Informative sensor selection on clustered sensors

Saed Sa’deh Juboor, Sook-Ling Chua and Lee Kien Foo
Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia
E-mail: slchua@mmu.edu.my

Abstract. Many researchers have focused their work on recognising activities in smart homes, with the aim to support the occupants by monitoring their daily activities and identify any abnormalities. In order to recognise human daily activities, sensors are installed in the home to collect information about the occupant. Activities can then be inferred from a sequence of sensor observations output from the house. However, one challenge still remains: which sensors are useful to effectively recognise the occupant’s activities. Many traditional filter-based methods have been proposed in the literature for sensor selection but these methods may not consider sensor inter-redundancy. Motivated by this, this paper addresses the sensor selection problem using clustering and then apply the filter-based method on the clustered sensors. The effectiveness of the method is evaluated using two well-known public smart home datasets. The results showed that our proposed method not only able to reduce the number of sensors needed but also able identify the set of sensors that are useful for activity recognition.

1. Introduction

The smart home is an automated home, which uses a variety of sensors around the house to monitor the human daily activities, particularly the older adults who are living alone in the home. The aim of the smart home is to improve the overall quality of life by allowing greater independence and detecting potential anomalous activities. Many types of sensors have been proposed in the literature to recognise human daily activities. Unobtrusive state-change sensors, RFID, pressure mat and motion sensors are commonly used to collect information about the occupant. These sensors are attached to household objects such as doors, windows, cabinets, etc., and are activated when the occupant does his/her activities like opening cupboards, toileting, closing the refrigerator door, etc. Activities can be inferred from a sequence of sensor observations output from the house.

One of the challenges in activity recognition is how to select sensors that are more relevant for recognising the activities of the occupant. Many research attempt to learn from a large number of sensors with the aim that the classifier obtains a good representation of the occupant’s activities. However, studies have shown that this is not always the case since training a classifier on irrelevant and/or redundant sensors affects the overall recognition and computational performance. Furthermore, it requires a larger training data to accurately learn the occupant’s activities [1].

An irrelevant sensor is a sensor that does not help to discriminate the different classes of activities. Even if a sensor is relevant to an activity i.e., with high correlation with the activity class, the sensor could be redundant. In this paper, we view redundancy as of two types: intra-redundancy and inter-redundancy. Intra-redundancy refers to sensors, which are highly
correlated with other sensors that are relevant to the same activity but the removal of one of the sensors will not affect the recognition accuracy. For example, a sensor attached to the microwave oven and a sensor attached to the kitchen stove. Removing either one of the sensor will not affect the classification of ‘preparing meal’ activity.

Inter-redundancy, on the other hand, refers to sensors which are highly relevant to different classes of activities. If other sensors are able to differentiate among these activities then the sensor can be removed. For example, sensor on the bathroom light may be included in multiple activities such as somebody could be in the bathroom showering, grooming, or toileting. If other sensors can help to distinguish among these activities, then the sensor attached to the bathroom light can be removed without affecting the recognition performance.

This paper aims to address the inter-redundancy problem by first cluster the sensors and then apply filter-based methods on the clusters for sensor selection. The proposed method is validated on two distinct smart home datasets and compared with baseline methods.

2. Related work
This section reviews the related methods that address the sensor selection problem for activity recognition. There are two general approaches to sensor selection. The filter-based approach evaluates the sensors based on some heuristics and often used as a pre-processor to determine the set of informative sensors for a learning algorithm. Sensors are selected based on their relevance and discriminating power with regard to the activities. The wrapper-based approach, on the other hand, uses an induction algorithm to estimate the sensor subsets [2] [3]. Such an approach requires extensive computation since it relies on a learning algorithm to evaluate each and every sensor subsets.

Dobrucali and Barshan [4] used mutual information criterion to select informative wearable sensors based on sensor types, measurement axes and sensor locations for recognizing daily sport activities. Alzubaidi et.al. [5] combined a genetic algorithm with mutual information to select the informative features for breast cancer diagnosis. In their work, genetic algorithm is applied on the features to select important feature subsets and then applied mutual information on the selected feature subsets to select the most relevant features for diagnosis.

Chua and Foo [6] applied information gain for sensor selection, where sensors with non-zero information gain were selected as informative. Chahuara et al. [7] also used information gain to select important features from audio and home automation sensors for activity recognition. Lee and Cho [8] applied information gain to identify informative mobile sensors to recognize both short-term and long-term activities. In another work done by Chetty et al. [9], they used information gain to select important features from smart phone data, where features with high information gain values are selected for activity recognition. However, these works do not consider the inter-redundancy among the sensors.

Among the works that are closely related to our work are Mishra and Sahu [10], and Taylor et al. [11]. Mishra and Sahu [10] first used k-means to cluster the features (genes) of microarray data. Signal-to-noise ratio is applied to rank the features for each cluster. Their work aims to get a subset of top ranked informative genes features that have less intra-redundant features. Taylor et al. [11] proposed a two-stage feature selection to remove both intra-and inter-redundancies features on telemetry signal data obtained from vehicle. In their work, they used maximum redundancy feature selection method, where the first stage was applied on features from individual signal and the second stage was applied on the combined selected features results from the first stage.

3. Proposed method
We approach the inter-redundancy problem by applying clustering on the sensor observations. With clustering, sensors that share similar characteristics will be grouped together. We reason
that an activity consists of a number of actions. For example, ‘making tea’ activity may involve actions such as ‘taking milk from the fridge’, ‘turning on the kettle’ and ‘taking cup from the cupboard’. Based on this reasoning, our method applies the hierarchical agglomerative clustering to cluster the sensors into cluster hierarchy. The resulting set of clusters represents the occupant’s activities. We can then applied the filter-based methods on each cluster to evaluate the importance of the sensors for recognising that particular activity.

In hierarchial agglomerative clustering, the algorithm starts by treating each sensor observation as a cluster. Then, it merges the most similar observations into a new cluster and repeat the process until all the observations are merged into one cluster. The similarity between clusters are calculated using average-linkage method, where two clusters with the smallest average linkage distance are combined. The average-linkage distance between any two sensors $S_1$ and $S_2$ from two clusters $C_1$ and $C_2$ is defined as follows:

$$ d(S_1, S_2) = \frac{1}{|C_1|.|C_2|} \sum_{S_1 \in C_1} \sum_{S_2 \in C_2} d(S_1, S_2) $$ (1)

Since the sensor observations output from the house are in binary form, we used Jaccard distance as the distance measure, where $d(S_1, S_2)$ is:

$$ d(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} $$ (2)

$|S_1 \cap S_2|$ is the number of shared members between both sets $S_1$ and $S_2$, and $|S_1 \cup S_2|$ is the total number of members in both sets $S_1$ and $S_2$.

We then applied the filter-based method on each cluster to find the set of informative sensors. The sensor with the highest information-theoretic value is selected from each cluster. In our work, we applied two information-theoretic measures: (1) information gain and (2) mutual information.

Information gain calculates the expected information for each sensor. Therefore, sensors that have high relevance for activity class will be selected. $Gain(R, S)$ of sensor $S$ relative to a collection of examples $R$ is:

$$ Gain(R, S) = Entropy(R) - \sum_{s \in Values(S)} \frac{|R_s|}{|R|} Entropy(R_s) $$ (3)

where $Values(S)$ are the possible values of sensor $S$ and $R_s$ is the subset of $R$ for which sensor $S$ has value $s$, i.e., $R_s = \{ r \in R | S(r) = s \}$.

Mutual information measures the dependency between sensor $S$ and activity $R$, which is:

$$ MutualInfo(S, R) = \sum_{s \in Values(S)} \sum_{r \in (R)} \log \left( \frac{P(s, r)}{P(s)P(r)} \right) $$ (4)

where $P(s, r)$ is the joint probability distribution of $s$ and $r$. $P(s)$ and $P(r)$ are the marginal distributions.

Algorithm 1 shows the algorithm of our method for sensor selection using filter-based method on clustered sensors.
Algorithm 1 Sensor selection using filter-based method on clustered sensors.

Require: sensor observations \( S \)

Ensure: \( k \) = number of clusters

\( \text{Cluster}_k \leftarrow \text{Apply hierarchical clustering on } S \)

for \( i = 1 \) to \( k \) do

\( \text{Score}_i \leftarrow \text{Apply filter-based method on } \text{Cluster}_i \)

\( \text{Selected}_i \leftarrow \text{Select sensor with the highest information-theoretic value from } \text{Score}_i \)

end for

4. Data source and evaluation methods

To test the effectiveness of our method, we used two public smart home datasets: (1) MIT PlaceLab and (2) van Kasteren.

The MIT PlaceLab [12] designed a smart home with 77 state-change sensors installed in the home. A subject lived in the home for 16 days and data were annotated with activities by the subject himself, which provides us the ground-truth information about the occupant’s activities. There are 5 activities of daily living observed in this dataset. These activities are toileting/showering, doing laundry, dressing/grooming, washing/putting away dishes and preparing meal. Of the total sensors installed in the home, 49 state-change sensors were used to collect information on these activities. For this dataset, we used leave-2 days-out cross validation method i.e., 14 days were used for training and 2 days for testing, which result in 8 training-test splits.

For van Kasteren’s dataset [13], a total of 14 state-change sensors was installed in a three-room apartment and data was collected over a period of 24 days. There were four activities observed in this dataset, which are leaving house, toileting/showering, go to bed and preparing meal. Since the number of activities and sensors used in this dataset are smaller compared to MIT PlaceLab, we used leave-4 days-out cross validation method i.e., 20 days for training and 4 days for testing. This result in 6 training-test splits.

For each training set, we trained 2 classifiers – hidden Markov models (HMM) and naïve Bayes classifier (NB). The main purpose of training two different classifiers is to validate the results of sensor selection and not to compare between classifiers. We used 4 metrics to measure the performance of the classifiers: accuracy, precision, recall and \( F \)-measure. Accuracy is the ratio of the total number of activities correctly recognised over the total number of activities examined. Precision is the percentage of inferred activities correctly recognised, while recall is the percentage of correctly recognised ground truth activities. \( F \)-measure is the average of precision and recall.

5. Experiments and results

We have conducted two experiments on both smart home datasets discussed in Section 4. The first experiment is to compare the performance of classifiers trained on sensors selected using our method with classifiers trained on the full set of sensors. The second experiment is to compare our method with classifiers trained on sensors selected by running information gain or mutual information directly on the whole dataset without clustering (baseline methods).

5.1. Comparison between proposed method and full set sensors

In this experiment, we first perform hierarchical clustering and then find the informative sensors by running (i) information gain (IG) and (ii) mutual information (MI) on each cluster. In order to validate our method, we ran different number of clusters for each dataset. For MIT PlaceLab dataset, we clustered the sensors into 5, 13 and 21 clusters. Since van Kasteren dataset has fewer number of sensors, we have clustered the data into 7, 9 and 11 clusters.
We ran IG and MI on each of the resulting cluster separately and select the sensor with highest IG or MI value from each cluster. For all the clusters, we noticed that the same sensor has the highest IG and MI value. Thus, IG and MI selected the same set of top sensor. The summary of the resulted sensor set are shown in table 1. HMM and NB are trained on each set of sensors. We compared the performance of these sensors with the same classifiers trained on the full set of sensors (49 sensors for MIT PlaceLab and 14 sensors for van Kasteran) for activity recognition. The results are shown in figure 1 for MIT PlaceLab and figure 2 for van Kasteran.

Table 1. Sensors selected with proposed method on (a) MIT PlaceLab and (b) van Kasteren datasets.

| (a) MIT PlaceLab Dataset | (b) van Kasteren Dataset |
|--------------------------|--------------------------|
| 5 Clusters | 4 Clusters | 4 Clusters |
| 13 Clusters | 7 Clusters | 7 Clusters |
| 21 Clusters | 11 Clusters | 11 Clusters |
| Kitchen containers | Front door | Microwave |
| Dishwasher | Freezer | Washing machine |
| Freezer | Toilet flush | Front door |
| Toilet flush | Kitchen cabinet 2 | Washing machine |
| Kitchen door | Freezer | Fridge |
| Bedroom drawer 1 | Kitchen cabinet 2 | Bedroom door |
| Foyer closet | Kitchen cabinet 3 | Bedroom door 1 |
| Bedroom drawer 2 | Kitchen cabinet 1 | Pans cupboard |
| Kitchen light | Toilet flush | Pans cupboard |
| Kitchen containers | Kitchen cabinet 4 | Bathroom door |
| Dishwasher | Bedroom drawer 1 | Fridge |
| Kitchen door 2 | Foyer closet | Toilet door |
| Office light switch | Bedroom drawer 2 | Cups cupboard |
| Shower faucet | Kitchen light | Freezer |
| Bathroom exhaust fan | Kitchen containers | Bathroom door |

Figure 1 shows that the recognition performance of both HMM and NB trained on top sensor selected from 13 or 21 clusters outperform HMM and NB trained on the full set of 49 sensors for MIT PlaceLab. When trained on only 5 top sensor selected from 5 clusters, performance of HMM and NB are compatible to the full set, with slightly lower precision and accuracy for NB. As for van Kasteran, HMM trained on top sensor selected from 7 or 11 clusters is compatible to the full set of 14 sensors. The performance of HMM trained on top sensor from only 4 clusters is not as good as the full set. Performance of NB trained on top sensor selected from 4, 7 or 11 clusters are compatible to the full set, except lower precision and F-measure for top sensor from 4 clusters.

5.2. Comparison between proposed method and baseline methods
In this experiment, we applied IG and MI directly on the full set of sensors in both datasets. The sensors are then ordered in descending order according to their IG or MI values. Table 2 shows the list of the top 21 sensors (out of 49) for MIT PlaceLab dataset and top 11 sensors (out of 14) for van Kasteren dataset.

For a fair comparison, the classifiers’ performance is compared in the following way: classifiers trained on the top 5 sensors (IG or MI value) is compared to classifiers trained on top sensor with highest IG or MI value from each of the 5 clusters. The same comparison is carried out for the other number of clusters.

Figures 3 and 4 show the recognition performance of HMM and NB trained on 5, 13 and 21 sensors selected based on IG or MI values, with and without clustering on MIT PlaceLab.
dataset. All the boxplots clearly show that the classifiers perform better when trained using our method i.e., top sensor selected from clusters, as compared to sensors selected when IG and MI were applied directly on the full set of data without clustering.

Figures 5 and 6 show the recognition performance of HMM and NB trained on 4, 7 and 11 sensors selected based of IG or MI values, with and without clustering on van Kasteren dataset. The classifiers perform equally well for 4 and 11 sensors. HMM trained on 7 sensors with highest
Figure 2. Boxplot showing recognition performance of (a) hidden Markov model and (b) naïve Bayes classifier on sensor set selected using information gain or mutual information on clustered sensors, compared to the full set of 14 sensors on van Kasteren dataset.

IG values from the full set performed slightly better than HMM trained on the set of top sensor selected from 7 clusters (figure 5). However, the opposite result is shown in figure 6 when the sensors are selected using MI values, where HMM trained on the top sensor selected from 7 clusters outperform HMM trained on 7 sensors with the highest MI values from the full set. NB classifier also achieved a higher precision and $F$-measure when trained on top sensor selected from 7 clusters as compared to baseline method.
Table 2. Sensors selected by running IG or MI directly on the full set of sensors for (a) MIT PlaceLab dataset and (b) van Kasteren dataset.

(a) MIT PlaceLab Dataset

| IG          | MI          |
|-------------|-------------|
| Dishwasher  | Dishwasher  |
| Toilet flush| Kitchen door|
| Kitchen door| Toilet flush|
| Bathroom sink faucet - hot | Laundry dryer |
| Laundry dryer | Washing machine |
| Freezer     | Bathroom sink faucet - hot |
| Refrigerator| Refrigerator |
| Washing machine | Kitchen drawer |
| Kitchen drawer | Bathroom sink faucet - cold |
| Bathroom medicine cabinet 1 | Bathroom medicine cabinet 1 |
| Bathroom medicine cabinet 2 | Bathroom medicine cabinet 2 |
| Shower faucet | Shower faucet |
| Bathroom cabinet | Bathroom cabinet |
| Jewellery box | Kitchen cabinet 1 |
| Kitchen cabinet 1 | Kitchen cabinet 2 |
| Bedroom drawer | Toaster |
| Kitchen cabinet 2 | Jewellery box |
| Toaster      | Bedroom drawer |
| Bathroom door | Kitchen cabinet 3 |
| Kitchen cabinet 3 | Bathroom Door |

(b) van Kasteren Dataset

| IG          | MI          |
|-------------|-------------|
| Front door  | Fridge      |
| Bathroom door | Front door   |
| Toilet flush | Bathroom door |
| Fridge      | Toilet flush |
| Toilet door | Plates cupboard |
| Bedroom door | Toilet door |
| Plates cupboard | Groceries cupboard |
| Groceries cupboard | Cups cupboard |
| Cups cupboard | Bedroom door |
| Freezer     | Freezer     |
| Pans cupboard | Pans cupboard |

6. Discussion
For the MIT PlaceLab dataset, 13 or 21 sensors selected using our method performed better than the full set of 49 sensors and performed equally well even with 5 sensors. For van Kasteren dataset, our method is compatible to the full set of 14 sensors when we used 7 or 11 clusters. However, the performance is not as good when we used 4 clusters. The main reason was that in the van Kasteren dataset, there are only 4 activities with 14 sensors where each activity is represented by a distinct set of sensors. Furthermore, there are less activity variations observed in this dataset. This is not the case in MIT PlaceLab where 49 sensors were used to capture 5 activities, which resulted in heavy intra- and inter-redundancies among sensors. The good performance obtained on MIT PlaceLab dataset showed that our method has an advantage in selecting sensor set that is less redundant.

When comparing the performance of classifiers trained on sensor set selected using our method with sensor set selected by applying the filter-based method directly on the full set of data, our method again performed better on MIT PlaceLab, which suggest that our method is able to select sensor set that is relevant to the activities in smart home, but with less redundancies.

In all the experiments, we have chosen 5, 13 and 21 clusters for MIT PlaceLab and 4, 7 and 11 clusters for van Kasteran. The minimum number of clusters is chosen to match the number of activities observed in each of the smart home dataset and the rest of cluster numbers are chosen arbitrary.

7. Conclusion
In this paper, we addressed the sensor selection problem in smart home by combining the strengths of clustering algorithm with filter-based feature selection method. Rather than applying filter-based methods directly on the sensors, we first applied clustering to minimise the inter-class similarity among the sensors. Filter-based methods are then applied on each cluster to evaluate the importance of the sensors for activity recognition. The proposed method is
applied on two distinct smart home datasets. Two classifiers were trained on the selected sensor set. Results showed that sensor set selected with our method has a better performance when comparing to the full set sensors used for activity recognition. Our method also outperformed the baseline method when applying filter-based method without clustering. Determining the set of informative sensors not only reduce the cost of sensors deployment in a smart home for activity
Figure 4. Boxplot showing recognition performance of (a) hidden Markov model and (b) naïve Bayes classifier on sensor set selected using mutual information with clustering (W) and without clustering (O) on MIT PlaceLab dataset.

recognition but also improve the recognition performance of the learning algorithms. With the encouraging results, our plan is to further research on determining the optimum number of clusters.
Figure 5. Boxplot showing recognition performance of (a) hidden markov models and (b) naïve Bayes classifier on sensor set selected using information gain with clustering (W) and without clustering (O) on van Kasteren dataset.
Figure 6. Boxplot showing recognition performance of (a) hidden Markov model and (b) naïve Bayes classifier on sensor set selected using mutual information with clustering (W) and without clustering (O) on van Kasteren dataset.
Acknowledgments
This research is supported by the Ministry of Higher Education, Malaysia under the Fundamental Research Grant Scheme (No: FRGS/1/2016/ICT02/MMU/02/3).

References
[1] Koller D and Sahami M 1996 Toward optimal feature selection Proc. of the 13th Int. Conf. on Mach. L. pp 284–292
[2] Chua S L and Foo L K 2017 Sensors 17 1902
[3] Attal F, Mohammed S, Dedabishvili M, Chamroukhi F, Ouahellou L and Amirat Y 2015 Sensors 15 31314–31338
[4] Dobrucal O and Barshan B 2013 Sensor-activity relevance in human activity recognition with wearable motion sensors and mutual information criterion Proc. of the 28th Int. Symp. on Comp. and Inf. Sci. (Springer, Cham) pp 285–294
[5] Alzubaidi A, Cosma G, Brown D and Pockley A G 2016 Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information Int. Conf. on Inter. Tech. and Gam. (IEEE) pp 70–76
[6] Chua S L and Foo L K 2015 Procedia Comput. Sci. 69 116–124
[7] Chahuara P, Fleury A, Portet F and Vacher M 2016 J. of Ambient Intell. Smart Environ. 8 399–422
[8] Lee Y S and Cho S B 2016 Pattern Anal. Appl. 19 1181–1193
[9] Chetty G, White M and Akther F 2015 Procedia Comput. Sci. 46 1181–1187
[10] Mishra D and Sahu B 2011 Int. Jour. of Sci. & Eng. Res. 2 1–7
[11] Taylor P, Griffths N, Bhalerao A, Popham T, Zhou X and Dunoyer A 2013 Redundant feature selection for telemetry data 9th Int. Wosh. on Agt and Data Min. Interact. - Vol. 8316 (Springer-Verlag) pp 53–65
[12] Tapia E M, Intille S S and Larson K 2004 Activity recognition in the home using simple and ubiquitous sensors Pervasive pp 158–175
[13] van Kasteren T, Noulas A, Englebienne G and Kröse B 2008 Accurate activity recognition in a home setting Proc. of the 10th Int. Conf. on Ubi. Comp. (ACM) pp 1–9