Effects of Smart Home Dataset Characteristics on Classifiers Performance for Human Activity Recognition

Iram Fatima, Muhammad Fahim, Young-Koo Lee, Sungyoung Lee
Department of Computer Engineering, Kyung Hee University, Korea
{iram.fatima, fahim, yklee,sylee}@oslab.khu.ac.kr

Abstract. Over the last few years, activity recognition in the smart home has become an active research area due to the wide range of human centric-applications. A list of machine learning algorithms is available for activity classification. Datasets collected in smart homes poses unique challenges to these methods for classification because of their high dimensionality, multi-class activities and various deployed sensors. In fact the nature of dataset plays considerable role in recognizing the activities accurately for a particular classifier. In this paper, we evaluated the effects of smart home datasets characteristics on state-of-the-art activity recognition techniques. We applied probabilistic and statistical methods such as the Artificial Neural network (ANN), Hidden Markov Model (HMM), Conditional Random Field (CRF), and Support Vector Machines (SVM). The four real world datasets are selected from three most recent and publically available smart home projects. Our experimental results show that how the performance of activity classifiers are influenced by the dataset characteristics. The outcome of our study will be helpful for upcoming researchers to develop a better understanding about the smart home datasets characteristics in combination with classifier's performance.

1 Introduction

A smart home is an intelligent agent that perceives state of resident and the physical environments using sensors. Recent advancements in the field of machine learning and data mining have enabled activity recognition research using smart homes sensing data to play a direct role in improving the general quality of health care. It is one of the best solutions to provide a level of independence and comfort in the homes of elderly people rather than requiring them to reside at health care centers [1]. The advancement of sensor technology has proven itself to be robust, cost-effective, easy to install and less intrusive for inhabitants. This fact is supported by a large number of applications developed using activity recognition to provide solutions to a number of real-world problems such as remote health monitoring, life style analysis, interaction monitoring, and behavior mining [2] [3].

A diverse set of machine learning and data mining algorithms have been previously used to identify the performed activities from the smart home datasets. The quest to optimize the performance of classifiers has a long and varied history. The diverse characterized data of smart homes require intelligent machine learning and data mining algorithms for automated analysis in order to make logical inferences from the stored raw data that may results in activity classification [4]. With the passage of time researchers found refinements that result in more accurate classification on comparable datasets. We study the relationship between the distribution of data, on the one hand, and classifier performance, on other. It is shown that predictable factors such as the available amount of training data, the spatial variability of data samples, deployed sensors in smart homes and the total activity occurrences in the dataset influence the performance of classifiers to a significant degree.

To select an appropriate classifier for certain type of data, there is a need to understand the behavior of classifiers on different data characteristics. Despite the great work and diversity in the existing classification methods, no significant work is done so far to assist a researcher in selecting a suitable classification technique for a particular nature of smart home dataset. The process of selecting an appropriate classifier is still a trial and error process that clearly depends on the relationship between the classifier and the data. The focus of this study is to facilitate the researchers in order to understand the effects of dataset characteristics on different classifiers for activity recognition. A particular dataset cannot be classified with same accuracy from all classifiers. Some vital dataset characteristics are dataset duration, performed activities, deployed sensors, activated sensors for a particular activity, total occurrences of single activity, and closely correlated activities. We compared state-of-the-art classifiers such as Artificial Neural network (ANN) [5], Hidden Markov Model (HMM) [2], Conditional Random Field (CRF) [3], and Support Vector Machines (SVM) [6]. These four selected schemes are applied on four datasets selected from three most significant smart home projects such as CASAS [7], ISL [8] and House_n[9] smart homes. The main subject of this paper is to provide a systematic and unbiased evaluation of the existing activity classification schemes to resolve the uncertainties associated with the choice of classifier and the nature of smart home dataset. The results show that neither of the classifier is best for all datasets, the classification accuracy of each classifier depends on the underline data characteristics. We also illustrate that dataset characteristics highly affect the classifiers' individual class level assignments along with their overall performances.

The rest of the paper is organized as follows. We describe related work in Section 2. In Section 3, we discussed the smart home datasets with their important characteristics for activity recognition. In Section 4, we introduced four classifiers with their preferred settings for our experiments. The analyzed results of the CASAS, ISL, and House_n smart home datasets are presented in Section 5. Finally, conclusion and future works are given in Section 6.

2. Related Work

Several studies have been conducted to determine effective and accurate activity classification methods for smart home datasets. In [10] authors study the impact of semi-Markov models classification accuracy using datasets from ISL smart homes. They consider availability of labeled data, the importance of training time and speedy inference for experimental purpose. In their analysis they showed that CRF outperforms other semi-Markov models. The authors in [6] apply SVM to identify daily living activities on their health smart home dataset. They selected a set of features from dataset according to their domain of interest before using multi-SVM for effective activity classification as compare to other classifiers. The work in [11] applied the ANN for to cluster analysis of human activities of daily living within their own developed smart home environment. Specifically their approach is GSOM-based data mining to cluster analysis of human activities effectively.

The authors in [12] proposed a data mining framework to recognize activities based on raw data collected from CASAS smart home. The framework synthesizes the sensor information and extracts the useful features as many as possible. They compared several machine learning algorithms on the selected features to compare the performance of activity recognition. They discussed the performance of the machine learning algorithm on the basis of their selected feature based on information gain and mRMR. The authors in [13] employ data mining techniques to look at the problem of sensor selection for activity recognition in smart homes along with classifier selection. They examine the issue of selecting and placing sensors in a CASAS smart home in order to maximize activity recognition accuracy. In [2] authors used ISL smart home dataset to show the potential of generative and discriminative models for recognizing activities. They presented that CRFs are more sensitive to overfitting on a dominant class than HMM.

The commonly observed methodologies in literature for smart home datasets are with only limited number of algorithms from the machine learning repository and select the one which gives relatively better results for their particular domain. No existing work has intensions to analyze the classifiers to show the effects of data characteristics. Our study will help the researchers in choosing an appropriate classifier based on a particular type of dataset.

3. Smart Home Datasets

Smart home datasets are generally associated with high-dimensional features and multiple classes. To comprehensively evaluate the performance of various classification schemes on smart home datasets, we analyzed four datasets from three smart home projects. We selected *Tulum2009* and *Twor2009* from CASAS smart home project. The dataset duration is 83 and 46 days respectively and deployed sensors are 20 and 71 respectively. From ISL and House_n smart homes, *House A* and *Subject1* datasets are evaluated. The duration of these datasets is 24 and 16 days respectively and deployed sensors are 14 and 28 respectively. The detail analysis and our calculated data dimensions of datasets are shown in Table 1 and 2.

Table 1 The CASAS datasets Twor2009, and Tulum2009. The 'Num' column shows activity count, the 'Time' column shows activity time in minutes, and the 'Sensor' column shows activity sensor events

Twor2009				Tulum2009			
Activities	**Num**	**Time**	**Sensor**	**Activities**	**Num**	**Time**	**Sensor**
Idle	-	8240.93	73043	Idle	-	102986.4	203408
Bed Toilet Transition	39	94.55	2241	Wash Dishes	71	1204.84	24869
Meal Preparation	118	6207.32	41730	Watch TV	528	4955.43	52222
R1 Sleeping in Bed	35	18428.36	29503	Enter Home	73	119.42	604
R2 Sleeping in Bed	35	18572.11	29604	Leave Home	75	101.58	1854
Cleaning	2	49.75	1540	Cook Breakfast	80	1440.31	33435
R1 Work	59	5902.51	45675	Cook Lunch	71	972.74	24527
R2 Work	44	2530.03	17955	Group Meeting	11	1847.04	31084
R1 Bed to Toilet	34	337.06	2298	R1 Eat Breakfast	66	932.87	20077
R1 Personal Hygiene	45	663.32	5818	R1 Snack	491	4461.85	81183
R2 Personal Hygiene	39	1029.47	8237	R2 Eat Breakfast	47	497.06	13649
Study	9	922.71	8133	-	-	-	-
Wash_Bathtub	1	33.09	219	-	-	-	-
Watch TV	31	3228.77	17879	-	-	-	-

Analyzed data characteristics show that each dataset is different from others in respective total time duration, deployed sensors, activity count, activity time, and activity sensor events. All these data attributes effect internal processing of classifiers based on their design intensions.

Table 2 The ISL & House_n datasets House A, and Subject1. The 'Num' column shows activity count, the 'Time' column shows activity time in minutes, and the 'Sensor' column shows activity sensor events

HouseA				Subject1			
Activities	**Num**	**Time**	**Sensor**	**Activities**	**Num**	**Time**	**Sensor**
Idle	-	5817.23	23	Idle	-	20930.6	731
Leaving	33	19664.27	84	Toileting	82	161.5	323
Toileting	114	155.38	402	Washing Dishes	7	42.23	67
Showering	23	136.38	59	Preparing Breakfast	14	182.42	147
Brush Teeth	10	9.78	22	Preparing Lunch	17	524.37	497
Sleeping/Go to bed	24	7914.37	183	Preparing Dinner	8	136.72	122
Prepare Dinner	9	306.47	128	Preparing a Snack	13	58.43	66
Snack	12	24.33	50	Preparing a Beverage	15	55.47	77
Prepare Breakfast	20	39.42	122	Dressing	24	88.7	121
Eating	1	22.56	0	Bathing	18	343.93	224
Drink	20	12.23	63	Grooming	37	216.98	302
Load Washingmachine	3	4.01	7	Cleaning	8	149.67	145
Load Dishwasher	5	31.85	15	Doing Laundry	19	146.12	172
Unload Dishwasher	4	15.23	27	Going out to Work	12	2.87	25
Store Groceries	1	1.183	3	-	-	-	-
Unload Washingmachine	4	3.27	9	-	-	-	-
Receive Guest	3	424.93	65	-	-	-	-

4. Classifiers for Activity Recognition

In this section, we briefly introduce the applied classifier for the domain of activity recognition with preferred settings for our experiments. The detail of each classifier is given as:

Artificial Neural Networks (ANN): It is an information processing network of artificial neurons connected with each other through weighted links. In activity recognition, multilayer neural network with back propagation learning algorithm is utilized to recognize the human activities [5]. The structure of the network, number of hidden layers, and number of neuron in each layer effects the learning of different activities. The activation of the neurons in the network depends on the activation function []. We train multi-layer neural network through back propagation learning method and weights are updated by the following equation:

$$\Delta w_{ki} = -c\left[-2\sum_{j}\left\{\left(y_{j(desired)} - y_{j(actual)}\right)f'\left(act_j\right)w_{ij}\right\}f'(act)_i x_k\right] \quad (1)$$

Where Δw is the weights adjustment of the network links. In our network, we used one hidden layer, twenty neurons, tangent sigmod function as an activation function as given below:

$$\varphi(v) = \tanh\left(\frac{v}{2}\right) = \frac{1 - \exp(-v)}{1 + \exp(-v)} \quad (2)$$

Learning of the network is limited to maximum 1000 epochs. The multi-layer neural network can be seen as an intuitive representation of a multi layer activity recognition system. The number of correctly classified activities depends on the number of training instances during the learning phase.

Hidden Markov Model (HMM): It is a generative probabilistic graph model that is based on the Markov chains process [2]. Model is based on the number of states and their transition weight parameters. Parameters are learned thorough observation and following parameters are required to train the model:

$$\lambda = \{A, B, \pi\} \quad (3)$$

Where λ is graphical model for activity recognition, A is a transition probability matrix, B represents the output symbol probability matrix, and π is the initial state probability [2]. We used Baum-Welch algorithm to determine the states and transition probabilities during training of HMM. The i^{th} classification of an activity is given as:

$$\lambda_i = \{A_i, B_i, \pi_i\}, \qquad i = 1, \ldots, N \quad (4)$$

Conditional Random Fields (CRF): It is a discriminative probabilistic graph model for labeling the sequences. The structure of the CRF is similar to HMM but learning mechanism is different due to absence of the hidden states [2]. In CRF model, the conditional probabilities of activity labels with respect to sensor observations are calculated as follows:

$$p(y_{1:T}|x_{1:T}) = \frac{1}{Z(x_{1:T}, w)} exp\left\{\sum_{j=1}^{N_f} w_j F_j(x_{1:T}, Y_{1:T})\right\} \quad (5)$$

In equation 5, Z denotes normalized factor and $F_j(x_{1:T}, Y_{1:T})$ is a feature function. To make the inference in the model, we compute the most likely activity sequence as follows:

$$y^*_{1:T} = argmax_{y'_{1:T}} p(y'_{1:T}|x_{1:T}, w) \quad (6)$$

Support Vector Machine (SVM): SVM is statistical learning method to classify the data through determination of a set of support vectors and minimization of the average error [6]. It can provide a good generalization performance due to rich theoretical bases and transferring the problem to a high dimensional feature space. For a given training set of sensors value and activity pairs, the binary linear classification problem require the following maximum optimization model using the Lagrangrian multiplier techniques and Kernel functions as:

$$\text{Maximize (w.r.t } \alpha) \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=0}^{n} \sum_{j=1}^{n} \alpha_i y_i \alpha_j y_j K(x_i, x_j) \tag{7}$$

$$\text{Subject to:} \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \tag{8}$$

Where K is the kernel function that satisfies $K(x_i, x_j) = \Phi^T(x_i)\Phi(x_j)$. In our case, we used radial basis function (RBF) for recognizing the activities.

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{(2\sigma^2)}\right) \tag{9}$$

Activity recognition is multi-class problem so we adopt “one-versus-one” method to classify the different activities. Classification of the final activity class is based on the voting mechanism and maximum vote of a class determined the activity label.

5. Results and Evaluation

In this section, we show the effect of data dimensions through demonstrating how performances of activity recognition techniques are influenced by the dataset characteristics. We split the dataset using the ‘leave one day out’ approach; therefore, the sensor readings of one day are used for testing and the remaining days for training. Figs. 1, 2, 3 and 4 show the experimental results for the *Tulum2009, Twor2009, House A* and *Subject1* datasets characteristics respectively. In each dataset, for each activity, accuracies of ANN, HMM, CRF and SVM are illustrated in the following paragraphs.

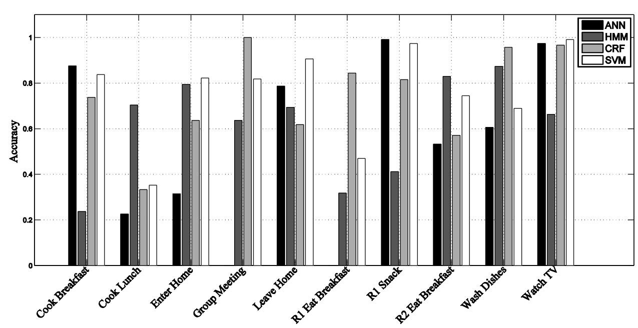

Fig. 1. *Tulum2009* activity based accuracy of classifiers

In our experiments ANN shows high diversity in its performance. It performs better on the set of activities whose training instances are high in the dataset while its performance is insignificant for the recognition of those activities whose training examples are few in the dataset. Overall performance of ANN on *Tulum2009* is 81.09% and it correctly classified “R1 Snack” activity; however, it could not recognize the “Group Meeting” activity. The training instances for these activities are 491 and 11 respectively that affects the ANN classification process. In case of *Twor2009*, training instances for “Meal Preparation” are 118 and ANN outperforms all other classifier on the identification of this activity. While on the same dataset, it could not classify “Study” activity due to less training instances. For *House A*, the overall performance of ANN is low 41.11% as compared to other datasets. However, ANN is the only classifier that 100% classify the “Toileting” activity as its training instances (i.e., 114 samples) are very high as compare to other activities. In case of Subject 1 too, training instances of “Toileting” activity are more (i.e., 82 samples) and ANN performance is better than other classifiers for the recognition of this activity. Although the overall accuracy of ANN varies from dataset to dataset but the better performance of ANN is consistently depend on the high number of training instance in the dataset for a particular activity.

For HMM, the number of deployed sensors effect the activity class distribution by observing their variation during the performed activities. For example, HMM performs well 57.83% in case of *House A* due to small number of deployed sensors (i.e., 14 sensors). It correctly classified “Take Shower”, “Unload Dishwasher” and “Store Groceries”. In case of *Tulum2009,* it outperforms other classifier for “Cook Lunch” and “R2 Eat Breakfast” activities with overall accuracy 56.84%, the number of deployed sensors are 20 in this case. For *Subject 1* and *Twor2009* accuracy of HMM is not significant, the number of deployed sensors in these smart homes are 28 and 71. The large number of miss classified activities are the result of HMM distributions modeling as they are observed in the dataset.

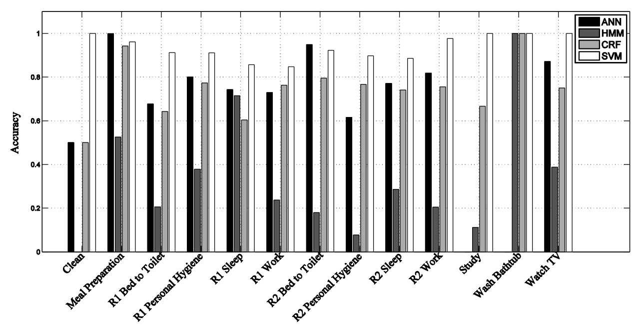

Fig. 2. *Twor2009* activity based accuracy of classifiers

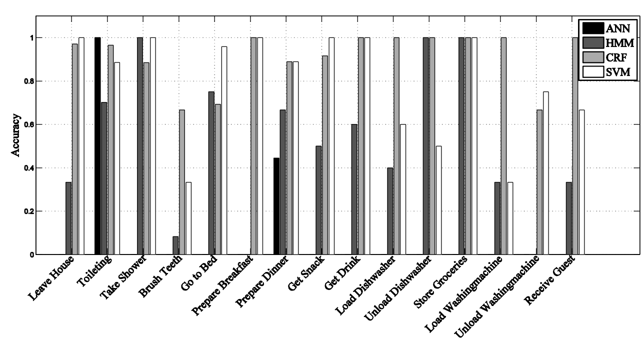

Fig. 3. *House A* activity based accuracy of classifiers

The performance of Conditional Random Fields (CRF) is affected by a set of data characteristics. Its performance does not depend only on single data dimension like senor count or activity occurrence however; its internal processing is based on conditioning of a set of data attributes. CRF outperforms all classifiers in case of *Tulum2009* for "Group Meeting", "R1 Eat Breakfast" and "Wash Dishes". However, other classifiers are better for "Cook Lunch", "Enter Home" and "Leave Home", "R1 Snack" and "R2 Eat Breakfast". In case of *House A*, for "Brush teeth", "Load washingmachine" and "Receive Guest" CRF is superior. For *Subject 1* its performance is high only for "Washing dishes" however for "Cook Lunch", "Enter Home" and "Leave Home", "R1 Snack" and "R2 Eat Breakfast" other classifiers performed better. In case of *Twor2009*, CRF shows low performance for all activities except "Wash Bathtub".

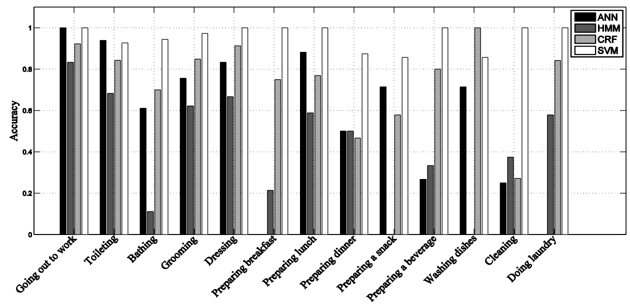

Fig. 4. *Subject 1* activity based accuracy of classifiers

SVM efficiently identified activities in case of *Subject 1* and *Twor2009*, it outperforms all classifier in these datasets except for "Washing Dishes" and "R2 Bed to Toilet" respectively. The performance of SVM is high if performed activities in the dataset are highly discriminative however it is hard for SVM to differentiate between activities that are closely correlated in various data dimensions. Due to this reason in *House A* and *Tulum2009*, for some activities other classifiers are better than SVM as discussed in above paragraphs. For example, in *Tulum2009* it confused "R1 Eat Breakfast" with "R1 Snack" activity both activities are very similar to each other, the second most confused activity is "Cook Lunch". SVM performance is affected if the performed activities are closely interrelated in respect to data dimensions.

Table 3 Overall classifiers accuracy

Classifier / *Dataset*	*ANN*	*HMM*	*CRF*	*SVM*
Tulum2009	0.8109	0.5684	0.8374	0.8889
Twor2009	0.7983	0.3421	0.7780	0.9307
HosueA	0.4111	0.5783	0.9230	0.8919
Subject 1	0.6836	0.5200	0.7745	0.9563

The overall comparison results of different classifiers are presented in Table 3. It specifies the overall accuracy associated with each of the dataset over the four learning techniques. The above results and statistics clearly show that dataset characteristics highly affect the classifiers' individual class level assignments and thus their overall performances.

6. Conclusion and Future Work

In this paper, we have presented the influence of smart home dataset characteristics on the performance of the classifiers. We conclude that the nature of a given dataset plays an important role on the classification accuracy of algorithms; therefore, it is imperative to choose an appropriate algorithm for a particular dataset. We have identified some general characteristics of a dataset that can be useful in selecting the most suitable algorithm as per the nature of underlying dataset. To assess the performance of the four machine learning methods, we applied each classifier on four different datasets from three smart home projects. We analyzed the results of the experiments and provided the explanation of those results for each classifier. It facilitates the researchers to understand the variability in classifiers performances influenced by dataset characteristics.

In future, we would like to devise a framework that can recommend the most suitable classifier for the candidate dataset by analyzing the patterns in the dataset.

Acknowledgement

This research was supported by the MKE (Ministry of knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency)" (NIPA-2009-(C1090-0902-0002)).

References

1. Mihail, P., and Elena, F. Linking Clinical Events in Elderly to In-home Monitoring Sensor Data: A Brief Review and a Pilot Study on Predicting Pulse Pressure, Journal of Computing Science and Engineering. pp. 180-199 (2008)
2. Kasteren, T.L.M. van. Englebienne, G., and Krose, B.J.A.: An Activity Monitoring System for Elderly Care using Generative and Discriminative Models. In Journal of Personal and Ubiquitous Computing. (2010)
3. Krose, B.J.A., Kasteren, T.L.M., Gibson, C.H.S., and Dool. T.: CARE: Context Awareness in Residences for Elderly. In the proceeding of 6th International Conference of the International Society for Gerontechnology. pp. 101-105 (2008)
4. Munguia, E., Tapia, Intille, S. S., and Larson, K.: Activity recognition in the home setting using simple and ubiquitous sensors. Lecture Notes in Computer Science. pp.158-175 (2004)
5. Helal, S. Kim, E. and Hossain, S.: Scalable Approaches to Activity Recognition Research. In proceedings of the workshop of How to do good activity recognition research? Experimental methodologies, evaluation metrics, and reproducibility issues. (2010)
6. Fleury,A. Vacher, M. and Noury, N.: SVM-Based Multimodal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms, and First Experimental Results. IEEE Transactions on Information Technology in Biomedicine (2010)
7. Cook. D., CASAS Smart Home project. [Online]. http://www.ailab.wsu.edu/casas/ [June 10, 2012]
8. Kasteren T., Noulas A., Englebienne G., Ben K.: Accurate Activity Recognition in a Home Setting. In proceeding of International Conference Ubiquitous Computing. pp. 21-24. (2008)
9. K Larson. House_n. [Online]. http://architecture.mit.edu/house_n/ [June 10, 2012]
10. Kasteren, T.L.M. van. Englebienne, G. and Krose, B. J. A.: Activity Recognition Using Semi-Markov Models on Real World Smart Home Datasets. In Journal of Ambient Intelligence and Smart Environments. pp.311–325 (2010)
11. Zheng, H.Wang, H. and Black, N.: Human activity detection in smart home environment with self-adaptive neural networks. In proceedings of the IEEE International Conference on Networking, Sensing and Control. pp. 1505-1510 (2008)
12. Chen, C., Das, B., and Cook, D. J.: A Data Mining Framework for Activity Recognition in Smart Environments. In proceeding of the International Conference on Intelligent Environments. pp. 80-83 (2010)
13. Cook, D. and Holder, L.: Sensor selection to support practical use of health-monitoring smart environments. In Journal of Data Mining and Knowledge Discovery. (2011)