Classification of Home Appliance by using Probabilistic KNN with Sensor Data

SeungJun Kang and Ji Won Yoon*
Korea University, Center for Information Security Technologies, Seoul Republic of Korea
E-mail: kangren@korea.ac.kr, jiwon_yoon@korea.ac.kr*

Abstract—To date, many researchers have been conducted studies to control an electrical power to construct a smart home system which automatically manipulates individuals. One of the recent topics is NILM(Non-intrusive Load Monitoring) system to infer the devices states. In NILM, the approaches have been focused on dealing only with the feature of the electrical power signals to identify the states of the running devices. However, it is hard to classify all of devices with such traditional approaches. To solve and increase the accuracy, we propose a new method to infer the device states by electrical power signal from the home appliances and also sensor data including temperature and humidity. In this paper, we compare the performance among PKNN(Probabilistic K-Nearest Neighbor) and other algorithms. We apply the three methods in PKNN and analyze the comparison through AUC(Area Under the ROC). Finally, we can find the optimized parameters for accurate classification in each method.

I. INTRODUCTION

Power saving is one of the biggest problems that our world faces. Too much consumption can lead to several problems such as lack of energy, overload of power plant to the maximum capacity, blackout, and so on. In 2003, there was the biggest blackout in U.S history [10]. According to the source, the power plants around the country are producing the same amount of power which millions of consumers are consuming. However, where the demand is not balanced, the power plant is pushed to produce the amount come up to the maximum capacity which can lead to the power down. These problem could also lead to the abundant waste of money. The most efficient way to economize the consuming habit is to reduce the usage of home appliance, especially contains electricity-guzzlers such as air-conditioner or heater. Surely, this minor action can reduce the power consumption up to 20%. If we cannot undertake the energy saving, the government needs to build more additional power plants to generate more electronic power. However, it cannot only produce more CO$_2$ in the atmosphere but also worsen the climate change. To achieve the energy saving and also maximize the effective power consuming, we need a system to monitor and control the power in a real-time based.

Appliance Load Monitoring (ALM) is a technique to measure the electricity consuming in several devices by sticking the sensor to the devices. However, it incurs extra hardware cost and extra time for the installation complexity [12]. NILM is a technique to infer the energy consumption from each home appliances device by analyzing the changes of consumed power from the combined total value of electrical power. It decomposes the total consumption power signals into separated power consumption signal for each device. However, this is a rather difficult process since many devices including small lights consumes relatively small power compared to other devices like heaters. Because of this, they can be ignored since some of weak signals are much weaker than even the noise of the strong signals. To overcome the problem, the usage of sub-metering can be one of the solutions [5]. Fortunately, along with the development of smart home, we can expect that the usage of sensors can be increased in the near future. Sensor data plays an important role in the smart home system. The amount of data produced by the sensor itself, such as temperature, humidity, etc., can be enourmously large. Thus, the classification of devices can be improved by using such additional sensor data. Eventually, this method can improve the performance of the previous NILM method where it only uses the power data. Furthermore We use PKNN [6] to infer the device’s status.

This paper consists of seven sections. Section one has been introduced. Section two shows the related works regarding NILM and PKNN. Section three explains the detail of PKNN. The feature selection for NILM classification is described in section four. Our proposed approach will be shown in section five, along with the experimental results in section six. Finally, this paper will be concluded in section seven.

II. RELATED WORK

NILM focused on two major features to analyze the changes in voltage and current on house appliances. The two primary features are steady state and transient state, respectively. The steady state uses the feature condition in all point in the plot based on the NILM principle. It uses a low-sampling without a preprocessing process (see transient state and event detection). However, when there is a small changing load, the system can recognize the change as a summation of other devices rather than the state of the original device. Thus, it is hard for achieving a complete disaggregation [1]. The transient state is a sudden change which disturbs the steady state. This state recognized as ON/OFF events belonging to same appliances which grouped together [2]. This method is highly efficient for identifying many devices with a small load simultaneously. However, this transient state needs a preprocessing process for a high sampling rate. Apart from those two major features, we need an efficient method to obtain a high accuracy classification of home appliances with small
power load. One of the proposed methods is event detection. Event detection is a process of looking the changes (event) in the signal to determine the device state. The event detection is also known as a transition state. The output data depends on the sensor data change even, such as the weather change. Many research regarding the disaggregation in NILM by using event detection has been conducted [1], [3], [5], [11].

Many classification algorithms by using two features (steady state and transient state) has been proposed. Hidden Markov Model (HMM) technique is one of the algorithm among them which has been utilized in NILM [7], [8]. However, it was difficult to use steady state in HMM. HMM requires an offline learning method with less scalability [12]. To solve the problem, some papers have utilized the K-nearest neighbor (KNN) method to increase the scalability. In this paper, we attempt to increase the accuracy performance of KNN. We are using Probabilistic K-nearest neighbor (PKNN) for classifying the device state utilized in a smart home system.

III. PROBABILISTIC KNN

In order to improve traditional KNN, there have mainly been two variations developed. One of the well-known variations is weighted KNN, which assigns different weights to neighbors. The other approach is PKNN [6], which is a different method with respect to the weighted KNN. The concept of weighted KNN is to give a high value to higher weight among of the nearest neighbors based on distance [4]. However, in PKNN, which calculates the probability of selected label, the ratio of the total label is also necessary. PKNN is a classification method which calculates the training data’s neighbor by its probability. Here, assume that there are \( k \) nearest neighbor labels among which each has its probability value. Among those labels, we calculate each probability and find the Maximum Likelihood (ML) label through training data. Thus, we assign the new label as ML for the new data. The equation of PKNN is shown in equation (1) below

\[
p(Y|X, \beta, k) = \prod_{i=1}^{N} \frac{\exp\{(\beta/k) \sum_{j=1}^{k} \delta_{y_i,j}\}}{\sum_{q=1}^{Q} \exp\{(\beta/k) \sum_{j=1}^{k} \delta_{y_i,j}\}}
\]

where \( Y \) is the label and \( Y_x \) is the label of data \( x \). \( X \) is the training data set, \( N \) is a total training data size and \( k \) is the threshold of the neighbor. \( \beta \) is the regression coefficient which denotes the weight value among the neighbors. \( \delta(a, b) \) is a dirac function which gives value 1 when \( a \) and \( b \) are equal and 0 when they are different. Meanwhile, \( j \) is the nearest neighbor consecutive order from the nearest \( j_1 \) to \( j_k \). \( Q \) is the variable of total label, also the combination of home appliances status. By examining the equation, the denominator denotes the probability of label from the neighbor data among of the total training set. Numerator denotes the probability of each label of neighbor data. Hence, the probability of the corresponding label in general is divided by the total probability of neighbor data and classify the label by observing the ML value of each possible neighbor labels.

Note that, in this paper, we propose a new feature selection technique for NILM rather than classification technique. After obtaining the features, we use probabilistic KNN (PKNN).

IV. FEATURE SELECTION FOR NILM CLASSIFICATION

A. Full combination based approach

One of the basic approaches is to maintain full combinations of each device’s states as a label. If we assume that there are \( D \) devices and only two state \( ON/OFF \) applied, then we need \( 2^D \) combinations of states. For instance, we will have \( 2^3 = 8 \) combinations for \( D = 3 \), i.e. \( (OFF,OFF,OFF), (OFF,OFF,ON), \ldots, (ON,ON,ON) \). This means that we can regard each combination as a particular label so there are \( 2^D \) labels for \( D \) devices. In this case, our problem in an analytical point of view becomes \( 2^D \) multiple lable classification for \( D \) devices. Therefore, it also needs the multiple training set which all devices depend on the combination. However, in this case, we need extremely large size of training dataset for figuring out the \( 2^D \) labels or classes.

B. Independently factorized approach

To reduce the size of training data set on above full combination based approach, we can use an independently factorized approach which forms the combination of devices by using only the feature of \( ON \) state for each devices. \( ON \) state of each devices follows the Gaussian’s central limit theorem of normal distribution when there are over 30 examples. Therefore, in this case, we need only \( D + 1 \) labels compared to \( 2^D \) for full combination based method. For three devices \( D = 3 \), we have only four labels or classes \( \mathcal{L}_1 = (ON,OFF,OFF) \) for the first device, \( \mathcal{L}_2 = (OFF,ON,OFF) \) for the second device, \( \mathcal{L}_3 = (OFF,OFF,ON) \) for the third device and \( \mathcal{L}_0 = (OFF,OFF,OFF) \) for initial state. With only four states, we can make other states by simple ‘and’ or ‘addition’ operation by

\[
(ON,OFF,ON) = \mathcal{L}_1 \cup \mathcal{L}_3 \cup \mathcal{L}_0
\]

Therefore, this can reduce training set size but might increase computing time for additional combination. In this approach, we estimate the feature values since the devices are assumed independent. Here, the electronic consumption separately and independently gathered in the training step while all devices are jointly considered in the full combination based approach.

C. Event Detection based Method

Event detection uses the feature obtained from alteration rate when device’s state has changed. For example, when the humidifier shows change from OFF to ON, if the power alteration is 38, then the value can be used a feature. That is, this approach only requires features when the device state are altered, thus this approach does not require to store dataset for all the time so the size of training set could be small. However, event detection approach has still serious problems. One of the problems is that the high power devices sometimes produce a higher noise which can be mistakenly recognized
as low power devices. Hence, it may be mistakenly inferred even though there was no alteration in the device state.

From this point of view, we are using not only feature and electronic power, but also sensor data which are obtained from temperature and humidity sensors. Event detection method produces more specific results using sensor data, since we can obtain additional information like alteration rate and time of temperature and humidity. To calculate the event detection, we need to compute the alternation rate of the power as below

\[ W_e = \int_t^{t+\Delta t} V(t) \times I(t) \, dt \]  \hspace{1cm} (2)

where \( V(t) \) and \( I(t) \) are voltage and current of electronic consumption power respectively. \( W_e \) is the power alteration rate in a time interval \( \Delta t \).

\[ \Delta t = L(t) \neq L(t+\Delta t) \]

V. PROPOSED APPROACH

In order to increase the accuracy rate and reduce the miscalculation, we propose our approach based on PKNN and sensory data. We use PKNN to classify the home appliance in the smart home system. However, the traditional PKNN calculates the probability from the total sample which requires a high computational time. In this paper, we use an approximation PKNN by reducing the sample; it only uses some part of the total sample in order to reduce the computational time. This approach uses three feature: steady state, transient state, and event detection. Thus, the total number of feature use in this approach are number of device + 3 feature data which shown as below.

\[ f = [\text{power}, \text{temperature}, \text{humidity}, S_1, S_2, ..., S_D] \]

where \([\text{power}]\) is handled by using the steady state, \([\text{temperature}, \text{humidity}]\) are handled by using the transient state, and \([S_1, S_2, ..., S_D]\) are the event detection data. The event detection data are extracted from each home appliances’ state \(S_i, i = 1, ..., D\). The event detection are using the power alteration after it altered by using the equation below.

\[ f_e = \min_i[W_e(i) - W_e(\text{new})] \]  \hspace{1cm} (3)

where \( W_e(i) \) is \( i \)-th device’s power alteration rate, \( W_e(\text{new}) \) is the new power alteration rate. \( f_e \) is the device state in event \( e \). The least difference between the trained event set and new event is calculated and chosen as the corresponding event. For example, the initial state of devices start from \((\text{OFF}, \text{OFF}, \text{OFF}, \text{OFF})\). When the device 3 event change into \(\text{ON}\), the device state feature is changed into \((\text{OFF}, \text{OFF}, \text{ON}, \text{OFF})\). Here, \(\text{ON}\) shows as 1 and \(\text{OFF}\) shows as 0. All of the feature \( f \) range are normalized by the range of \([0,1]\).

VI. EXPERIMENTAL RESULTS

Our experiment was conducted in windows 7 environment and using matlab R2014a. The result data were produced through smartplug and thermo-hygrometer with the sampling time based on 1 Hz. We use variable devices as follows: Air conditioner, light, humidifier, fan and computer. In all of the experiments, we used 10 fold cross-validation and given the PKNN nearest neighbor number \( k \) as 7. The experiment was conducted by calculating the average of five cases for each devices. The illustration of the system environment is shown in Fig. 1.

The experiment divided into four parts. First, we compared the accuracy of PKNN and other algorithms by only using power as its feature. We also obtained the result by additionally adopting sensory data. Second, we trained the sensor data \([\text{temperature}, \text{humidity}]\) with transient state and compared the accuracy of the sensor data only and the sensor data with transient state. Third, we analyzed the comparison between
traditional full combination and three approaches, such as full combination with sensor data, independently factorized and the proposed approach. Fourth, we searched for the optimal parameter $k$ which is essential for a proper and accurate classification.

The first part of experiment’s results is shown in Fig. 2. The $x$–axis shows the number of devices and $y$–axis shows the accuracy comparison between PKNN and other algorithms. The accuracy of most algorithms drops significantly as the number of devices increases. However, PKNN with both power data and sensory data can give higher accuracy (95%) for five devices than SVM, Naviebays and J48 [9]. The performance of PKNN increases significantly by using sensor data as feature, regardless the number of devices.

The second part of experiment’s results is shown in Fig. 3. This experiment gives the accuracy difference between sensor data (temperature and humidity) and sensor data with transient state. The $x$–axis shows the number of devices and $y$–axis shows the accuracy. The sensor data with transient state gives a better accuracy. During the experiment, the devices were turned ON/OFF at a fixed period of time. However, when there are four devices, each device is turned ON and OFF in a short time length, caused a frequent alteration. Thus, the accuracy can not significantly increased. However, in the real life environment, devices are not turned ON and OFF frequently. Yet, the result can be increased in the real life case. Fig. 3 also shows that the accuracy of sensor data with transient state is higher than without transient state when there is more than five devices. It can prevent the classification error and gives a better result.

In this experiment, the sensor data, temperature and humidity, is defined and utilized. However, the temperature and humidity are depend on the season and weather at the given time. For example, in summer, the temperature is higher and the humidity is lower compare to other season. This particular condition gives a different effect for the sensor data. Thus, the accuracy of devices can be influenced by the sensor data error. So, we need to use the alteration of the sensor data.

The third part of experiment’s results is shown in Fig. 4. It shows the accuracy for each five cases applied to each of total five devices. Full combination approach shows the accuracy approximately 80%. Independently factorized approach gives nearly 40% accuracy. Since it is unstable, it can not use a classification method. Here, it shows that it is not the sum value of total power from devices. In the approach using sensor feature, we obtain accuracy by more than 95%. Moreover, the proposed method provides nearly 100% accuracy during which almost no misclassification occurred.

The fourth part of the experiment’s result is shown in Table I. The table shows the performance of algorithms with varying $k$ for 1 to 15. The accuracy is the average of five cases applied to each device. All of the three methodologies shows high accuracy. Event detection has the optimization value 97.94 at $K = 7$. PKNN with sensor and the proposed method have respective accuracies 98.26 and 99.76 when $K = 1$.

In this paper, we conducted the classification of home appliance experiments by using PKNN with sensor data. The experiments were utilizing the NILM, PKNN and three different feature selection approaches to sensor data. PKNN with sensor data showed a higher accuracy when the number of devices exceeds to more than four devices. Meanwhile, other

<table>
<thead>
<tr>
<th>method</th>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>97.36</td>
<td>97.51</td>
<td>97.89</td>
<td>97.85</td>
<td>97.91</td>
<td></td>
</tr>
<tr>
<td>Sensor</td>
<td>98.26</td>
<td>98.22</td>
<td>97.97</td>
<td>97.77</td>
<td>97.62</td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>99.76</td>
<td>99.73</td>
<td>84.50</td>
<td>99.73</td>
<td>99.72</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>method</th>
<th>$k$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>97.87</td>
<td>97.94</td>
<td>97.89</td>
<td>97.92</td>
<td>97.90</td>
<td></td>
</tr>
<tr>
<td>Sensor</td>
<td>97.40</td>
<td>97.31</td>
<td>97.13</td>
<td>97.00</td>
<td>96.81</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>method</th>
<th>$k$</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>97.88</td>
<td>97.91</td>
<td>97.85</td>
<td>97.84</td>
<td>97.86</td>
<td></td>
</tr>
<tr>
<td>Sensor</td>
<td>96.76</td>
<td>96.59</td>
<td>96.47</td>
<td>96.30</td>
<td>96.24</td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>99.36</td>
<td>99.29</td>
<td>99.24</td>
<td>99.23</td>
<td>99.18</td>
<td></td>
</tr>
</tbody>
</table>

TABLE I

OPTIMAL PARAMETER $k$

VII. CONCLUSION

In this paper, we conducted the classification of home appliance experiments by using PKNN with sensor data. The experiments were utilizing the NILM, PKNN and three different feature selection approaches to sensor data. PKNN with sensor data showed a higher accuracy when the number of devices exceeds to more than four devices. Meanwhile, other
algorithm decreased significantly at the same time. Moreover, the proposed approach produced a higher classification rate accuracy average of 99.76%, compare to other approaches. The accuracy remains stable above 99% while optimizing the parameter k of PKNN in the proposed approach. By using PKNN and sensor data along with a higher classification accuracy, the pattern of home appliances can be detected easily.

ACKNOWLEDGMENT

This research has been supported by the R&D grant from Korea Institute of Energy Technology Evaluation and Planning.

REFERENCES