Neural Processing Letters (2020) 51:1771–1787
https://doi.org/10.1007/s11063-019-10185-8

Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient

Yaqing Liu[1,2] · Yong Mu[1] · Keyu Chen[1] · Yiming Li[1] · Jinghuan Guo[1]

Published online: 7 January 2020

Abstract
In the case of a smart home, the ability to recognize daily activities depends primarily on the strategy used for selecting the appropriate features related to these activities. To achieve the goal, this paper presents a daily activity feature selection strategy based on the Pearson Correlation Coefficient. Firstly, a daily activity feature is viewed as a vector in Pearson Correlation Coefficient formula. Secondly, the relation degree between daily activity features is obtained according to weighted Pearson Correlation Coefficient formula. At last, redundant features are removed by the relation degree between daily activity features. Two distinct datasets are adopted to mitigate the effects of the coupling of the dataset used and the sensor configuration. Three different machine learning techniques are employed to evaluate the performance of the proposed approach in activity recognition. The experiment results show that the proposed approach yields higher recognition rates and achieves average improvement F-measures of 1.56% and 2.7%, respectively.

Keywords Activity recognition · Feature selection · Pearson Correlation Coefficient · Smart home

1 Introduction

The main purpose of Ambient Assisted Living (AAL) is to support independent living and alleviate some of the problems associated with aging. It is widely considered to be an effective solution to some of the problems associated with supporting an aging population [1, 2]. As an application of AAL, smart home is designed to improve energy management and the safety, comfort, and convenience of the residents at low costs. Smart home features a number of ambient sensors which are fixed in rooms. Corresponding ambient sensors are activated when the resident engages in daily activities. Energy and living quality improvements premise the daily activity recognition. Hence, the foundation of smart homes is based on the ability to correctly recognize the user's daily activities [3].

✉ Jinghuan Guo
guojinghuan@dlmu.edu.cn

[1] School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

[2] Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Zigong 643000, China

Daily activity recognition includes stages of raw sensor data collection, preprocessing and segment, feature extraction and selection, classifier training and data classification [4]. The performance of activity recognition depends on the daily activity features extracted and selected. There have been a number of studies on feature extraction [5]. However, there have been few studies on feature selection with respect to daily activities. To select the features that are actually useful for recognizing the daily activities of users, we propose a daily activity feature selection strategy based on the Pearson Correlation Coefficient.

The major contributions of this paper can be summarised as follows. This paper proposes an improved Pearson Correlation Coefficient and applies it to daily activity feature selection. Two distinct datasets are adopted to mitigate the effects of the coupling of the dataset used and the sensor configuration. Three different machine learning techniques are employed to evaluate the performance of the proposed approach in activity recognition. The experiment results show that the proposed approach yields higher recognition rates.

The rest of the paper is organized as follows. Section 2 describes related works. In Sect. 3, the process of activity recognition is described. Further, the proposed daily activity features selection approach is explained in Sect. 4. Section 5 describes the tests performed to evaluate the proposed strategy and the results obtained. Section 6 discusses the results obtained. Finally, Sect. 7 lists the conclusions of the study.

2 Related Work

Methods for activity recognition in smart homes can be classified as being based on methods that are driven either by knowledge or by data [6–18].

Knowledge-driven methods focus on establishing activity recognition models based on the domain knowledge of experts in the relevant fields and then recognizing the activities through logical reasoning. The activity recognition model established in the case of a knowledge-driven method is usually based on logic theories. A real-time continuous activity recognition model based on multisensor data streams in knowledge-driven smart homes was proposed by Chen et al. [19]. In addition, knowledge-driven approaches often use ontologies as tools. An ontology-based model for recognizing the activities of the elderly was proposed by Latfi et al. [20]. An ontology that is generated automatically based on the features of an "activities of daily life" classifier for activity recognition has also been proposed [21]. To improve the accuracy of activity recognition, Gayathri et al. [22] used a probabilistic ontology and Rodriguez et al. [23] used a fuzzy ontology to model the contexts of daily activities. Therefore, the knowledge-driven approach is a top-down one. However, it is poor in terms of dealing with uncertainty and time information. Safyan et al. [24] proposed a complete activity model derived from a generic activity model for sequential and parallel activities by exploiting different knowledge engineering techniques, ontology-based temporal formalisms, and data-driven techniques. Chiang et al. [25] proposed a feature-based knowledge transfer framework for cross-environment activity recognition. Meditskos et al. [26] used the web ontology language (OWL) to capture the domain relationships between low-level observations and high-level activities, while context aggregation and activity interpretation were performed based on context-aware fusion.

In contrast, data-driven approaches collect data from a large number of sensor streams and organize them to form information. All the related information is then refined, and machine learning algorithms are trained using this information to obtain an automated decision model based on the data [27]. Brdiczka et al. proposed a

framework for acquiring and developing different hierarchical context models in intelligent environments [28]. In addition, a real-time algorithm for automatically identifying physical activities has also been proposed [29]. The data-driven approach is thus a bottom-up one with a strong ability to deal with uncertainty and time information.

Data-driven methods can be classified into two types: generative methods and discriminative methods. With respect to the generative mode, Patterson et al. proposed several hidden Markov models (HMMs) for activity recognition [30]. Lu et al. proposed a method for extracting the latent features from sensor data using a beta process HMM [31]. Van Kasteren et al. proposed a multilayer HMM for activity recognition [32]. In the case of discriminative methods, a new and effective feature selection algorithm for the m-estimates-based conditional random field was proposed for identifying the most important features for activity recognition [33]. Fahad et al. [34] proposed a binary solution based on a support vector machine for simplifying the problem of classifying correct/incorrect assignments. Although the model works better when dealing with uncertain and incomplete data, it requires a very large training dataset for optimization. Exploiting the developments in neural networks, Bourobou et al. employed a relevant and efficient unsupervised learning method called the k-pattern clustering and an artificial neural network based on Allen's temporal relationships to recognize and predict user activities [35]. Several deep learning models have also been used for activity recognition. For instance, a back propagation neural network was proposed for representing and recognizing human activities from observed sensor sequences [36]. Chen et al. [37] proposed using deep learning techniques to automatically extract high-level features from binary sensor data. Hassan et al. [38] proposed the Deep Belief Network model for human activity recognition. An ensemble of Long Short-Term Memory (LSTM) networks [39] has also been proposed for activity recognition. The LSTM recurrent neural network is used to analyze sensor readings from accelerometers and gyroscopes and to identify human activity on this basis. In addition, position-aware methods have been developed for improving recognition accuracy [40].

For both knowledge-based approaches and data-driven approaches, the activity recognition process consists of four stages. Raw data related to sensor events are collected in the form of a stream whenever a daily activity occurs. In the second stage, this stream is separated into a number of subsequences. Each subsequence corresponds to one instance of an entire activity. In the third stage, a set of daily activity features is determined. In the last stage, a model is built for activity recognition.

The approach proposed in this study contributes to the third stage. The performances of activity recognition depend on both the feature extraction and the classification stages in the context of streaming data [5]. Approaches for feature extraction usually discretize data from the sensors into time slices of constant or variable length, and each time slice is labeled with only one activity [5]. In most cases, feature selection follows feature extraction because feature selection is able to optimize feature space. On one hand, similar features are almost repeated to activity recognition improvement. On the other hand, different features don't equally contribute to activity recognition improvement and even some features restrict activity recognition performance. In addition, space features of daily activity are extracted from sensor events stream. Feature selection leads that some features are discarded, which will prompt that corresponding sensors should be redundant in smart homes. In conclusion, feature selection improves not only activity recognition but also saves the use of unrelevant sensors in smart homes.

3 Smart Homes and Daily Activity Recognition

In a smart home, noninvasive sensors (e.g., infrared motion sensors and temperature sensors) are installed in different parts of the house. When a daily activity of a resident is underway, the corresponding sensors are activated, and they emit sequential data. In Fig. 1, the sensor events activated by the daily activity "Toilet" are shown. When a sensor is activated, the activation date, activation time, name of the sensor, and the emitted data value are stored. For example, the first activated sensor is "MA015", which recorded the value "ON" at time "11:09:18.9523" on 2012-07-20 for activity "Toilet" (see Fig. 1).

Daily activity recognition aims to deduce which daily activities are underway when a sequence of sensors are activated. This is usually a classification task. The daily activity recognition process typically consists of three stages.

In the first stage, a sequence of sensor events is divided into a number of subsequences. Each subsequence corresponds to an entire daily activity. In the second stage, the features related to the daily activity are determined. Generally, these features can be divided into temporal features and space features [41]. The start time and duration of an activity instance are common temporal features while the data from the installed sensors are common space features. The temporal and spatial features are used to characterize the daily activity instances. In the last stage, the daily activity recognition model (classifier) is first built and then trained using the daily activity instances. Finally, the trained recognition model is employed to assign a daily activity label to each of the test activity instances.

In this study, we focused on feature selection related to the daily activities.

Fig. 1 Sequence of events related to activated sensor for daily activity "toilet."

2012-07-20 11:09:18.9523 MA015 ON **Toilet begin**
2012-07-20 11:09:20.094157 LS015 45
2012-07-20 11:09:20.156992 MA015 OAF
2012-07-20 11:09:21.205007 LS015 43
2012-07-20 11:09:21.252048 MA015 ON
2012-07-20 11:09:23.281336 LS015 42
2012-07-20 11:09:25.563333 MA015 OAF
2012-07-20 11:09:31.335903 LS015 41
2012-07-20 11:09:31.387414 MA015 ON
2012-07-20 11:09:35.624681 LS015 42
2012-07-20 11:09:35.711798 MA015 OAF
2012-07-20 11:09:36.565344 MA015 ON
2012-07-20 11:09:38.245902 LS015 49
2012-07-20 11:09:41.634772 LS015 40
2012-07-20 11:09:41.702395 MA015 OAF
2012-07-20 11:09:42.184108 LS015 42
2012-07-20 11:09:42.242239 MA015 ON
2012-07-20 11:09:45.008733 LS015 59
2012-07-20 11:09:45.069491 MA015 OAF
2012-07-20 11:09:45.642422 MA015 ON
2012-07-20 11:09:47.640922 LS015 43
2012-07-20 11:09:48.749309 LS015 41
2012-07-20 11:09:48.811984 MA015 OAF
2012-07-20 11:09:52.714222 MA015 ON
2012-07-20 11:09:53.255856 LS015 42
2012-07-20 11:09:53.837118 MA015 OAF
2012-07-20 11:09:54.372784 LS015 41
2012-07-20 11:09:54.43453 MA015 ON
2012-07-20 11:09:55.14713 LS015 63
2012-07-20 11:09:59.079955 LS015 57
2012-07-20 11:09:59.128578 MA015 OAF **Toilet end**

4 Daily Activity Feature Selection

Formally, let $S=\{s_1, s_2, \ldots, s_n\}$ be the set of sensors installed in a smart home. The feature set of daily activities is defined as $F=\{st, du\}\cup S$. The means of *st* and *du* denote the start time and duration of a daily activity, respectively. For a daily activity, the start time and duration are extracted as the values of the features *st* and *du*. In Fig. 1, the values of *st* and *du* for the daily activity "Toilet" are "11:09:18.9523" and 2,376,558 s, respectively. For $s\in S$, the frequency that *s* is activated in a daily activity is assigned to the feature value of *s*. In Fig. 1, 15 is the feature value assigned to sensor "LS0" (LS015).

The features related to a daily activity are not always correlated with each other. Most previous approaches used all the temporal and spatial features available. However, this can lead to the overfitting of the activity recognition model. Overfitting means that the learned model fits well training data rather than test data. For activity recognition, some learned features are relevant to each other and can fit training data. However, these relevant features are perhaps too rigid to recognize test data well, which leads to overfitting. Reducing relevant features is helpful to loose the learned model and then eliminate overfitting to a certain extent. To avoid the problem of overfitting, we propose using the Pearson Correlation Coefficient (PCC) to reduce the number of daily activity features.

The PCC, which can be calculated using the expression given in Eq. (1), is used to evaluate the linear correlation between two variables *X*, *Y*. The function *COV(X, Y)* is the covariance of *X* and *Y*. σ_X and σ_Y are the deviations of *X* and *Y*, respectively, while μ_X and μ_Y are the respective means. $\rho_{X,Y}$ ranges from $+1$ to -1. A value of $+1$ implies that *X* is completely positively linearly correlated to *Y*. On the other hand, a value of 0 indicates that *X* is not linearly correlated to *Y* at all. Finally, a value of -1 implies that *X* is completely negatively linearly correlated to *Y*. In most cases, *X* and *Y* show an extremely strong correlation to each other when $\rho_{(X,Y)}$ is greater than 0.8. Further, *X* and *Y* can be said to be strongly correlation to each other when $\rho_{(X,Y)}$ is greater than 0.6.

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X-\mu x)(Y-\mu y)\right]}{\sigma_X \sigma_Y} \tag{1}$$

For a daily activity feature $f\in F$, Algorithm (1) is used to determine whether *f* should be kept in *F* or not. Given a set of daily activity features *f*, F_f denotes the set of daily activity features that are each (extremely) strongly correlated with *f*. For any $f_1, f_2\in F_f$, $|F_{f1}|$ is different from $|F_{f2}|$. Different $f^*\in F_f$ would have different effects on *f*. The smaller $|F_f^*|$ is, the stronger will its influence be on *f*. Therefore, a weighted Pearson Correlation Coefficient (WPCC) is proposed. WPCC can be calculated as given in Eq. (2).

$$WPCC(X) = \sum_{j=1}^{n} \left(\frac{\rho_{X,Y_j}}{|F_{Y_j}|} \right) * |F_X|, \quad Y_j \in F_X \tag{2}$$

For Algorithm (1), two important thresholds, namely, α and β, are necessary. Here, α is the threshold degree of linear correlation. For two daily activity features $f, f^*\in F$, α is used to determine whether f^* is correlated to *f*. For a daily activity feature *f*, β is the threshold number of daily activity features that are correlated to *f*. Algorithm (1) aims to find the distinct daily activity features. thus, the smaller $|F_f|$ or *WPCC(f)* is, the more distinct *f* will be. Hence, β is used to determine whether *f* is a distinct daily activity feature.

Table 1 Frequencies that daily activity features are activated

	f_1	f_2	f_3	f_4
a_1	1	1	3	8
a_2	2	3	6	1
a_3	2	2	5	7

Table 2 $\rho_{X,Y}$ values of daily activity features

	f_1	f_2	f_3	f_4
f_1	1	0.866	0.944	− 0.60
f_2	0.866	1	0.981	− 0.92
f_3	0.944	0.981	1	− 0.83
f_4	− 0.60	− 0.924	− 0.835	1

Algorithm 1. *solvePCCSet*

Input: $\{a_1, a_2, \ldots, a_m\}$, a set of daily activities
F, set of daily activity features.
$f \in F$, a daily activity feature.
α, threshold degree of linear correlation
β, threshold number of daily activity features that are correlated to f.
w, w is $|F_f|$ or $WPCC(f)$

Output: F.

1. $F_f \leftarrow \varnothing$; //$F_f$ is a set of daily activity features. $f^* \in F$ is (extremely) strongly correlated with f.
2. $X \leftarrow (frequency(a_1, f), frequency(a_2, f), \ldots, frequency(a_m, f))$;
3. //$frequency(a_i, f)$ is used to determine the frequency that f is activated when a_i is underway.
4. **for each** f^* **in** $F/\{f\}$
5. $Y \leftarrow (frequency(a_1, f^*), frequency(a_2, f^*), \ldots, frequency(a_m, f^*))$;
6. **if** $\rho_{X,Y} \geq \alpha$ **then**
7. $F_f \leftarrow F_f \cup \{f^*\}$
8. **end if**
9. **end for**
10. **if** $w \geq \beta$ **then**
11. $F \leftarrow F/\{f\}$
12. **end if**
13. **return** F

To explain Algorithm 1, an example is given in Table 1. $F = \{f_1, f_2, f_3, f_4\}$ is an initial set of daily activity features. $A = \{a_1, a_2, a_3\}$ is a training set. For $a \in A$ and $f \in F$, the frequency that f is activated when a is underway is shown in Table 1. The $\rho_{X,Y}$ values for the daily activity features are given in Table 2. It can be seen from the table that f_1 is extremely strongly correlated to f_2 and f_3. If $|F_f|$ is assigned to w and a value of 1 is assigned to β, F

will change to $F=\{f_2, f_3, f_4\}$. $WPCC(f_1)$ can be determined as $(0.866/2+0.944/2)*2=1.81$. If $WPCC(f_1)$ is assigned to w and a value of 1 is assigned to β, F will change to $F=\{f_2, f_3, f_4\}$.

5 Evaluation

5.1 Datasets

To verify the proposed approach, we used two common datasets: "hh101" and "hh102." These datasets have been published by the Washington State University [42]. Table 3 describes the basic statistics for these two datasets. The values listed under "Sensors" correspond to the number of sensors used and their respective types. Similarly, the values listed under the column "Activity Categories" correspond to the number of activity classes involved, while the values listed under the column "Daily Activities" correspond to the number of activity instances involved. Finally, the values on the right side of the table are the times at which the data were collected.

For the "hh101" dataset, the following identifier categories were considered.

(1) Identifiers with names starting with "M" indicate infrared motion sensors—M001–M012.
(2) Identifiers with names starting with "T" indicate temperature sensors—T101–T105.
(3) Identifiers starting with "BA" indicate sensor battery levels—BATP001–BATP016, BATP102–BATP103, BATV001–BATV016, BATV102–BATV103.
(4) Identifiers starting with "D" indicate magnetic door sensors—D001-D002.
(5) Identifiers starting with "LS" indicate light sensors—LS001–LS016.
(6) Identifiers starting with "MA" indicate wide-area infrared motion sensors—MA013–MA016.

The atomic activities involved include "*Bathe*"("*B*"), "*Dress*"("D"), "*Enter_Home*"("*E_H*"), "*Eat_Dinner*"("*E_D*"), "*Leave_Home*"("*L_H*"), "*Eat_Breakfast*"("*E_B*"), "*Groom*"("*G*"), "*Sleep*"("*S*"), "*Read*"("*R*"), "*Toilet*"("*T*"), "*Wash_Breakfast_Dishes*"("*W_B_D*"), "*Wash_Dinner_Dishes*"("*W_D_D*"), "*Watch_TV*"("*W_T*"), "*Work_At_Table*"("*W_A_T*"), "*Cook*"("*C*"),"and *Eat*"("*E*").

For the "hh102" dataset, the following identifier categories were considered.

(1) Identifiers with names starting with "M" indicate infrared motion sensors—M001–M002, M004–M008, M011–M012, M015–M019, M021–M022.
(2) Identifiers with names starting with "T" indicate temperature sensors—T101–T105.

Table 3 Statistical information related to datasets "hh101" and "hh102"

Dataset	Sensors	Activity categories	Daily activities	Measurement time
"hh101"	75 (6 categories)	16	1645	60 days
"hh102"	111 (7 categories)	12	787	58 days

(3) Identifiers starting with "BA" indicate sensor battery levels—BATP001–BATP011, BATP013, BATP014, BATP016–BATP023, BATP105, BATV001–BATV023, BATV102–BATV105.
(4) Identifiers starting with "D" indicate magnetic door sensors—D002–D003, D005–D006.
(5) Identifiers starting with "LS" indicate light sensors—LS001-LS023.
(6) Identifiers starting with "MA" indicate wide-area infrared motion sensors—MA003, MA009, MA010, MA013–MA014, MA020, MA023.
(7) Identifiers starting with "L" and "LL" indicate light switches—L001–L005, LL001, LL005.

The activities involved include "*Bathe*"("*B*"), "*Dress*"("*D*"), "*Eat_Breakfast*"("*E_B*"), "*Eat_Dinner*"("*E_D*"), "*Groom*"("*G*"), "*Sleep*"("*S*"), "*Take_Medicine*"("*T_M*"), "*Toilet*"("*T*"), "*Wash_Breakfast_Dishes*"("*W_B_D*"), "*Wash_Dinner_Dishes*"("*W_D_D*"), "*Watch_TV*"("*W_T*"), and "*Work_At_Table*"("*W_A_T*").

5.2 Evaluation Setup

The activity recognition performance of the proposed approach was evaluated while employing all the features. Three classifiers, namely, Naive Bayes (NB), C4.5, and Random Forest (RF), were used as the activity recognition models. The tool employed was Weka 3.8 [43], and 10 cross-validation attempts were made to account for the datasets. The performance indicators used were the accuracy, precision, and F-measure.

5.3 Results

The accuracy, precision, and F-measure values corresponding to the proposed method for the "hh101" and "hh102" datasets are listed in Tables 4 and 5, respectively. Each table is divided into three groups, which are AF, PCC, and WPCC. All Features (AF) means that the daily activities were recognized using all the daily activity features $F = \{st, du\} \cup S$ which is defined in Sect. 4. For AF, there are no α and β. PCC and WPCC means that the daily activities were recognized using selected features from F using Algorithm 1. When a value of 0.6 was assigned to α, PCC was considered "strong correlation PCC (SPCC)" and WPCC was called "extremely strong correlation PCC (ESPCC)." When a value of 0.8 was assigned to α, PCC was said to be "strong correlation WPCC (SWPCC)" and WPCC was called as "extremely strong correlation PCC (ESWPCC)." For PCC, one of the following values was assigned to β: 2, 5, and 10. For WPCC, one of the following values was assigned to β: 1, 2, 5, 10, and 15.

The average accuracy, precision, and F-measure values are shown in Figs. 2, 3, 4 and 5. The average accuracy values obtained using SPCC, ESPCC, SWPCC, and ESWPCC were higher than those obtained using AF for dataset "hh101" when either C4.5 or NB were employed. The average accuracies obtained using SPCC and SWPCC were higher than the average accuracy obtained using AF for dataset "hh101" when RF was employed. The average accuracy values obtained using ESPCC and ESWPCC were slightly lower than the average accuracy value obtained using AF for dataset "hh101" when RF was employed.

The average accuracy values obtained using SPCC, ESPCC, SWPCC, and ESWPCC were superior to the average accuracy value obtained using AF for dataset "hh102" irrespective of the classifier employed. Further, the average F-measure values obtained using

Table 4 Accuracy, Precision, F-measure values for recognition using three classifiers in case of dataset "hh101"

Metric	Feature	α	β	NB (%)	C4.5 (%)	RF (%)
Accuracy	AF	–	–	*74.16*	*83.58*	86.86
	PCC	0.8	5	74.77	84.37	86.56
			2	75.13	84.8	87.05
			10	75.19	84.43	86.99
		0.6	5	76.1	85.95	87.35
			2	75.56	85.77	87.78
	WPCC		10	74.46	84.13	86.8
			5	*74.16*	84.55	*86.26*
		0.8	2	75.13	84.8	87.05
			1	75.13	84.8	87.05
			10	75.31	83.7	86.44
			5	75.5	85.22	87.29
		0.6	2	75.86	85.22	87.23
			1	75.56	85.77	87.78
Precision	AF	–	–	80	*82*	84.2
	PCC	0.8	5	*79.3*	82.7	84.5
			2	79.6	83.1	85
			10	79.7	82.7	85
		0.6	5	80	84.3	85.5
			2	*79.3*	84.4	85.7
	WPCC		10	80.2	82.6	84.4
			5	*79.3*	83	*83.9*
		0.8	2	79.6	83.1	85
			1	79.6	83.1	85
			10	80.4	*82*	84.5
			5	79.7	83.6	85.3
		0.6	2	79.7	83.8	84.8
			1	*79.3*	84.4	85.7
F-measure	AF	–	–	74.8	*82.7*	85.1
	PCC	0.8	5	74.9	83.4	85.1
			2	75.2	83.8	85.6
			10	75.3	83.5	85.5
		0.6	5	75.9	85	86.1
			2	75.3	84.9	86.4
	WPCC		10	75.1	83.3	85.2
			5	*74.5*	83.7	84.5
		0.8	2	75.2	83.8	85.6
			1	75.2	83.8	85.6
			10	75.8	*82.7*	*84.9*
			5	75.5	84.3	85.8
		0.6	2	75.6	84.3	85.7

For any of accuracy, precision and F-measures, italics value is minimum and underline value is maximum when same classifier is used

Table 5 Accuracy, Precision, F-Measure values for recognition using three classifiers in case of dataset "hh102"

Metric	Feature	α	β	NB (%)	C4.5 (%)	RF (%)
Accuracy	AF	–	–	*75.98*	82.71	*85*
	PCC	0.8	10	77.12	82.97	86.14
			5	78.65	83.35	87.16
			2	76.49	83.6	86.65
			15	76.11	82.71	86.14
			10	77.63	83.22	87.03
		0.6	5	76.74	82.84	86.65
			2	78.52	83.86	88.18
	WPCC		15	76.74	*82.59*	86.27
			10	77.5	82.97	86.65
			5	77.5	83.73	86.4
		0.8	2	78.14	83.48	86.53
			1	77.76	83.48	86.91
			15	75.73	82.84	87.16
			10	77.63	83.1	87.03
			5	77.12	83.22	86.53
		0.6	2	78.39	83.48	87.03
			1	79.79	83.35	88.56
Precision	AF	–	–	*81.6*	82.4	*83.2*
	PCC	0.8	10	82.5	82.6	85.4
			5	85	82.6	86.6
			2	84.6	82.9	85.7
		0.6	15	81.8	82.4	85.1
			10	85.8	82.7	86.2
			5	85	82.4	85.8
			2	83.6	83.3	87.3
	WPCC		15	82.3	*82.2*	85.5
			10	83.1	82.6	85.6
			5	83.2	83.2	85.8
		0.8	2	84.6	82.5	85.5
			1	84.7	82.6	86.7
			15	*81.6*	82.5	86.2
			10	85.4	82.6	86.3
			5	85.3	82.7	85.7
		0.6	2	83.9	83.1	86.5
			1	86	82.9	88.2

Table 5 (continued)

Metric	Feature	α	β	NB (%)	C4.5 (%)	RF (%)
F-measure	AF	–	–	*77.2*	82.5	*82.8*
	PCC	0.8	10	78	82.7	83.9
			5	79.5	82.8	85.5
			2	77.9	83.1	85.2
		0.6	15	77.3	82.5	84.2
			10	78.9	82.8	85.7
			5	78.1	82.5	85.2
			2	77.6	83.5	86.9
	WPCC		15	77.7	*82.4*	84
			10	78.4	82.7	85
			5	78.4	83.4	84.4
		0.8	2	79	82.9	84.8
			1	78.8	82.9	85.4
			15	77	82.6	85.3
			10	78.9	82.8	85.5
			5	78.4	82.8	85.1
		0.6	2	78.2	83.1	85.8
			1	79.6	83	87.6

For any of accuracy, precision and F-measures, italics value is minimum and underline value is maximum when same classifier is used

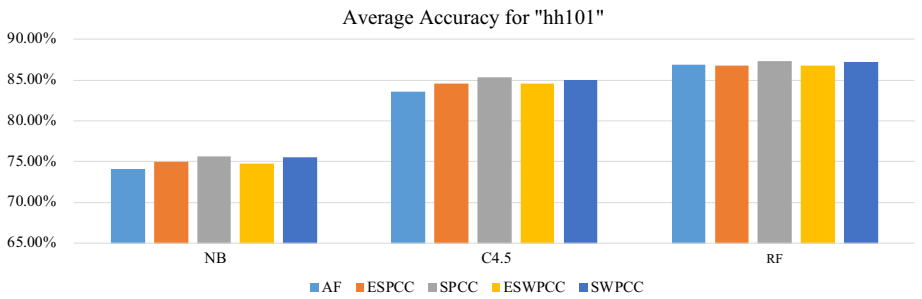

Fig. 2 Average accuracy for "hh101"

SPCC, ESPCC, SWPCC, and ESWPCC were higher than the average F-measure value obtained using AF. This was true for both datasets and all classifiers.

The results demonstrated the effectiveness of the proposed approach. The results clearly show that the distinct sensors were selected and the noisy sensors were discarded and then, robust classifiers were achieved from the distinct sensors.

In addition, the changes in the F-measure value for the pruned number of daily activity features are shown in Figs. 6, 7, 8, 9, 10, 11, 12 and 13. Please note that the y-axis does not start at 0%, which aims to make these trend lines show clearly. For dataset

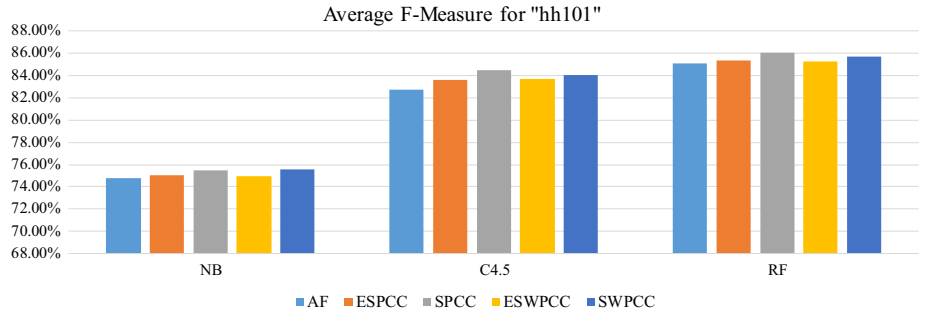

Fig. 3 Average F-measure for "hh101"

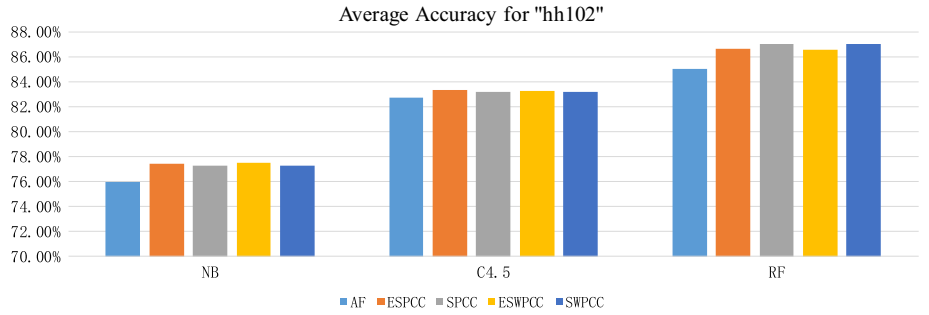

Fig. 4 Average accuracy for "hh102"

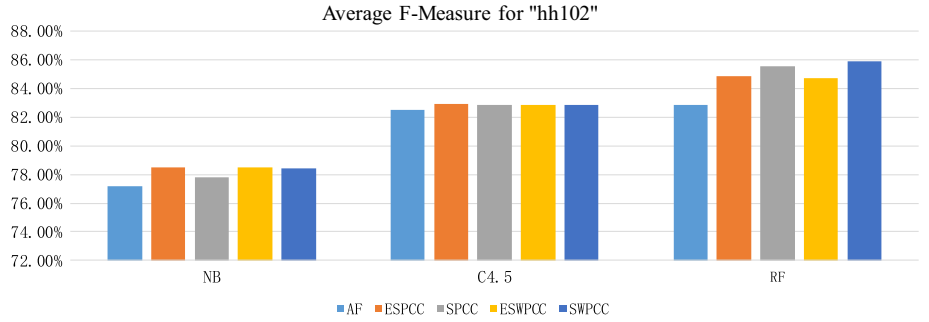

Fig. 5 Average F-measure for "hh102"

"hh101," the F-measure values obtained using SPCC, ESPCC, SWPCC, and ESWPCC remained almost unchanged even after the pruning of the number of daily activity features. On the other hand, for dataset "hh102," the F-measure values obtained using SPCC, ESPCC, SWPCC, and ESWPCC changed significantly and non-monotonously with the pruning of the number of daily activity features. Owing to the different principles of the employed classifiers, the pruned features had different degrees of impact on training classifiers. In Figs. 6, 7, 8, 9, 10, 11, 12 and 13, steep curves were achieved

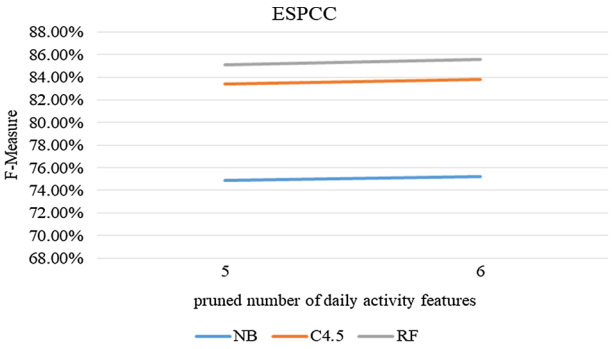

Fig. 6 Changes in F-measure with pruning of number of daily activity features for "hh101" using ESPCC

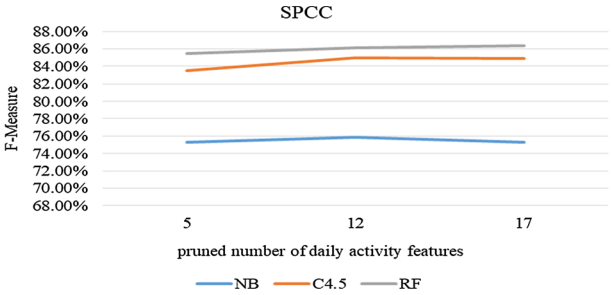

Fig. 7 Changes in F-measure with pruning of number of daily activity features for "hh101" using SPCC

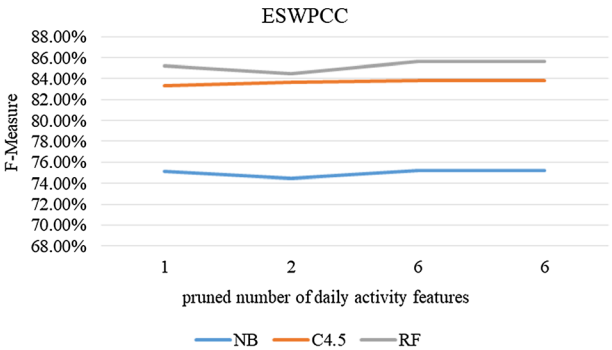

Fig. 8 Changes in F-measure with pruning of number of daily activity features for "hh101" using ESWPCC

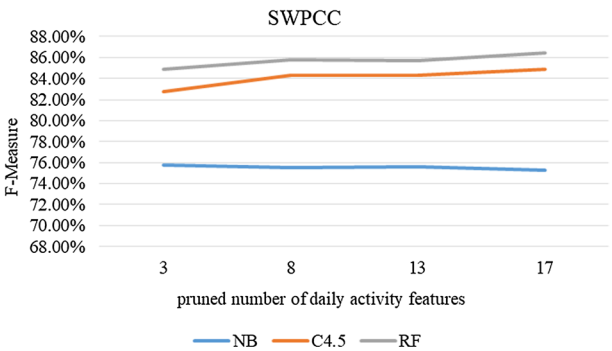

Fig. 9 Changes in F-measure with pruning of number of daily activity features for "hh101" using SWPCC

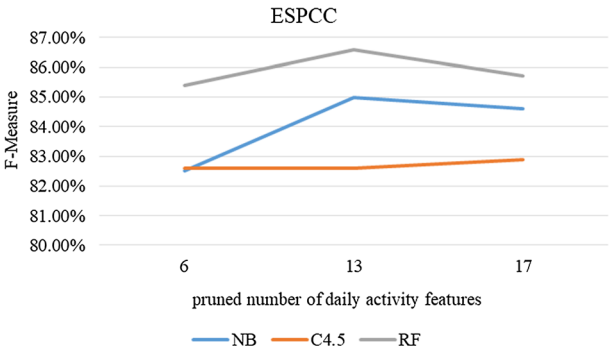

Fig. 10 Changes in F-measure with pruning of number of daily activity features for "hh102" using ESPCC

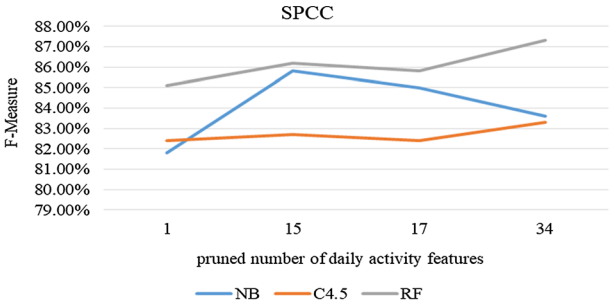

Fig. 11 Changes in F-measure with pruning of number of daily activity features for "hh102" using SPCC

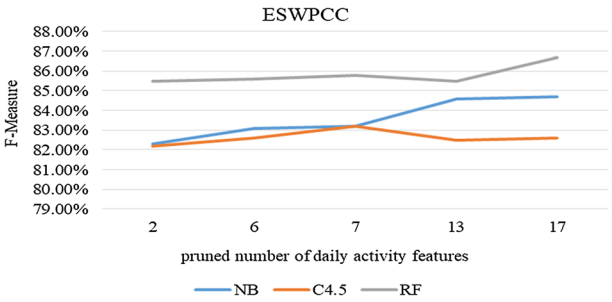

Fig. 12 Changes in F-measure with pruning of number of daily activity features for "hh102" using ESWPCC

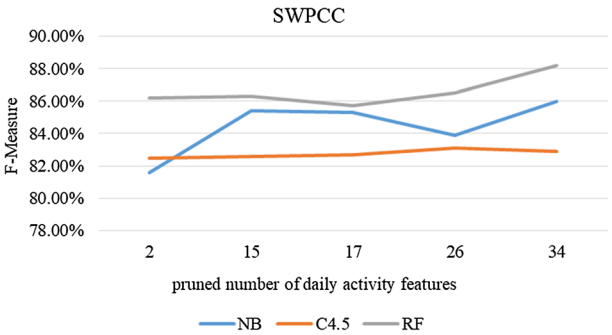

Fig. 13 Changes in F-measure with pruning of number of daily activity features for "hh102" using SWPCC

because some classifiers were sensitive to pruned features. Gentle curves were achieved because some classifiers were not sensitive to the pruned features.

6 Discussion

The slopes in Figs. 6, 7, 8, 9, 10, 11, 12 and 13 reflect how the pruning of the number of daily activity features affects F-measure of activity recognition. In Figs. 6, 7, 8 and 9, F-measures hardly change with the pruning of the number of daily activity features on different classifiers. In Figs. 10, 11, 12 and 13, F-measures hardly change with the pruning of the number of daily activity features on C4.5. F-measures change between 82 and 86% on NB and change between 85 and 88% on RF. In addition, F-measures don't increase or decrease monotonically with the pruning of the number of daily activity features on different classifiers. Hence, the proposed algorithm can be used at the cost of possible insignificant fluctuation of F-measures.

As shown in Figs. 2, 3, 4 and 5, the proposed algorithm slightly improves activity recognition performance. However, Figs. 6, 7, 8, 9, 10, 11, 12 and 13 show that the proposed algorithm can decrease used features, which prompts that corresponding sensors should be dismantled in smart homes. Hence, the proposed algorithm is worth being used in smart home because it can simplify used sensors in smart home.

7 Conclusion

The technology of daily activity recognition has been used in detection and analysis of resident health events, automated clinical assessment, home energy management, etc. It is able to support independent living and alleviate some of the problems associated with aging [44]. It can help to diagnose some behavior problems, e.g. Alzheimer's disease [45]. Also, it can improve smart homes energy management by recognizing daily activity which resident is carrying out [46]. For example, the water heating system should be turned off when the resident is out of home. Therefore the algorithm proposed in this study could help in this endeavor.

Daily activity recognition is worthy work. In this study, we proposed a daily activity feature selection approach based on the Pearson Correlation Coefficient for activity recognition. Three classifiers (Naïve Bayes, Random Forest, and C4.5) were used to evaluate the proposed approach for two distinct datasets (hh101 and hh102). The results showed that the proposed approach can not only finitely improve the activity recognition performance but also simplify sensors in smart homes.

Acknowledgements We thank all the reviewers for their useful comments for improving the manuscript. This work was supported by the National Natural Science Foundation of China (No. 61976124); the Fundamental Research Funds for the Central Universities (No. 3132018194); the Open Project Program of Artificial Intelligence Key Laboratory of Sichuan Province (No. 2018RYJ09); the CERNET Innovation Project (No. NGII20181203).

References

1. Chan M, Campo E, Estève D, Fourniols JY (2008) A review of smart homes- present state and future challenges. Comput Methods Programs Biomed 91(1):55–81
2. Ruyter BD, Zwartkruispelgrim E, Aarts E (2010) Ambient assisted living research in the carelab. Interactions 14(4):30–33

3. Machot FA, Mosa AH, Ali M, Kyamakya K (2018) Activity recognition in sensor data streams for active and assisted living environments. IEEE Trans Circuits Syst Video Technol 28(10):2933
4. Feuz KD, Cook DJ (2017) Collegial activity learning between heterogeneous sensors. Knowl Inf Syst 53(2):337–364
5. Yala N, Fergani B, Fleury A (2017) Towards improving feature extraction and classification for activity recognition on streaming data. J Ambient Intell Humaniz Comput 8(2):177–189
6. Guo SK, Liu YQ, Chen R, Sun X, Wang XX (2019) Improved SMOTE algorithm to deal with imbalanced activity classes in smart homes. Neural Process Lett 50(2):1503–1526
7. Liu YQ, Yi XK, Chen R, Zhai ZG, Gu JX (2018) Feature extraction based on information gain and sequential pattern for english question classification. IET Softw 12(6):520–526
8. Liu YQ, Wang XX, Zhai ZG, Chen R, Zhang B, Jiang Y (2019) Timely daily activity recognition from headmost sensor events. ISA Trans 94:379–390
9. Guo SK, Chen R, Wei MM, Li H, Liu YQ (2018) Ensemble data reduction techniques and multi-RSMOTE via fuzzy integral for bug report classification. IEEE Access 6:45934–45950
10. Deng W, Xu J, Zhao H (2019) An improved ant colony optimization algorithm based on hybrid strategies for scheduling problem. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2897580
11. Chen R, Guo SK, Wang XZ, Zhang TL (2019) fusion of multi-RSMOTE with fuzzy integral to classify bug reports with an imbalanced severity distribution. IEEE Trans Fuzzy Syst. https://doi.org/10.1109/TFUZZ.2019.2899809
12. Deng W, Zhao H, Yang X, Xiong J, Meng S, Bo L (2017) Study on an improved adaptive pso algorithm for solving multi-objective gate assignment. Appl Soft Comput 59:288–302
13. Guo SK, Chen R, Li H, Zhang TL, Liu YQ (2019) Identify severity bug report with distribution imbalance by CR-SMOTE and ELM. Int J Softw Eng Knowl Eng 29(2):139–175
14. Deng W, Zhao H, Li Z, Li G, Yang X, Wu D (2017) A novel collaborative optimization algorithm in solving complex optimization problems. Soft Comput 21(15):4387–4398
15. Zhao H, Zheng J, Xu J, Deng W (2019) Fault diagnosis method based on principal component analysis and broad learning system. IEEE Access. https://doi.org/10.1109/access.2019.2929094
16. Li H, Gao GF, Chen R, Ge X, Guo SK, Hao LY (2019) The influence ranking for testers in bug tracking systems. Int J Softw Knowl Eng 29(1):93–113
17. Zhao H, Yao R, Xu L, Yuan Y, Li G, Deng W (2018) Study on a novel fault damage degree identification method using high-order differential mathematical morphology gradient spectrum entropy. Entropy 20(9):L682
18. Yang C, Liu H, Mcloone S, Chen CL, Wu X (2018) A novel variable precision reduction approach to comprehensive knowledge systems. IEEE Trans Cybern 48(2):661–674
19. Chen L, Nugent CD, Wang H (2012) A knowledge-driven approach to activity recognition in smart homes. IEEE Trans Knowl Data Eng 24(6):961–974
20. Latfi F, Lefebvre B, Descheneaux C (2007) Ontology-based management of the telehealth smart home, dedicated to elderly in loss of cognitive autonomy. In: CEUR workshop proceeding, June 2007
21. Salguero AG, Espinilla M (2018) Ontology-based feature generation to improve accuracy of activity recognition in smart environments. Comput Electr Eng 68:1–13
22. Gayathri KS, Easwarakumar KS, Elias S (2017) Probabilistic ontology based activity recognition in smart homes using markov logic network. Knowl Based Syst 121:173–184
23. Rodrguez ND, Cullar MP, Lilius J, Calvo-Flores MD (2014) A fuzzy ontology for semantic modelling and recognition of human behaviour. Knowl Based Syst 66:46–60
24. Safyan M, Qayyum ZU, Sarwar S, Garcia-Castro R, Ahmed M (2019) Ontology-driven semantic unified modelling for concurrent activity recognition. Multimed Tool Appl 78(2):2073–2104
25. Chiang YT, Lu CH, Hsu YJ (2017) A feature-based knowledge transfer framework for cross-environment activity recognition toward smart home applications. IEEE Trans Hum Mach Syst 47(3):310–322
26. Meditskos G, Kompatsiaris I (2017) iknow: Ontology-driven situational awareness for the recognition of activities of daily living. Pervasive Mob Comput 40:17–41
27. Bao L, Intille SS (2004) Activity recognition from user-annotated acceleration data. In: 2nd international conference on pervasive computing, Linz and Vienna, Austria, April 2004
28. Oliver B, Crowley JL, Patrick R (2009) Learning situation models in a smart home. IEEE Trans Syst Man Cybern Part B Cybern 39(1):56
29. Tapia EM, Intille SS, Haskell W, Larson K, Wright JA, King A, Friedman RH (2007) Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In: IEEE international symposium on wearable computers, Boston, MA, USA, October 2007
30. Patterson DJ, Fox D, Kautz H, Kautz H (2005) Fine-grained activity recognition by aggregating abstract object usage. In: Ninth IEEE international symposium on wearable computers, Osaka, Japan

31. Lu L, Cai QL, Zhan YJ (2017) Activity recognition in smart homes. Multimed Tools Appl 76(22):24203–24220
32. Kasteren TLMV, Englebienne G, Krse BJA (2011) Hierarchical activity recognition using automatically clustered actions. In: 2nd international joint conference on ambient intelligence, The Netherlands, November, 2011
33. Vail DL, Veloso MM, Lafferty JD (2007) Conditional random fields for activity recognition. In: International joint conference on autonomous agents & multiagent systems, May 2007
34. Fahad LG, Khan A, Rajarajan M (2015) Activity recognition in smart homes with self verification of assignments. Neurocomputing 149:1286–1298
35. Bourobou ST, Yoo Y (2015) User activity recognition in smart homes using pattern clustering applied to temporal ann algorithm. Sensors 15(5):11953–11971
36. Fang H, Lei H (2012) BP neural network for human activity recognition in smart home. In: International conference on computer science & service system, August 2012
37. Chen G, Wang A, Zhao S, Liu L, Chang CY (2018) Latent feature learning for activity recognition using simple sensors in smart homes. Multimed Tools Appl 77(12):15201–15219
38. Hassan MM, Huda S, Uddin MZ, Almogren A, Alrubaian M (2018) Human activity recognition from body sensor data using deep learning. J Med Syst 42(6):99
39. Yu G, Thomas P (2017) Ensembles of deep LSTM learners for activity recognition using wearables. In: Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies
40. Chen WH, Baca CAB, Tou CH (2017) Lstm-rnns combined with scene information for human activity recognition. In: IEEE international conference on E-health networking, Dalian, China, October 2017
41. Krishnan NC, Cook DJ (2014) Activity recognition on streaming sensor data. Pervasive Mob Comput 10(Pt B):138–154
42. WSU CASAS Datasets http://ailab.wsu.edu/casas/datasets.html. Accessed 2 Feb 2016
43. Weka 3.8. https://sourceforge.net/projects/weka/. Accessed 29 Apr 2016
44. Sprint G, Cook DJ, Fritz RS, Schmitter-Edgecombe M (2016) Using smart homes to detect and analyze health events. Computer 49(11):29–37
45. Dawidi PN, Cook DJ, Schmitter-Edgecombe M (2016) Automated clinical assessment from smart home-based behavior data. IEEE J Biomed Health Inform 20(4):1188–1194
46. Thomas BL, Cook DJ (2016) Activity-aware energy-efficient automation of smart buildings. Energies 9:624