 10-fold cross validation has completed

Feature vectors of K-means algorithm compose of active power, reactive power, maximum transient current (rms) and total harmonic distortion. Since the optimal number of clusters k has to be specified for K-means clustering, CH index, DB index and SH index are calculated. Therefore, the performance of K-means algorithm composes of three numbers based on different indexes. The results of CH index, DB index and SH index for varying number of clusters k (1-20) is tabulated in Table II. The optimal numbers of clusters using CH index, DB index and SH index are 18, 15 and 20 respectively. Largest value of CH index is the best to choose. Smallest value of DB index is the best to choose. Largest value of SH index is the best to choose. Table III-IV shows that accuracy of each fold of cross validation and each appliance respectively. The accuracy using CH index, DB index, SH index are 55.5%, 42.1% and 67.7% respectively.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>CH index</th>
<th>DB index</th>
<th>SH index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>2</td>
<td>0.158</td>
<td>0.118</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>0.142</td>
<td>0.145</td>
<td>-0.999</td>
</tr>
<tr>
<td>4</td>
<td>0.122</td>
<td>0.282</td>
<td>-0.999</td>
</tr>
<tr>
<td>5</td>
<td>0.089</td>
<td>0.174</td>
<td>-0.993</td>
</tr>
<tr>
<td>6</td>
<td>0.061</td>
<td>1.137</td>
<td>-0.988</td>
</tr>
<tr>
<td>7</td>
<td>0.152</td>
<td>0.331</td>
<td>-0.994</td>
</tr>
<tr>
<td>8</td>
<td>0.204</td>
<td>0.048</td>
<td>-0.990</td>
</tr>
<tr>
<td>9</td>
<td>0.163</td>
<td>0.063</td>
<td>-0.990</td>
</tr>
<tr>
<td>10</td>
<td>0.159</td>
<td>4.758</td>
<td>-0.988</td>
</tr>
<tr>
<td>11</td>
<td>0.114</td>
<td>6.809</td>
<td>-0.982</td>
</tr>
<tr>
<td>12</td>
<td>0.176</td>
<td>0.552</td>
<td>-0.988</td>
</tr>
<tr>
<td>13</td>
<td>0.209</td>
<td>0.043</td>
<td>-0.984</td>
</tr>
<tr>
<td>14</td>
<td>0.234</td>
<td>0.086</td>
<td>-0.988</td>
</tr>
<tr>
<td>15</td>
<td>0.232</td>
<td>0.023</td>
<td>-0.988</td>
</tr>
<tr>
<td>16</td>
<td>0.295</td>
<td>0.585</td>
<td>-0.982</td>
</tr>
<tr>
<td>17</td>
<td>0.394</td>
<td>0.045</td>
<td>-0.988</td>
</tr>
<tr>
<td>18</td>
<td>0.585</td>
<td>0.033</td>
<td>-0.984</td>
</tr>
<tr>
<td>19</td>
<td>0.354</td>
<td>0.516</td>
<td>-0.989</td>
</tr>
<tr>
<td>20</td>
<td>0.522</td>
<td>10.816</td>
<td>-0.977</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fold</th>
<th>Accuracy (CH index)</th>
<th>Accuracy (DB index)</th>
<th>Accuracy (SH index)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.412</td>
<td>0.294</td>
<td>0.647</td>
</tr>
<tr>
<td>2</td>
<td>0.529</td>
<td>0.353</td>
<td>0.588</td>
</tr>
<tr>
<td>3</td>
<td>0.647</td>
<td>0.353</td>
<td>0.706</td>
</tr>
<tr>
<td>4</td>
<td>0.82</td>
<td>0.529</td>
<td>0.765</td>
</tr>
<tr>
<td>5</td>
<td>0.563</td>
<td>0.688</td>
<td>0.625</td>
</tr>
<tr>
<td>6</td>
<td>0.688</td>
<td>0.625</td>
<td>0.625</td>
</tr>
<tr>
<td>7</td>
<td>0.563</td>
<td>0.438</td>
<td>0.563</td>
</tr>
<tr>
<td>8</td>
<td>0.563</td>
<td>0.313</td>
<td>0.688</td>
</tr>
<tr>
<td>9</td>
<td>0.438</td>
<td>0.438</td>
<td>0.813</td>
</tr>
<tr>
<td>10</td>
<td>0.313</td>
<td>0.188</td>
<td>0.75</td>
</tr>
<tr>
<td>Overall</td>
<td>0.555</td>
<td>0.421</td>
<td>0.677</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Appliance</th>
<th>Accuracy (CH index)</th>
<th>Accuracy (DB index)</th>
<th>Accuracy (SH index)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile charger</td>
<td>0.45</td>
<td>0.15</td>
<td>0.65</td>
</tr>
<tr>
<td>Notebook</td>
<td>0.55</td>
<td>0.3</td>
<td>0.65</td>
</tr>
<tr>
<td>Fan</td>
<td>0.65</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Iron brush</td>
<td>0.85</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>Electric stove</td>
<td>0.4</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>Hair dryer</td>
<td>0.5</td>
<td>0.5</td>
<td>0.458</td>
</tr>
<tr>
<td>Light bulb</td>
<td>0.55</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Vacuum cleaner</td>
<td>0.5</td>
<td>0.35</td>
<td>0.95</td>
</tr>
<tr>
<td>Overall</td>
<td>0.555</td>
<td>0.421</td>
<td>0.677</td>
</tr>
</tbody>
</table>

V. CONCLUSION

The proposed ASIS using K-means algorithm has accuracy of 67.7% using SH index for determination of optimal number of clusters. This provides evidence that same type of electric appliances have similar characteristic in the feature vector components, active power, reactive power, total harmonic distortion and maximum transient current (rms). The appliance signature can enhance the performance of demand response. Energy auditing allows end-users and utility analyze the total energy consumption especially the energy profile of individual
appliances. For end-users, they can propose their own strategies to reduce the use of electricity based on the detailed energy profile. For example, if they study the energy profiles of the appliances carefully, they can conclude that some appliances waste lots of energy and thus trying to replace with a more energy effective products (another brand or series). They only know the total energy consumption without the proposed sustainable system. For utility, a more effective energy management can be proposed and implemented based on additional information of energy consumption pattern of users. Appliance signature gives information to utility what appliances every end-users is using. Therefore, demand response can perform accurately. As a result, total energy consumption is reduced in entire world which protects the environment. In the future, ASIS is planned to modify so that it can improve the accuracy like increasing the dimension of feature vector. Also, other algorithm, like support vector machine, neural network, mean-shift clustering could be attempted to investigate whether the performance is better than the proposed K-means clustering.

ACKNOWLEDGMENT

Gratitude is expressed to Center for Smart Energy Conversion and Utilization, City University of Hong Kong and Citycom Technology Ltd. for providing the testing sites and testing facilities.

REFERENCES