Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-020-01836-z

ORIGINAL RESEARCH

A novel feature selection method based on comparison of correlations for human activity recognition problems

Athina Tsanousa[1,2] · Georgios Meditskos[2] · Stefanos Vrochidis[2] · Lefteris Angelis[1]

Received: 13 October 2019 / Accepted: 21 February 2020

Abstract
In human activity recognition studies it is important to identify an optimal set with the minimum number of features that will potentially improve the recognition rate. In the current paper we introduce a promising feature selection method that exploits the differences on the correlation structure of the features, between the different classes of the target variable. Using the recordings of triaxial accelerometers and gyroscopes, we extracted several features and created subsets according to the activities performed. For each subset, we calculated the pairwise correlation coefficients of the features and compared the feature correlations of different subsets. By identifying the significantly different correlations we ranked the variables participating in those correlations based on their frequency of appearance and thus created a subset of features that will optimize the performance of a classification algorithm. The method allows the researcher to select the desired number of features to be included in the classification. Two publicly available datasets were used to evaluate the performance of the proposed methodology in binary and multiclass classification problems. The evaluation revealed quite promising results of the methodology that was compared to the performance of the whole feature set and of a feature selection method that has been extensively used in activity recognition studies.

Keywords Feature selection · Activity recognition · Wearable sensors · Machine learning

1 Introduction

Recognition of daily activities is a field that gains increasing attention and is found in many applications, like ambient assisted living systems or localization tasks. Human activity recognition (HAR) can be conducted with the exploitation of cameras (vision-based) or the analysis of sensor recordings (sensor-based) (Chen et al. 2012). Sensor-based HAR applications are more common, due to the availability of specific sensors embedded in all smartphones and devices. Such sensors are characterized as wearable sensors, a category that includes also the ones attached to the human body. Wearable sensors provide measurements in three orthogonal directions and signal is translated in three vectors, responding to the axes of the Cartesian system (Chen et al. 2012).

Accelerometers and gyroscopes are the inertial sensors known to be more competent in recognizing activities. Especially accelerometers are the sensors with the best performance when they are used individually, which is the reason they are more frequently found in relevant literature (Ustev et al. 2013, Chen et al. 2012, Wang et al. 2018). Gyroscopes are usually found in combination with accelerometers and they assist in the recognition of activities not well detected only by accelerometers (Ustev et al. 2013). Accelerometers capture the magnitude and direction of an object in motion; while gyroscopes measure the angular velocity of rotation hence they detect the object's orientation (Wang et al. 2018). Due to the different nature of the two sensors, it was considered more appropriate to apply the proposed methodology separately on each one.

✉ Athina Tsanousa
atsanous@csd.auth.gr

Georgios Meditskos
gmeditsk@iti.gr

Stefanos Vrochidis
stefanos@iti.gr

Lefteris Angelis
lef@csd.auth.gr

[1] School of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

[2] Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece

Published online: 06 March 2020

A generic human activity recognition chain consists of several steps. The first step is to collect or extract the raw data, which in cases of sensor-based HAR consist of three variables responding to each axis of the Cartesian system. The next step is the application of filters and/or the normalization of the data to eliminate signal noise. Next, a time window is selected to extract features. The need for a time window arises from the fact that the duration of a performed activity is bigger than the sensor's sampling rate, making it quite difficult to recognize an activity by selecting samples of specific time points. In order for two signals to be comparable, features should be extracted from each time window. Feature extraction is supposed to retain valuable information from the signals (Lara et al. 2012). An extended list of commonly extracted features that are generally categorized as time domain and frequency domain can be found in (Lara et al. 2012). Following feature extraction, a feature selection method can be applied to identify which features contain useful information for the recognition of activities (Lara et al. 2012). The HAR framework is concluded with the classification process, where a classification algorithm is applied to recognize the activities. In case there is the need to combine information from different sensors, fusion methods are employed. Fusion, which is the combination of information, can be applied at the end of the process to combine the classification results of different sensors or at earlier stages, where it combines the features of different sensors, before the classification step (Mangai et al. 2010).

Focusing on the feature selection stage of the process, feature selection methods are used for three main reasons:

1. To remove redundant features and select the optimal feature set
2. To improve the classification accuracy
3. To decrease the computational time.

A quite common feature selection method in HAR applications is the Correlation Based Feature Selection (CFS) that was suggested in Hall and Smith (1999). In CFS, the correlation of each feature with the target variable is calculated, as well as the pairwise correlations of all features. First, features that are highly correlated with the target variable are selected by the algorithm as optimal and thus create a subset. That subset is further eliminated by removing features highly correlated with each other, resulting in the final optimal set of features that will be used in the classification algorithm. In the original paper, the authors compared their method with a wrapper feature selection one and tested their performance on twelve datasets, suitable for binary and multiclass classification although not from the activity recognition field. The CFS algorithm outperformed the wrapper method in most of the datasets and when combined with two classification algorithms, Naïve Bayes and C4.5 trees. Another method found in HAR applications, is the Fast Correlation Based Feature Selection (FCBF). FCBF that is also based on correlation structures is a fast algorithm that usually selects a small number of features. The algorithm that was introduced in Yu and Liu (2003), shares the same concept as the CFS algorithm. However, FCBF doesn't use pairwise correlation to measure relevance or redundancy among the features but symmetrical uncertainty. The authors evaluated the performance of their method in terms of speed, number of features and classification accuracy in comparison with some other feature selection methods. Although FCBF did not outperform the others in most of the datasets used, it has proven a very effective feature selection method. Another method found in HAR applications is minimum redundancy and maximum relevance (MRMR) (Peng et al. 2005). The authors utilized the minimum and maximum mutual information between features and classes to develop the minimum Redundancy and Maximum Relevance criterion for incremental feature selection.

In the current paper we suggest a novel feature selection method, the Comparison of Correlations Based Method (CCBM) that is based on the differences found between pairwise feature correlations of different groups of data. More specifically, correlation matrices of all features are computed separately for each category of the target variable and the correlation coefficients of the different categories are then compared and tested for significant pairwise differences. The features are afterwards ranked according to their presence in the significantly different correlation coefficients and a number of them are selected to perform the classification task. The method is applied in binary and multiclass classification. For the multiclass case, the pairwise comparisons of correlations were conducted for all possible pairs of activities and the p values of the comparisons of the same pairs of variables were adjusted by the number of the combinations of the activity subsets, using the Bonferroni correction. The comparison of correlations has been recently studied in Tsanousa et al. (2019) to identify differences in correlation structures of gene expressions in different groups of patients. The significantly different correlations were further analyzed with the use of Structural Equation Modeling. In this paper, the comparison of correlations is used to identify the optimal features that will be included in a classification model in HAR applications. Compared to other feature selection methods that utilize correlations, like the aforementioned CFS (Hall and Smith 1999) and FCBF (Yu and Liu 2003), the difference with the proposed method (CCBM) lies in the study of the correlations. CFS and FCBF focus on the correlation of features with the target variable and on the correlations between features. In the currently proposed CCBM, the target variable is used in order to divide the data into subsets and then study the correlation structure of these subsets.

To evaluate the proposed method, we tested its performance on two well-known datasets, the Human Activity Recognition (HAR) dataset (Anguita et al. 2013) and the Heterogeneity Human Activity Recognition (HHAR) dataset (Stisen et al. 2015). Both datasets contain sensor readings of subjects performing daily activities. Support Vector Machines (SVM) algorithm was used for the classification and the results were evaluated using the classification accuracy metric. The performance of the proposed method is compared with the performance of the whole feature set and of the CFS algorithm. Overall the contribution of this paper can be summarized in the following:

- The introduction of a novel feature selection method.
- Evaluation of the proposed method on HAR problems.
- The comparison of the performance of this method with another feature selection method that utilizes correlations.

The rest of the paper is organized as follows: an overview of related work is presented in Sect. 2. Theory and methodology are found in Sect. 3, while Sect. 4 presents the experimental results. Finally, the paper is concluded with Sect. 5.

2 Related work

Feature selection methods are employed in activity recognition studies to refine the initial dataset so as to overcome several issues. Analysis of reduced feature sets requires less computational time and resources. The devices used to record human activities create some limitations since they provide raw data and numerous features that cannot produce efficient classification results (Dobbins and Rawassizadeh, 2018, Maurer et al. (2006)). With the use of feature selection methods, the number of variables is decreased and the most useful ones are retained. Besides computational time and energy demands, large sets of features may result in overfitting (Chowdhury et al. 2017).

In machine learning applications of activity recognition problems, features are usually extracted from the raw data, since the analysis of the latter could lead to misleading results. In Luštrek and Kaluža (2009) machine learning algorithms are applied on six sets of extracted features. The authors categorized the extracted features as reference, body and angle attributes and examined their individual performance as well as their combinations, concluding that the reference attribute set performs best in most cases. Stisen et al. (2015), that along with their study provide a dataset used in the current paper, computed most of the typical features found in HAR applications, which are categorized as time domain features, frequency domain and ECDF (extracted from empirical cumulative distribution functions). The feature categories were used separately and in combination in the classification process. Increasing numbers of extracted features does not necessarily result in better classification results. Papers with a few extracted features are regularly found in literature, like Ravi et al. (2005) that extracted two time domain features (mean and standard deviation) and two frequency domain features (energy and correlation) from triaxial sensor readings and achieved very good classification results. A limited set of features like mean, entropy, energy and correlation is also found in Bao and Intille (2004), where the authors mainly examined the effect of the training set's composition to the classification results.

In activity recognition applications, where numerous features are extracted from the -mostly triaxial-sensors, a feature selection algorithm is needed to identify the optimal set of variables that improves the classification performance. In Dobbins and Rawassizadeh (2018) principal components analysis and correlation based feature selection (CFS) were implemented on the HHAR dataset (Stisen et al. 2015), in order to reduce the large number of time and frequency domain features the authors had extracted. The reduced feature sets were used to cluster the data, which are also utilized here, and in general the authors conclude that the CFS method improved the clustering of the activities more than the total feature set or the PCA. The CFS algorithm was also applied in (Chowdhury et al. 2017) to reduce the set of initially extracted features intended to be used in multiclass classifiers of two well-known datasets for human activity recognition. The authors do not compare the results obtained by the method with a baseline or the whole feature set, since the purpose of the paper was to propose a weighting fusion method. Maurer et al. (2006) utilize CFS to select optimal feature sets extracted from sensors of different body positions and compares their performance in terms of classification accuracy. CFS algorithm was also utilized in Capela et al. (2015) along with other two filter methods, Relief-F and FCBF. The three methods were applied on accelerometer and gyroscope data and their performance was evaluated using three classifiers-Naïve Bayes, SVM and decision trees. The authors do not focus on the classification performance of the three methods, but extensively discuss the common or different features of them. FCBF was also used in Wang et al. (2018), integrated in a classification framework, in order to remove redundant features. The classification framework is a hierarchical procedure where the feature selection method was applied on the training set of non-leaf nodes.

Jatoba et al. (2008) review pattern recognition methods, applied on accelerometer sensor systems that are employed in medical applications. After extracting several time and frequency domain features, the authors apply MRMR to eliminate the feature set. Fish et al. (2012), use MRMR algorithm to develop another feature selection method that is applied on data from a body sensor network that consists of 14 triaxial accelerometers. Their method is used to eliminate a set of 434 extracted features, and it is applied on each node of a tree-based classifier. The classification accuracy of their system is compared to that of random forests algorithm. MRMR is also found in Suto et al. (2016), along with FCBF, CFS and other feature selection methods. The paper provides an overview on feature extraction and feature selection methods used in HAR. The feature selection methods are applied separately on the sensor recordings of each subject and their performance is evaluated using neural networks and decision trees. The authors conclude that the results are quite subject-dependent, but overall CFS seems to achieve higher recognition rates in most cases.

Several feature selection methods have been proposed during activity recognition studies. Uddin and Uddiny (2015) suggest a guided random forest to select features during an activity recognition chain. The authors train a random forest to obtain importance scores for the variables and then use these scores to build a guided random forest. Finally they evaluate the constructed forest with the use of a random forest classifier. A feature selection method based on Choquet integral is proposed in Jarraya et al. (2017). The Choquet integral was chosen because of its non-linear and non-additive nature. This methodology takes into consideration the interactions between activities, a property not found in regular feature selection methods. The feature selection algorithm proposed in Peng et al. (2011), uses a non-linear full regression model to examine the relations of the whole feature set and the class variable. Utilizing the sum of squared errors the authors measured the informativeness of features. A two-stage subset selection algorithm founds the optimal subset of features by minimizing locally the sum of squared errors.

The current paper proposes a novel feature selection method for HAR problems and compares its performance with the whole feature set and with CFS that is also utilizing correlations and has been extensively used in HAR literature. Unlike Suto et al. (2016), the subjects are not studied separately. The proposed methodology produces comparable results to CFS. It also allows the researcher to select the number of features to be included in the classifier, like MRMR method. Furthermore, the proposed methodology introduces a different perspective in feature selection methods and explores the variability that may exist in the relationships of variables across different classes. Besides a discussion on the selected feature sets that is also found in other works, we discuss on the difference of feature correlation structures between similar and quite different activities.

3 Methodology

3.1 Tests for comparison of correlations

The correlation coefficients of different samples can be compared either as a whole, i.e. the correlation matrices, or individually, i.e. the coefficient c_{ij} of sample 1 is compared to c_{ij} of sample 2. In any case, the available methods for the comparison are based on the relation of the samples, which can be dependent and overlapping, dependent and non-overlapping or independent.

There are three formulas for the calculation of a correlation coefficient and the choice depends on the type of data. Pearson's correlation coefficient has the assumption of normality of the data and is the most common one (Pearson 1895). Spearman's rho (1904) and Kendall's Tau (1948) are non-parametric and suitable for data not normally distributed. The two latter coefficients use ranks to calculate the relation between two variables. The Pearson's correlation for two variables X and Y, with values x_i, y_i and respective averages $\bar{x}, \bar{y}$, is calculated by Eq. (1). Pearson's correlation is utilized in the current work and is also suitable for the tests for comparison of correlations described below. In the current application, r is calculated for all pairs of features of a dataset.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{1}$$

For the comparison of individual correlation coefficients of two independent samples, there are two available tests: (a) Fisher's z procedure (Fisher 1992) and (b) Zou's confidence intervals (Zou 2007). Fisher's z procedure uses the transformation of a correlation coefficient in order to perform the following tests for detection of significant difference: (a) the comparison of an observed correlation and a given theoretical value, (b) the comparison of two observed correlations. The method is similar to that used to test the significance of a correlation coefficient. The transformation z is given by the following formula:

$$z = \frac{1}{2}[\ln(1 + r) - \ln(1 - r)]. \tag{2}$$

The transformation z follows approximately the normal distribution. Its value tends to be equal to r when the latter is small but when it is close to 1, z tends to increase extremely.

The standard error of z is also utilized in Fisher's z procedure as a threshold to which the z values are compared. When the difference of z values does not exceed twice the standard error, there is no evidence of significantly different correlations. Fisher's z follows normal distribution approximately, even when the data are not normally distributed. The test statistic Z for the difference is given by the following formula, provided by Revelle and Revelle (2015):

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}. \tag{3}$$

The other available method for comparison of correlation coefficients is proposed in Zou (2007). Zou argues that the use of confidence intervals should be preferred, since they provide the analyst with the ability to estimate the amount of the effect as well as the existence of it. Older papers like Olkin and Finn (1995) support this argument, suggesting that confidence intervals are a more suitable method for comparison of correlations. Zou has proposed approaches for independent samples, overlapping and non-overlapping correlations and for comparison of independent multiple squared correlations.

Zou's confidence intervals are based on simple asymptotic (SA) confidence intervals, which were quite popular, but with some restrictions due to the requirement to fulfill some conditions, like (a) changes of the underlying parameter do not affect the sampling distribution of the estimated parameter and (b) the sampling distribution is approximately standard normal. When these conditions do not hold, results may be quite inaccurate. Modified asymptotic methods (MA) that create confidence intervals easier to compute than the bootstrap confidence intervals, were proposed by Zou. MA confidence intervals provide more accurate results than SA ones in some cases. Simulation studies were used to evaluate their performance (Zou 2007).

For two independent samples, which is the case of interest in the current study, the procedure is as follows:

(a) for each correlation coefficient r first create a confidence interval, using the Fisher's r to z transformation, by Eq. 4, where $l', u' = z \mp 1.96\sqrt{\frac{1}{n-3}}$

$$l = \frac{exp\left(2l'\right) - 1}{\exp(2l') + 1}, u = \frac{exp\left(2u'\right) - 1}{\exp(2u') + 1}. \tag{4}$$

(b) Then calculate the lower and upper bounds of the final confidence interval for the difference of correlation coefficients by (Eqs. 5 and 6).

$$L = r_1 - r_2 - \sqrt{(r_1 - l_1)^2 + (u_2 - r_2)^2}, \tag{5}$$

$$U = r_1 - r_2 - \sqrt{(u_1 - r_1)^2 + (r_2 - l_2)^2}, \tag{6}$$

where (r_1, r_2) are the correlation coefficients, (l_1, u_1) is the confidence interval for r_1 and(l_2, u_2) for r_2. The existence of zero in the final confidence interval, suggests that the null hypothesis should not be rejected, thus the two correlation coefficients are not significantly different. The comparison of correlations was performed using the R package 'cocor' (Diedenhofen and Musch 2015).

3.2 Proposed framework

This section describes the steps of the proposed framework. The novel feature selection method is employed to recognize daily activities from sensor recordings. Features in HAR applications are extracted from the raw signals of sensor data using a time window. Although the extracted features in an activity recognition problem are related since they derive from the same raw variables and thus are expected to be correlated in a certain degree, correlation based feature selection methods have been widely used in the field. The current method, however, does not concentrate just on the correlation of variables, but on how that correlation changes over different subsets of data, currently denoting different activities.

As already mentioned, the proposed method is based on the existence of diversity in correlation structures of different subsets. The subsets of interest are created by the different classes of the categorical target variable of a dataset, therefore in HAR applications each subset responds to an activity. The method can be directly applied to binary classification problems and it has been adapted for multiclass classification as well. The steps of the proposed feature selection method for binary classification are described below:

1 Following the extraction of features from raw sensor signals, the dataset is segmented into subsets so that each subset responds to a level of the categorical target variable, which is the activity label. Each one of these subsets is then split to train and test sets to train and evaluate the classification algorithms.

2 The proposed feature selection method, namely CCBM, is applied on the train sets of each activity subset. First, for each activity related subset, the pairwise correlations between features are computed. Then, by using Fisher's test and Zou's confidence intervals, the correlations of the same features of two activity subsets are compared and tested for statistically significant difference. When performing multiple simultaneous comparisons, the alpha values of individual comparisons are not considered appropriate and need to be adjusted in order to take into consideration all the tests. From the few suitable methods available for this procedure, we selected Bonferroni correction, which is simple and widely used. The Bonferroni correction adjusts the alpha level of each individual comparison by dividing with the $n$number of comparisons (Dunn 1958), thus the number of variable pairs. After the adjustment of p values with the Bonfer-

roni correction, the significantly different correlations between two different activities are identified.
3 From each comparison, the variables that are found in many significantly different correlations of the two activity subsets are considered to be informative for the recognition of these activities. Those variables are ranked according to their participation in significantly different correlations. The variable that participates in the most significantly different correlations is in the highest position and the others follow in a descending order. The desired number of the ranked features is selected according to the researcher's criteria and used in the classification process.

The procedure is graphically depicted in Fig. 1.

In order to adapt the proposed feature selection method to multiclass problems, we utilized the Bonferroni correction of p values for multiple comparisons, however in this case, the comparisons refer to all of the combinations of activity subsets. First, subsets that correspond to each activity were created and we calculated the pairwise correlation coefficients of all features for each subset. Afterwards, we conducted the comparisons of correlations between all pairs of activity related subsets. Then we created vectors of the p values that correspond to the comparisons of the same pair of variables from each comparison of activity related subsets' correlations. Those p values were then adjusted, using the Bonferroni correction (Dunn 1958), by the number of all combinations of subsets. Since there is not only one decision on whether the correlation of a pair of variables is significantly different, a threshold that exceeds the median number of comparisons with significant difference was selected in order to characterize a correlation coefficient as such. The pairs of variables that show different correlation structure are then selected and ranked according to their number of appearance in different correlations, in the same way it was used in the binary case.

4 Application and Results

The proposed method is applied on two publicly available datasets that include sensor measurements of subjects performing daily activities. The human activity recognition (HAR) dataset (Anguita et al. 2013), that contains recordings of six daily activities and has already extracted features and the Heterogeneity human activity recognition (HHAR) dataset (Stisen et al. 2015) that includes recordings of six daily activities with accelerometers and gyroscopes embedded in smartphones and smartwatches. In order to demonstrate the application of the proposed framework to binary classification problems, the datasets are segmented by combining pairs of activities.

The performance of the proposed framework is evaluated by comparing its classification performance to the performance of all variables and of the feature set selected by the CFS method, which is also based on correlations and is frequently used in HAR problems. For the evaluation of the performance, we report the accuracy value of the classification algorithm (Eq. 7). SVM were employed to recognize the activities. Each sensor is examined separately. The same stands for devices with different sampling frequencies, existing in the HHAR dataset. The feature selection algorithms were applied on the train sets. The extracted features, whether they already existed or calculated for the application, are only time domain. Frequency domain features were avoided since some of them include correlation and it was not considered appropriate to apply the proposed method on them.

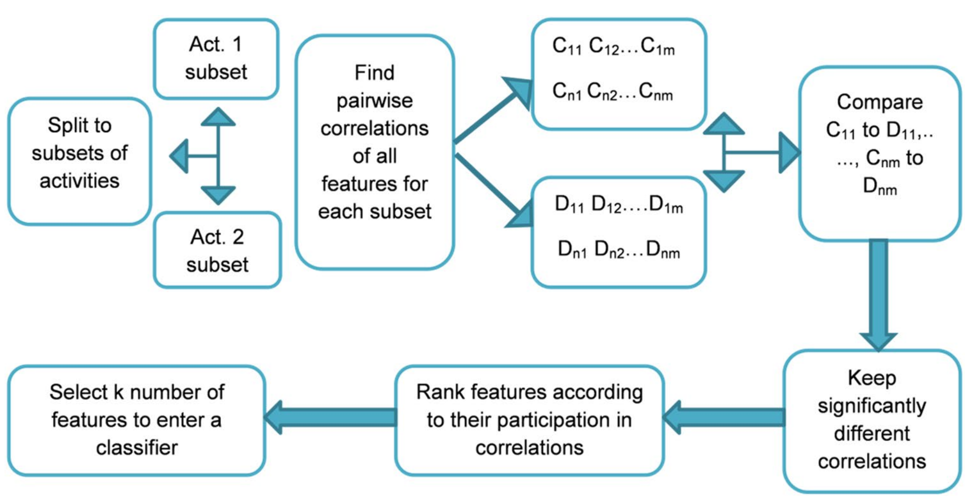

Fig. 1 Flowchart of the analysis' steps. C_{nm} and D_{nm} refer to the feature correlation coefficients of the activities 1 and 2 subsets respectively

$$Accuracy = \frac{True\ Positives + True\ Negatives}{TP + TN + FP + FN}. \tag{7}$$

In the following, we present the experimental results on each dataset. To demonstrate the performance of the proposed method in the binary classification applications, one activity was chosen as the 'control' group and was combined with each one of the other activities to create subsets suitable for the binary classification problem.

4.1 Application of CCBM on the UCI HAR dataset for binary classification

The UCI HAR dataset contains 561 features extracted from accelerometers and gyroscopes. For the current work, we utilized only the time domain features extracted from each of the three sensor axes, namely mean, standard deviation (std), median absolute deviation (mad), maximum and minimum values. Thus, the whole feature set included 30 features, obtained from gyroscope body and jerk signals. For more information, we refer the reader to Anguita et al. (2013). Using the proposed feature selection method, CCBM, different sizes for the feature set were examined, including maximum 15 features. The 30 participants performed the following six activities: stand, lay, sit, walk downstairs (WD), walk upstairs (WU) and walk. Standing was selected as the control activity with which all others were combined.

Table 1 shows the classification accuracy of the SVM algorithm with various feature sets: a) all features b) feature sets of different sizes selected by the proposed method (CCBM) and c) features selected by the CFS method. Specifically for CCBM, we demonstrate the results of three feature sets, using 15, 5 and 2 features respectively. The classification was performed using 'leave-one-user-out' validation, i.e. one user was kept out for testing and the rest were used for training the classification algorithm. The procedure was repeated until all participants were used as testsets and the final classification accuracy reported was the average of all runs. In the binary classification of similar static activities like 'stand vs sit' or 'stand vs lay', CCBM with 15 features and the whole feature set provided better recognition rates. As it can be seen in Table 1 for these two pairs of activities, decreasing the selected features of CCBM decreases the classification accuracy. For the rest of the comparisons (stand vs WD, stand vs WU, stand vs walk) CFS and CCBM with less features achieved the higher recognition rates. For these three pairs that contain activities quite different by nature, one static and one dynamic, less features suggested by CCBM were required to improve the classification accuracy.

Table : .

To elaborate more on the performance of the proposed method, we observe the correlation structures of different activities. Figure 2 depicts the visual presentation of correlation coefficients of the time domain features of gyroscope for the activity 'sit' (upper triangle) and for the activity 'stand' (lower triangle). Darker and bigger circles depict stronger correlations while empty cells refer to insignificant correlation coefficients that are not depicted in the plot. Both activities show similar correlation patterns that could be the cause of the CCBM's decrease in performance when selecting fewer variables. On the contrary, Fig. 3, which includes the correlation plots of 'stand vs walk', reveals that although there are similar patterns in both subsets, especially regarding the insignificant correlations, the features in the subset of 'stand' have stronger correlations (darker and bigger circles) opposed to the subset 'walk', which can be an indication of the high performance achieved even when selecting two variables with CCBM.

Table 2 includes the top 5 selected features of CCBM and the selected features of CFS, for every pair of activities. For 'stand vs lay', four of the most important features according to CCBM are also selected by CFS. For the three last subsets, CFS selects only one feature, which is also selected by CCBM for the subsets 'stand vs WD' and 'stand vs Walk'. Comparing the selected features of CCBM for the various subsets, we don't observe related patterns, a finding that supports the importance of differences in correlation structures. Two of the five features are common in 'stand vs WD' and 'stand vs WU', which could be expected since walking upstairs or downstairs are similar activities.

Table 1 Classification accuracy using the gyroscope recordings

	Gyroscope				
	Stand vs lay	Stand vs sit	Stand vs WD	Stand vs WU	Stand vs walk
All	*0.7612*	0.8107	0.9322	0.9413	0.9565
CCBM (#15)	0.7423	*0.8135*	0.9508	0.9561	0.9651
CCBM (#5)	0.6610	0.7700	0.9750	*0.9776*	0.9867
CCBM (#2)	0.6000	0.6024	0.9896	0.9702	0.9888
CFS	0.7049	0.7650	*0.9949*	*0.9789*	*0.9939*

The number of features selected by CCBM is included in the parentheses

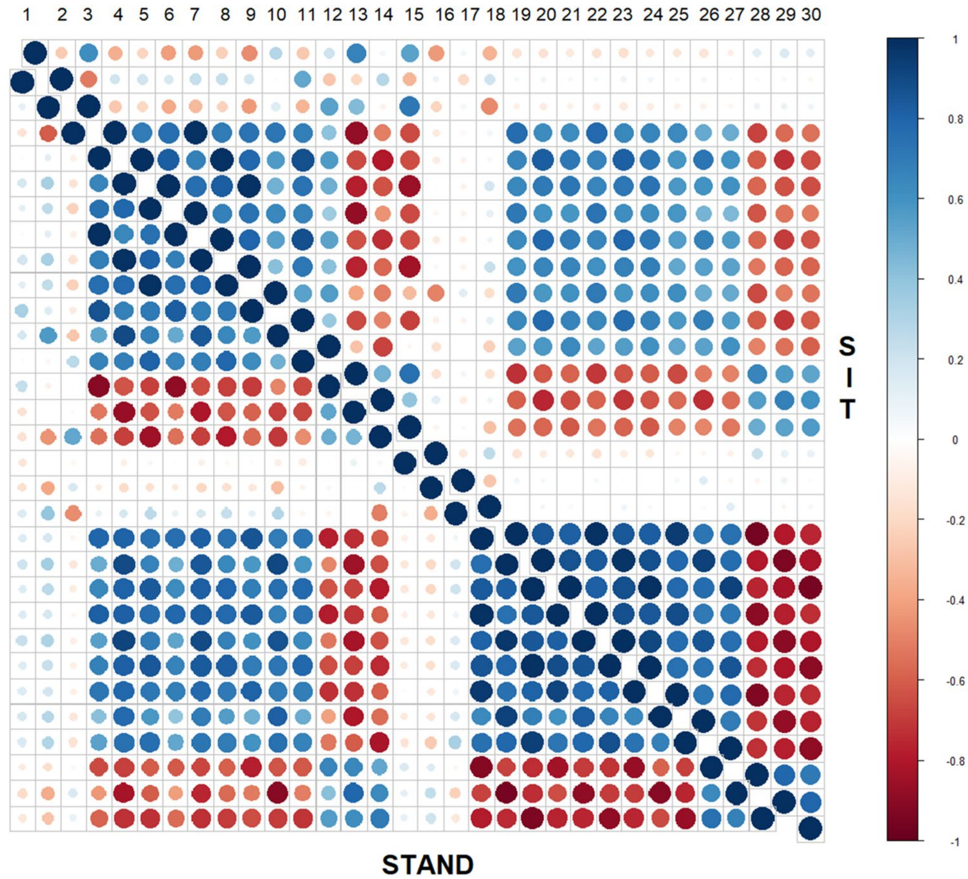

Fig. 2 Correlation plot of the gyroscope features in 'sit' subset (upper triangle) and 'stand' subset (lower triangle). The names of features can be found in "Appendix".

4.2 Application of CCBM on the HHAR dataset for binary classification

This dataset was chosen due to its simplicity in the format of the raw files and its plurality in devices. The HHAR dataset includes recordings from gyroscopes and accelerometers embedded in smartphones and smartwatches. Nine users participated in the experiment, carrying the smartphones attached on their waist and the smartwatches on their wrists. Furthermore, there were multiple devices of different manufacturers and with different sampling frequencies, thus giving us the advantage of creating many subsets to run our experiments. The dataset was initially intended for analyzing heterogeneity, which is not of interest in the current study (Stisen et al. 2015). Therefore, to eliminate diversity, each device was considered separately for the application. The sampling frequencies of the devices were the following: the LG Nexus 4 device had sampling frequency of 200 Hz, Samsung S3 150 Hz, S3 mini 100 Hz and Samsung Galaxy S+ (referred to as Samsung old in the dataset) had 50 Hz sampling rate.

For each device related subset included in this paper, the following time domain features were extracted from a sliding time window: mean, median, max, min, standard deviation and variance. The time windows were adjusted to the sampling frequency of each device in order to cover a period of two seconds with one second overlap. The datasets were again split into train and test sets respective to the users and leaving one user out for the test set on each run. The results of all runs were then averaged and reported in the tables below. SVM algorithm was again used for the classification.

The whole feature set consisted of 18 extracted variables and for the CCBM feature set we selected as minimum the top five of the ranked variables, which is almost one third of the initial dataset. That number of features was selected because it optimized the results in most of the runs of an experiment. The CFS algorithm selects the optimal number of variables automatically. Since we consider all users as test sets in different runs, the optimal number of variables was not stable across all runs for the CFS method.

4.2.1 Accelerometer

The results of the accelerometer recordings are reported separately for each phone model, due to the different sampling frequencies of the models. First we report the results for the Nexus 4 device with sampling frequency of 200 Hz. Table 3 shows the results of the binary classifications obtained from the pairwise combinations of "stand" with the rest of the activities. There are three feature sets: (a) all variables, (b)

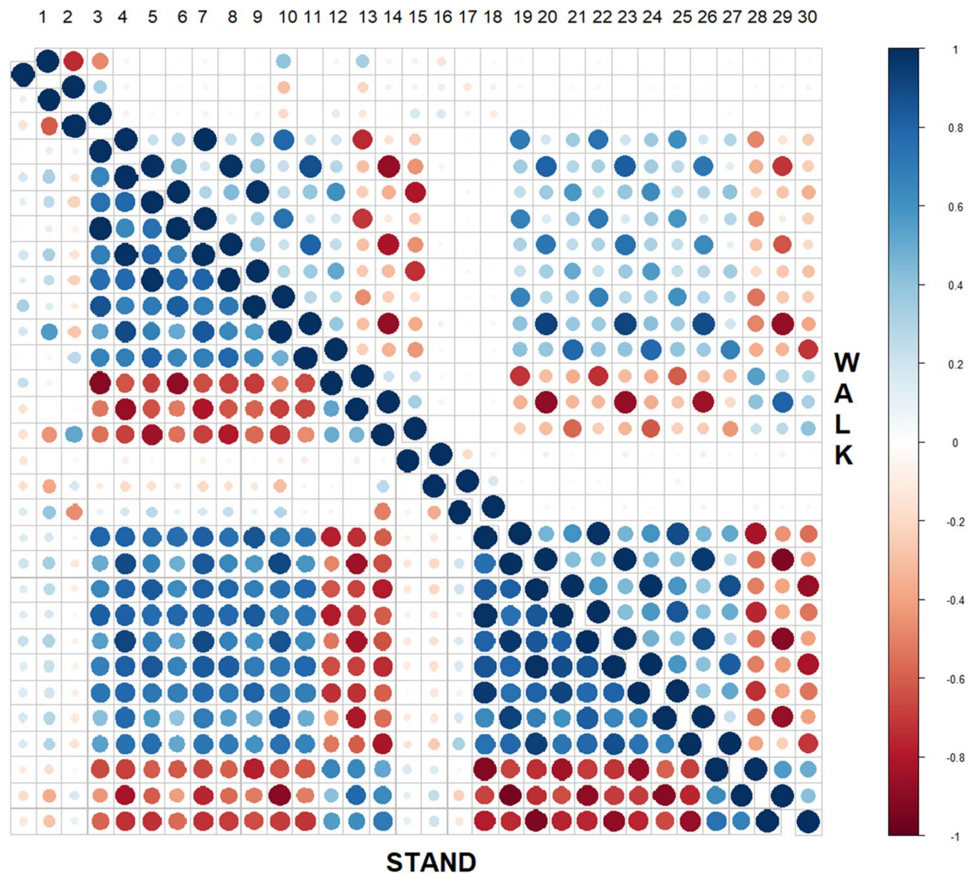

Fig. 3 Correlation plot of the gyroscope features in 'walk' subset (upper triangle) and 'stand' subset (lower triangle). The names of features can be found in "Appendix".

Table 2 Features selected from CCBM and CFS for each activities' subset

	Top #5 CCBM	CFS
Stand vs lay	Mean of body gyro. jerk-z, mean of body gyro.-y, mean of body gyro.-z, min of body gyro.-z, mean of body gyro. jerk-x	Mean of body gyro.-y, mean of body gyro.-z, mad of body gyro.-z, max of body gyro.-z, min of body gyro.-x, min of body gyro.-z, mean of body gyro. jerk-z, std of body gyro. jerk-x
Stand vs sit	Mean of body gyro.-x, max of body gyro.- z, max of body gyro.-x, mean of body gyro. jerk-y, std of body gyro. jerk-y	Mean of body gyro.-z, mad of body gyro.-x, max of body gyro.-x, min of body gyro.-x, std of body gyro. jerk-x
Stand vs WD	Mean of body gyro.-y, mad of body gyro.-x, max of body gyro.-z, min of body gyro.-z, mad of body gyro.-z	Mad of body gyro.-z
Stand vs WU	Mean of body gyro.-y, mad of body gyro.-x, max of body gyro.-x, min of body gyro.-x, std of body gyro.-x	Mad of body gyro. jerk-z
Stand vs walk	Mad of body gyro.-z, mad of body gyro. Jerk-y, max of body gyro. jerk-z, std of body gyro.-z, std of body gyro. jerk-y	Mad of body gyro. jerk-z

Table 3 Classification accuracy using the accelerometer recordings of the Nexus device

	Accelerometer-Nexus				
	Stand vs bike	Stand vs sit	Stand vs SD	Stand vs SU	Stand vs walk
All	0.8240	0.7806	0.9934	0.9798	0.9829
CCBM	0.8741	*0.8081*	0.9743	*0.9937*	0.9970
CFS	*0.9738*	0.778	*0.9976*	0.9903	*0.9998*

features selected from CCBM and (c) features selected from CFS. For most of the comparisons, the classification accuracy was improved when using the reduced set of CCBM instead of the whole dataset. The differences in the performance of CCBM vs the CFS were not important, except in the case of 'stand vs bike'.

The accelerometer recordings of the S3 device (Table 4) revealed superiority of the CFS method in most of the comparisons and lower performance of the CCBM method compared to the two other feature sets.

Using the accelerometer recordings of the S3 mini device (Table 5), it is observed that the classification accuracy is generally lower than the other devices although the sensor is the same. The CCBM feature set achieved an important increase in the classification accuracy of the whole feature set in the 'stand vs SD' subset. Overall, CFS outperformed the other two feature sets in most cases. This is also observed in Table 6 that contains the Samsung S+ recordings.

To further elaborate on the performance of the proposed method, we graphically observe the correlation structure of some sample datasets. We chose the accelerometer recordings of the S3 device to illustrate the correlation plots of two pairs of activities: (a) 'stand vs sit', where CCBM did not identify the most optimal variables to distinguish between the two activities as it is observed in Table 3 and (b) 'stand vs stairs up', where the reduced feature set of CCBM gave high classification accuracy. The correlation plots refer to the pairwise correlations of features calculated from all users. As it can be seen in Fig. 4 the correlation structure of sit (upper triangle) and stand (lower triangle) is quite similar. The same features have insignificant correlations (empty cells) in both subsets and both have many identical pairs of features strongly correlated. While the correlation plots of the subset of 'stairs up' activity is quite different from the control activity 'stand' (Fig. 5). There are fewer insignificant correlations in 'stairs up' subset and less strong correlations overall. Besides the correlation structures of two activities visualized here, variation has been observed among the results of the different subjects with all three feature sets.

4.2.2 Gyroscope

Using the gyroscope sensor and the Nexus device (Table 7), the performance of the CCBM feature set was again quite satisfying. In most cases, the whole feature set already performed very well and this performance was retained with the reduced feature sets of the two methods that required less running time. For the activities 'stand vs sit' that are the most difficult to classify according to all feature sets and devices that have been used, CCBM achieved the higher performance of the other two feature sets. This may be an indication that devices affect the performance of a sensor, a finding in line with Stisen et al. (2015).

Table 8 includes the accuracy obtained from the gyroscope recordings of the Samsung S3 mini model. The performance of CCBM is in most experiments satisfactory; although we still observe that the separation of the activities

Table 4 Classification accuracy using the accelerometer recordings of the Samsung S3 device

	Accelerometer–S3				
	Stand vs bike	Stand vs sit	Stand vs SD	Stand vs SU	Stand vs walk
All	0.8758	*0.8447*	0.9804	0.9767	0.9838
CCBM	0.8696	0.7498	0.9626	0.9904	0.9567
CFS	*0.9884*	0.7782	*0.9990*	*0.9995*	*0.9997*

Table 5 Classification accuracy using the accelerometer recordings of the Samsung S3 mini device

	Accelerometer–S3mini				
	Stand vs bike	Stand vs sit	Stand vs SD	Stand vs SU	Stand vs walk
All	0.8182	0.6905	0.8709	0.8926	0.9212
CCBM	0.8490	0.6647	*0.9967*	0.8907	0.8590
CFS	*0.9569*	*0.7579*	0.9936	*0.9938*	*0.9994*

Table 6 Classification accuracy using the accelerometer recordings of the Samsung S+ device

	Accelerometer -Samsung S+				
	Stand vs bike	Stand vs sit	Stand vs SD	Stand vs SU	Stand vs walk
All	0.8711	0.7614	0.9625	0.9629	0.9629
CCBM	0.8598	0.6122	0.9112	0.9707	0.9720
CFS	*0.9769*	*0.8441*	*0.9948*	*0.9965*	*0.9942*

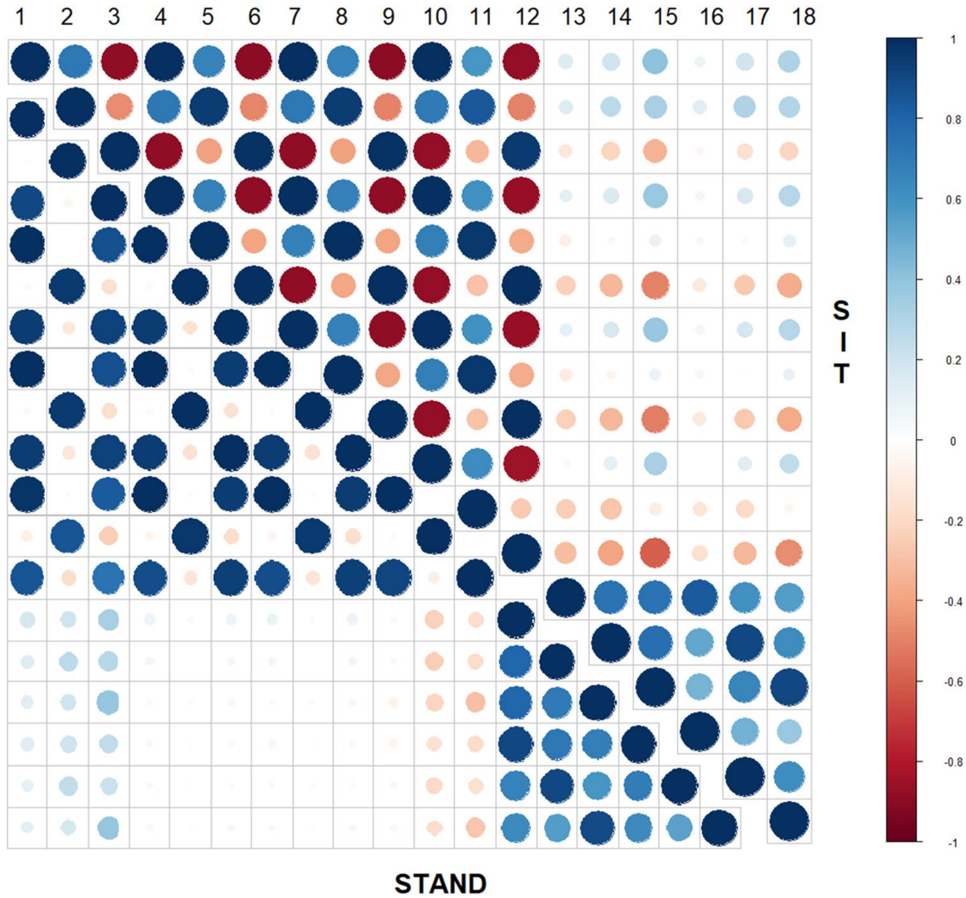

Fig. 4 Correlation plot of the accelerometer features (S3 device) in 'sit' subset (upper triangle) and 'stand' subset (lower triangle). The names of features can be found in "Appendix".

'stand vs sit' cannot be improved using either feature selection method.

4.3 Application of CCBM on both datasets for multiclass classification

As already described, in order to apply CCBM to multiclass problems, we conducted the comparisons of correlations for all possible combinations of activity related subsets and afterwards adjusted the p values corresponding to the same pairs of variables, by the number of those combinations. Both datasets used in this paper, consist of recordings of six activities, thus there were 15 combinations of activity subsets. A correlation was characterized as different if it showed significant difference in 14 or 15 comparisons. That decision reflects that a correlation coefficient is different in all or almost all of the comparisons. Table 9 shows the results for the implementation of methodology to both datasets. Leave-one-user-out validation was used again on each subset. For all the experiments included in Table 9 five variables were selected with CCBM. In the HAR application, CFS achieved the higher performance, although all three feature sets produced close accuracy values. In the application of HHAR subsets the best performance for half of the subsets, was achieved using the whole feature set. CCBM achieved higher or close accuracy values to CFS in the first three experiments. It is noteworthy though, that CCBM achieved close recognition accuracy to CFS, using almost half of the features, since for CCBM we selected the top five ranked features, while CFS selected on average around 10 features for each subset.

The duration of the experiments of the multiclass classification is included in Table 10 in order to compare the two feature selection methods, CCBM and CFS, in terms of time. Varying on the size and composition of the dataset, CFS required 3 min on average to select the optimal feature set while CCBM required 2 min to produce the ranked features. Since the nature of the two methods differs in the amount of variables entered in the classifier, we report the duration of the whole process. The cells include the average time needed for an experiment, starting from the feature selection method and until the classification results are produced. The results refer to experiments with nearly identical number of variables.

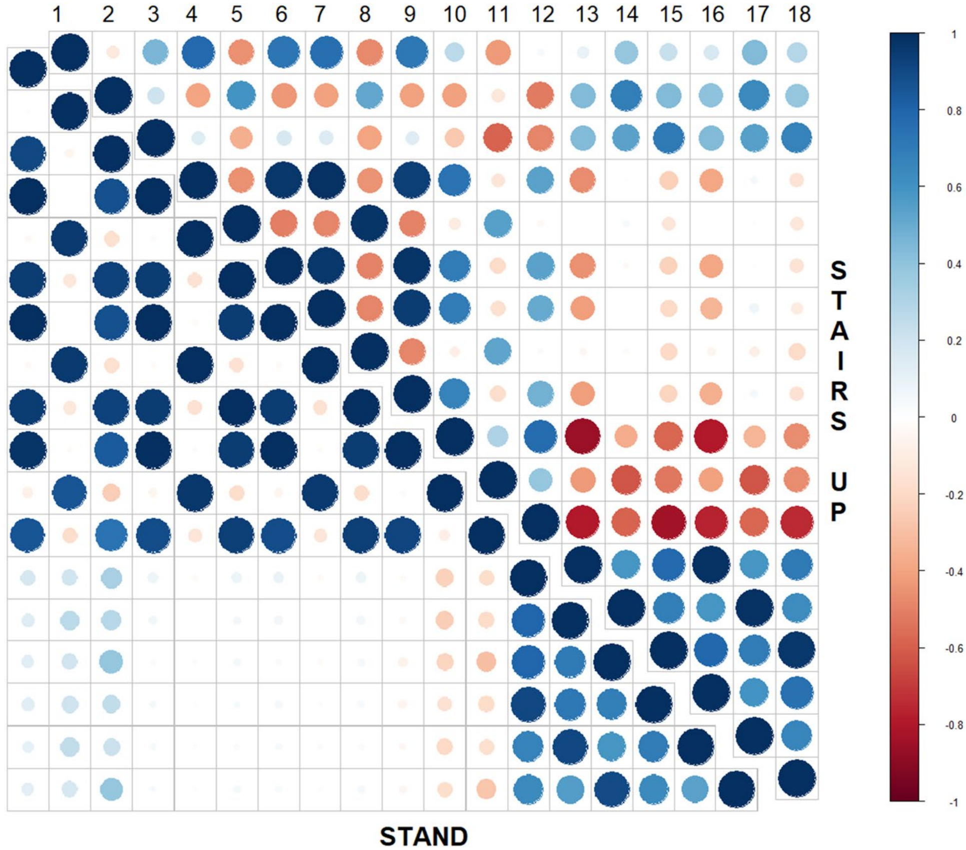

Fig. 5 Correlation plot of the accelerometer features (S3 device) in 'stairs up' subset (upper triangle) and 'stand' subset (lower triangle). The names of features can be found in "Appendix".

Table 7 Classification accuracy using the gyroscope recordings of the Nexus device

	Gyroscope-Nexus				
	Stand vs bike	Stand vs sit	Stand vs SD	Stand vs SU	Stand vs walk
All	0.9343	0.7490	0.9492	0.9610	0.9708
CCBM	0.9360	*0.7770*	0.9680	0.9626	0.9839
CFS	*0.9501*	0.7400	*0.9839*	*0.9873*	*0.9979*

Table 8 Classification accuracy using the gyroscope recordings of the Samsung S3 mini device

	Gyroscope–S3 mini				
	Stand vs bike	Stand vs sit	Stand vs SD	Stand vs SU	Stand vs walk
All	0.9100	*0.7223*	0.9496	0.9510	0.9651
CCBM	*0.9326*	0.7031	*0.9598*	0.9620	*0.9925*
CFS	0.9259	0.7173	0.9339	*0.9871*	0.9816

5 Conclusion

Feature selection is an important step in the activity recognition process where there are usually a large number of features involved. In the current paper we introduce a feature selection algorithm that is based on the comparison of correlation coefficients of features between different subsets of data, namely the CCBM. The proposed method is based on the observation that correlations of variables, and not only individual variables, may differ across different groups of data. The feature selection method is applied on subsets of two publicly available datasets for human activity recognition and its performance is tested both on binary and multiclass classification problems.

Correlation based feature selection methods, like CFS and FCBF, are regularly found in HAR literature and are reported to achieve high classification performance, although they may require more computational cost. Here, CFS was utilized and compared with our method, CCBM, since they

Table 9 Classification accuracy of the multiclass case. Number of features is included in the parentheses

	Feature sets		
	All	CCBM	CFS
HAR-gyroscope features	0.5159 (#30)	0.5077 (#5)	*0.5216*
HHAR-accel. Nexus	*0.6921 (#18)*	0.6526 (#5)	0.5859
HHAR-accel. S3	*0.7187 (#18)*	0.6374 (#5)	0.6374
HHAR-accel. S3 mini	*0.6771 (#18)*	0.5821 (#5)	0.6067
HHAR-accel. Samsung S+	0.7136 *(#18)*	0.5904 (#5)	*0.7284*
HHAR-gyro. Nexus	0.6289 *(#18)*	*0.6128* (#5)	*0.6333*
HHAR-gyro. S3 mini	0.5962 *(#18)*	0.5378 (#5)	*0.6001*

Table 10 Average duration of multiclass experiments starting from the feature selection method until the production of classification results.

	Duration of experiments	
	CCBM (min)	CFS (min)
HAR-gyroscope features	50	46
HHAR-accel. Nexus	40	40
HHAR-accel. S3	15	18
HHAR-accel. S3 mini	10	10.7
HHAR-accel. Samsung S+	111.6	82.25
HHAR-gyro. Nexus	112	123
HHAR-gyro. S3 mini	38	36

both use correlation structures to select an optimal set of features. The feature selection sets selected by both methods were compared to the whole dataset in terms of classification performance. The proposed method, CCBM, provided quite promising results, outperforming the other feature sets in many experiments or achieving similar performance. For the binary classification of similar activities, more features proposed by CCBM where required to improve the recognition rate, while the opposite was observed for the binary classification of diverse activities. For both datasets utilized, remarkable deviations in classification accuracy of different subjects was observed. Differences among subject were also observed in the feature sets provided by CCBM.

Overall, we can conclude that the proposed method seems very promising. Although such methods are quite data driven and their performance may vary on different datasets, the cases CCBM did not improve the recognition rate need to be further studied. More specifically, the effect of the overall correlation structure of two subsets on the feature selection needs to be taken into consideration when ranking the features. The focus of the current work was to explore to what extend the variables with different correlations between different groups, can distinguish between classes and how many of these variables are usually needed. For future work the two main issues that are currently under study are the selection of a more advanced method for ranking the variables and the modification of tests that compare correlations so that they can conduct multiple comparisons. Furthermore, it would be interesting to compare the performance of CCBM among different subjects and among devices with different sampling frequencies. The proposed method is not limited to human activity recognition problems and can be employed in any classification task.

6 Appendix

List of features of Figs. 2, 3	
1	tBodyGyro_mean_X
2	tBodyGyro_mean_Y
3	tBodyGyro_mean_Z
4	tBodyGyro_std_X
5	tBodyGyro_std_Y
6	tBodyGyro_std_Z
7	tBodyGyro_mad_X
8	tBodyGyro_mad_Y
9	tBodyGyro_mad_Z
10	tBodyGyro_max_X
11	tBodyGyro_max_Y
12	tBodyGyro_max_Z
13	tBodyGyro_min_X
14	tBodyGyro_min_Y
15	tBodyGyro_min_Z
16	tBodyGyroJerk_mean_X
17	tBodyGyroJerk_mean_Y
18	tBodyGyroJerk_mean_Z
19	tBodyGyroJerk_std_X
20	tBodyGyroJerk_std_Y
21	tBodyGyroJerk_std_Z
22	tBodyGyroJerk_mad_X
23	tBodyGyroJerk_mad_Y
24	tBodyGyroJerk_mad_Z
25	tBodyGyroJerk_max_X
26	tBodyGyroJerk_max_Y
27	tBodyGyroJerk_max_Z
28	tBodyGyroJerk_min_X
29	tBodyGyroJerk_min_Y
30	tBodyGyroJerk_min_Z

List of features of Figs. 4, 5	
1	max_x

List of features of Figs. 4, 5	
2	max_y
3	max_z
4	mean_x
5	mean_y
6	mean_z
7	median_x
8	median_y
9	median_z
10	min_x
11	min_y
12	min_z
13	sd_x
14	sd_y
15	sd_z
16	var_x
17	var_y
18	var_z

Acknowledgements This research has been cofinanced by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code:T1EDK-00686)

References

Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL (2013) A public domain dataset for human activity recognition using smartphones. In: Esann

Capela NA, Lemaire ED, Baddour N (2015) Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients. PLoS One 10(4):e0124414

Chen L, Hoey J, Nugent CD, Cook DJ, Yu Z (2012) Sensor-based activity recognition. IEEE Trans Syst Man Cybern Part C (Appl Rev) 42(6):790–808

Chowdhury AK, Tjondronegoro D, Chandran V, Trost SG (2017) Physical activity recognition using posterior-adapted class-based fusion of multiaccelerometer data. IEEE J Biomed Health Inf 22(3):678–685

Diedenhofen B, Musch J (2015) Cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 10(4):e0121945

Dobbins C, Rawassizadeh R (2018) Towards clustering of mobile and smartwatch accelerometer data for physical activity recognition. In: Informatics, vol 5, no 2. Multidisciplinary Digital Publishing Institute

Dunn OJ (1958) Estimation of the means of dependent variables. Ann Math Stat:1095–1111

Fish B, Khan A, Chehade NH, Chien C, Pottie G (2012) Feature selection based on mutual information for human activity recognition. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 1729–1732

Fisher RA (1992) Statistical methods for research workers. In: Breakthroughs in statistics. Springer, New York, NY, pp 66–70

Hall MA, Smith LA (1999) Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In: FLAIRS conference, vol 1999, pp 235–239

Jarraya A, Arour K, Bouzeghoub A, Borgi A (2017) Feature selection based on Choquet integral for human activity recognition. In: 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, pp 1–6

Jatoba LC, Grossmann U, Kunze C, Ottenbacher J, Stork W (2008) Context-aware mobile health monitoring: Evaluation of different pattern recognition methods for classification of physical activity. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp 5250–5253

Kendall MG (1948) Rank correlation methods

Lara OD, Labrador MA (2012) A survey on human activity recognition using wearable sensors. IEEE Commun Surv Tutor 15(3):1192–1209

Luštrek M, Kaluža B (2009) Fall detection and activity recognition with machine learning. Informatica 33(2)

Mangai UG, Samanta S, Das S, Chowdhury PR (2010) A survey of decision fusion and feature fusion strategies for pattern classification. IETE Tech Rev 27(4):293–307

Maurer U, Smailagic A, Siewiorek DP, Deisher M (2006) Activity recognition and monitoring using multiple sensors on different body positions. In: International workshop on wearable and implantable body sensor networks (BSN'06). IEEE, 4 p

Olkin I, Finn JD (1995) Correlations redux. Psychol Bull 118(1):155

Pearson K (1895) Note on regression and inheritance in the case of two parents. Proc R Soc Lond 58:240–242

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

Peng JX, Ferguson S, Rafferty K, Kelly PD (2011) An efficient feature selection method for mobile devices with application to activity recognition. Neurocomputing 74(17):3543–3552

Ravi N, Dandekar N, Mysore P, Littman ML (2005) Activity recognition from accelerometer data. In: AAAI, vol 5, no 2005, pp 1541–1546

Revelle W, Revelle MW (2015) Package 'psych'. The comprehensive R archive network

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101

Stisen A, Blunck H, Bhattacharya S, Prentow TS, Kjærgaard MB, Dey A et al (2015) Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In: Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems. ACM, New York, pp 127–140

Suto J, Oniga S, Sitar PP (2016) Comparison of wrapper and filter feature selection algorithms on human activity recognition. In: 2016 6th international conference on computers communications and control (ICCCC). IEEE, pp 124–129

Tsanousa A, Ntoufa S, Papakonstantinou N, Stamatopoulos K, Angelis L (2019) Study of gene expressions' correlation structures in subgroups of Chronic Lymphocytic Leukemia Patients. J Biomed Inform 95:103211

Uddin MT, Uddiny MA (2015) A guided random forest based feature selection approach for activity recognition. In: 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE, pp 1–6

Ustev YE, Incel DO, Ersoy C (2013) User, device and orientation independent human activity recognition on mobile phones: Challenges and a proposal. In: Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication. ACM, New York, pp 1427–1436

Wang A, Chen G, Wu X, Liu L, An N, Chang CY (2018) Towards human activity recognition: a hierarchical feature selection framework. Sensors 18(11):3629

Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03). Washington, D.C. pp 856–863 **(August 21–24, 2003)**

Zou GY (2007) Toward using confidence intervals to compare correlations. Psychol Methods 12(4):399