

# Analysis and Effect of Feature Selection Over Smartphone-Based Dataset for Human Activity Recognition

Ilham Amezzane<sup>1</sup>(✉), Youssef Fakhri<sup>1</sup>, Mohammed El Aroussi<sup>1</sup>,  
and Mohamed Bakhouya<sup>2</sup>

<sup>1</sup> Faculté des Sciences, Université Ibn Tofail, Kenitra, Morocco  
ilhammaj@gmail.com, fakhri@uit.ac.ma,  
mohamed.elaroussi@ieee.org

<sup>2</sup> International University of Rabat, Sala Aljadida, Morocco  
Mohamed.bakhouya@uir.ac.ma

**Abstract.** The availability of diverse and powerful sensors that are embedded in modern smartphones has created exciting opportunities for developing context-aware services and applications. For example, Human activity recognition (HAR) is an important feature that could be applied to many applications and services, such as those in healthcare and transportation. However, recognizing relevant human activities using smartphones remains a challenging task and requires efficient data mining approaches. In this paper, we present a comparison study for HAR using features selection methods to reduce the training and classification time while maintaining significant performance. In fact, due to the limited resources of Smartphones, reducing the feature set helps reducing computation costs, especially for real-time continuous online applications. We validated our approach on a publicly available dataset to classify six different activities. Results show that Recursive Feature Elimination algorithm works well with Radial Basis Function Support Vector Machine and significantly improves model building time without decreasing recognition performance.

**Keywords:** Human Activity Recognition · Smartphone sensors · Feature selection

## 1 Introduction

Human Activity Recognition (HAR) using Smartphones has been widely studied during recent years mainly because Smartphones are not intrusive and widely used in everyday life. Researchers are developing many new challenging application scenarios based on mobile phone sensors in various fields such as in healthcare (e.g., fitness, diabetes, elderly and obesity assisted surveillance), in smart buildings (e.g., context aware automatic indoor air quality and thermal comfort control) and in smart cities applications (e.g., traffic congestion). Actually, modern Smartphone devices have great capacity for collecting and classifying large amounts of multiple sensor readings. However, for real-time online implementation, data pre-processing and training steps

are required to be achieved under hardware resource constraints, such as memory and battery life. Many solutions have been studied so far in literature to overcome these limitations [1], for example, by reducing the amount of training data needed in the learning phase.

In this paper, we present a comparison study based on feature selection approaches in order to reduce the dimensionality of the training dataset while maintaining high recognition performance. The remainder of this paper is organized as follows. A brief overview of the related work is presented in Sect. 2. Section 3 presents the dataset, the experimental methodology and simulation results. Section 4 presents the conclusions and future work.

## 2 Related Work

Many research efforts have been done to implement HAR process on different Smartphone devices using various data sets. However, for real-time online implementation, data pre-processing and training steps are required to be achieved under hardware resources constraints. Recently, researchers focus mainly on mobile phone's onboard sensors for real-time online applications. For example, authors in [1] have reviewed research studies in this domain and stated that only few of them focused on online training in which classifiers can be trained in real time on mobile phones [2–5]. Authors, in [2], introduced “hardware friendly” adaptation of the classification algorithms in order to overcome the resources constraints. In [6], authors used dynamic and adaptive sensor selection to save battery energy. Other studies used adaptive sampling techniques [7] for the same goal.

Classifiers could play a key role in HAR process regarding energy consumption depending on their simplicity or complexity. Nonetheless, some of them have proven their suitability for Smartphone implementation, such as K-nearest neighbor (KNN), Support Vector Machine (SVM) and Decision Tree (DT) [1]. In fact, in the pre-processing phase, various features (aka. predictors or variables) are extracted from sensors readings. These features are used by the classifier later during training (aka. learning), validation (optionally) and testing phases. Moreover, in online activity recognition, two main types of features are generally used: time or/and frequency domain features. It has been shown in [8] that time domain features are cheaper than frequency domain features in terms of computation and storage costs. In practice, large feature sets may significantly slow down the learning process [9]. In addition, the “dimensionality curse” phenomenon states that the number of training data needed grows exponentially with the number of dimensions used [10]. Subsequently, the online training requires further intensive computation if locally undergone on Smartphones [11]. For this reason, the goal of Feature Selection (FS) methods is to select optimal subsets of variables in the pre-processing step. The main benefits are reducing the computation cost and storage requirements as well as training time [9].

There are three main approaches for feature selection in literature: (i) *Wrapper* methods for measuring the “usefulness” of the features guided by a classifier performance, (ii) *Filter* methods for measuring the “relevance” of the features independently of the classifier, and (iii) *Embedded* methods that are implemented by

algorithms having their own built-in FS methods for performing variable selection implicitly while the model is being trained.

Regarding the performance, an important source of influence on the HAR process is the classifier itself as stated in [11]. For this reason, in our comparison study, we have evaluated different classifiers belonging to different categories before and after feature selection. Our final selection included four classifiers: Linear Discriminative Analysis (LDA), Radial Basis Function Support Vector Machine (RBF SVM), K-nearest neighbors (KNN), and Random Forest (RF). In order to discriminate test data into labeled classes, the classifiers need to be trained first. Thus, we trained our selected classifiers using a 10-fold cross validation technique. The parameters of the best final models were preserved for testing on holdout data.

### 3 Experimental Methodology and Results

In this work, we used a publicly available dataset: “Human Activity Recognition Using Smartphones Data Set” [12], which has been used by the authors to conduct experiments using Support Vector Machine (SVM) classifier [13]. The latest update (15-Feb-2015) includes labeled data collected from 30 subjects who engaged in six different activities (standing, sitting, laying down, walking, walking downstairs and walking upstairs), while wearing Smartphones that embed accelerometer and gyroscope sensors. The list of all the measures applied to the time and frequency domain signals are available in [13]. A total of 561 features were extracted to describe each activity window (2.56 s) in the dataset. Two files for activity labels and subjects ID numbers are also available. The classification results of the original work [13] show an overall accuracy of 96% for the Test data. In our comparison study, we have used the same “Test” data. In addition, we partitioned the original “Training” data (7352 observations) into “Training/Validation” subsets to avoid overfitting during learning phases.

In the comparison study, we tried one wrapper algorithm called Recursive Feature Elimination (RFE) and one embedded algorithm specific to Random Forest (RF) classifier called Variable Importance (varImp). After running calculations, 20 variables have been selected with the latter while 50 have been selected with the former. After looking at the names of those features, we first noticed that the “50 features subset” include almost all the variables of the “20 features subset” except one. Moreover, we noticed that only 5 variables in the latter are of frequency domain, against 10 variables in the former. In both cases, this reduction in the number of frequency domain features is beneficial in terms of computation cost, because the original feature set contains many frequency domain features based on Fast Fourier Transform (FFT) [13] which demands extra computation [8]. Finally, we constructed two new data sets with the features selected from both methods before applying different classifiers, and conducted a comparison study in order to select the feature subset that works well regarding the recognition performance and the training time.

Because balanced class proportions assumption is verified in our data sets, good performance of each classifier can be measured by accuracy metric only, and good performance of each activity is obtained if that activity can be classified with high

precision (PR), recall (RC), and F-measure (F1) metrics. As shown in Table 1, we examined 20, 50 and 561 features datasets for the comparative performance for each classifier, in terms of classification accuracy (see Table 1 on Left side) and time taken to build the model in seconds (see Table 1 on Right side). Overall, LDA offered the highest performance, yielding 96% accuracy for the original feature vector. However, SVM is the best performing classifier for the “50 features subset” with 93% accuracy. However, for the sake of concision, we do not show the detailed metrics for each activity per classifier. Instead, Table 2 shows the averaged precision (macro-PR), recall (macro-RC) and F-measure (macro-F1) for the “50 features subset” classifiers.

**Table 1.** Comparison of classifier performance for different number of features (N.F): Left side: overall accuracy; Right side: model building time (in seconds).

N.F	LDA	KNN	SVM	RF	N.F	LDA	KNN	SVM	RF
<b>20</b>	0.90	0.87	0.90	0.84	<b>20</b>	1.5	1.6	4.7	10.7
<b>50</b>	0.92	0.91	0.93	0.91	<b>50</b>	2.0	1.8	6.9	23.4
<b>561</b>	0.96	0.90	0.93	0.91	<b>561</b>	21.1	3.0	43.0	276

**Table 2.** Averaged evaluation metrics for the “50 features subset” classifiers.

N.F (50)	Macro-PR	Macro-RC	Macro-F1
LDA	0.91	0.92	0.91
KNN	0.91	0.92	0.91
SVM	0.94	0.94	0.93
RF	0.92	0.92	0.92

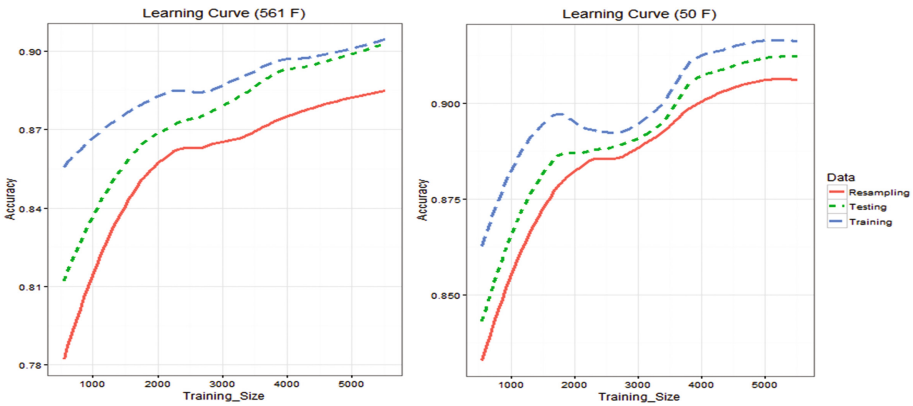
**Table 3.** Confusion matrix of best performing classifier on test data: (Predicted activities (*P*) vs Actual activities)

Activities	Laying	Sitting	Standing	Walking	Walking-downstairs (T)	Walking-upstairs (T)
Laying ( <i>P</i> )	537	0	0	0	0	0
Sitting ( <i>P</i> )	0	397	35	0	0	0
Standing ( <i>P</i> )	0	94	497	0	0	0
Walking ( <i>P</i> )	0	0	0	488	12	37
Walking-downstairs ( <i>P</i> )	0	0	0	6	370	10
Walking-upstairs ( <i>P</i> )	0	0	0	2	38	424

We have also tested our final SVM model on unseen and unlabeled data, which we kept strictly apart. We noticed that the accuracy has slightly decreased (92%), which means that there was almost no overfitting in the training phase. In Table 3, the confusion matrix is presented. We first notice a perfect classification between “moving” and “non-moving” activities. However, the classifier sometimes confuses and misclassifies one activity from another when there are inter-class similarities. Actually, it

confuses little bit between all types of walking activities, but little more between “Sitting” and “Standing”. On the contrary, the “Laying” activity is perfectly classified.

Furthermore, in order to figure out how well they perform over different sized versions of the training set, we have also simulated the learning curves of the final SVM models obtained before and after feature selection (see Fig. 1). For this purpose, the original Training data (7352 instances) was partitioned into Training set (75%) and Test set (25%), and 10-fold cross validation was used for resampling. As expected, reducing the feature space helped reducing the amount of training data needed to reach the same classification performance. For example, in order to reach 90% of accuracy, approximately 5000 instances were needed before feature selection (Fig. 1, Left side), while only 3500 instances were needed after feature selection (Fig. 1, Right side).



**Fig. 1.** Learning curves of the RBF SVM final models obtained: before feature selection (Left side), after feature selection (Right side).

## 4 Conclusions and Future Work

In this paper, we have conducted a comparison study using Smartphone accelerometer and gyroscope sensors data obtained from a publicly available HAR dataset. As the dimensionality of the original feature set is very high, we used feature selection approaches in order to reduce the feature space before classifying activities. Results show that, with RFE algorithm, only around 9% of the original feature set is needed to achieve the best tradeoff between classification accuracy, model building time, and confusion matrix. This comparison study is our starting point towards finding energy efficient techniques for real time HAR based on Smartphone sensors. Further research would involve an accelerated implementation of the proposed model, which might take advantage of specific computation platforms, alongside with energy consumption and performance analysis. HAR over smartphones is also under development for diabetic control and prediction of hypoglycemia [14].

## References

1. Shoaib, M., et al.: A survey of online activity recognition using mobile phones. *Sensors* **15**(1), 2059–2085 (2015)
2. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: Energy efficient smartphone based activity recognition using fixed-point arithmetic. *J. UCS* **19**, 1295–1314 (2013)
3. Frank, J., Mannor, S., Precup, D.: Activity recognition with mobile phones. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) *ECML PKDD 2011*. LNCS, vol. 6913, pp. 630–633. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-23808-6\\_44](https://doi.org/10.1007/978-3-642-23808-6_44)
4. Ouchi, K.; Doi, M.: Indoor-outdoor activity recognition by a smartphone. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, Pittsburgh, PA, USA, pp. 600–601, 5–8 September 2012
5. Kose, M., Incel, O.D.; Ersoy, C.: Online human activity recognition on smart phones. In: *Proceedings of the Workshop on Mobile Sensing: From Smartphones and Wearables to Big Data*, Beijing, China, pp. 11–15, 16 April 2012
6. Wang, Y., Lin, J., Annavam, M., Jacobson, Q.A., Hong, J., Krishnamachari, B., Sadeh, N.: A framework of energy efficient mobile sensing for automatic user state recognition. In: *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, Krakow, Poland, pp. 179–192, 22–25 June 2009
7. Yan, Z., Subbaraju, V., Chakraborty, D., Misra, A., Aberer, K.: Energy-efficient continuous activity recognition on mobile phones: an activity-adaptive approach. In: *Proceedings of the 2012 16th International Symposium on Wearable Computers (ISWC)*, Newcastle, Australia, pp. 17–24, 18–22 June 2012
8. Figo, D., Diniz, P.C., Ferreira, D.R., Cardoso, J.M.: Preprocessing techniques for context recognition from accelerometer data. *Pers. Ubiquitous Comput.* **14**, 645–662 (2010)
9. Kotsiantis, S.B.: Feature selection for machine learning classification problems: a recent overview. *Artif. Intell. Rev.* **42**, 157 (2014). <https://doi.org/10.1007/s10462-011-9230-1>
10. Verleysen, M., François, D.: The curse of dimensionality in data mining and time series prediction. In: Cabestany, J., Prieto, A., Sandoval, F. (eds.) *IWANN 2005*. LNCS, vol. 3512, pp. 758–770. Springer, Heidelberg (2005). doi:[10.1007/11494669\\_93](https://doi.org/10.1007/11494669_93)
11. Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv. (CSUR)* **46**, 33 (2014)
12. UCI Machine Learning Repository: Human Activity Recognition Using Smartphones Data Set (2012). <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>
13. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: A public domain dataset for human activity recognition using smartphones. In: *ESANN*, April 2013
14. De Florio, V., Bakhouya, M., Elouadghiri, D., Blondia, C.: Towards a smarter organization for a self-servicing society. In: *Proceedings of the 7th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, pp. 1–6 (2016)