Human Activity Recognition Using Extreme Gradient Boosting (XGBoost) and PCA Principal Component Analysis (PCA)

Sami Alperen Akgun
sami.alperen.akgun@gmail.com

***Abstract*—In this paper, XGBoost classifier were trained for HAR and 99.66% accuracy were obtained after feature extraction with PCA.**

***Index Terms*—XGBoost classifier, PCA, HAR, multi-class classification**

I. Introduction

Human Activity Recognition (HAR) has been studied extensively since early 2000s due to its promising applications such as elderly monitoring [1], weight-loss programs [2], digital assistants [3] and military purposes [4]. Thanks to all devices connected to each other since Internet of Things concept started, HAR became relatively easier to achieve with various low cost sensory inputs [5].

There are two main types of sensors that have been used for HAR purposes. The first type of sensors are called external sensors. They are usually placed in particular points in the environment. Inference of activities of users depends on their voluntary or semi voluntary interactions with the placed sensors. Researchers have implemented various successful HAR systems with these type of sensors like RGBD cameras [6] and sensors in smart homes [7]. The second type of sensors are wearable sensors which are worn by the user or attached to the user. Accelerometers, gyroscopes, magnetometers etc. on smartphones and smartwatches were employed to create efficient HAR systems [8]–[10].

In this project, we focused on wearable sensors due to their low-cost and availability. To create a classifier that is able to detect different human activities, a publicly available dataset "Human Activitiy Recognition Using Smartphones Data Set" [11] was used.

The rest of the paper was organized as follows. Section II explained the related work on HAR with wearable sensors. The background was explained in Section III. Section IV discussed methodology. Section V introduced results. Conclusion was made in Section VI following future work in Section VII.

II. Related Work

In this section, related work that uses wearable sensors for HAR purposes were analyzed from machine learning and pattern recognition perspective. There are many papers in this area to achieve better accuracy using wearable sensor data with various machine learning algorithms. For example, in [12], researchers employed wrist and ankle worn accelerometer to classify four different youth activity (sedentary, cycling, ambulation and other activities) with Support Vector Machine classifier. The implemented classifier were validated using leave-one-out cross validation. 95% accuracy were obtained for ankle data and 84.7% accuracy were obtained for wrist data.

In another paper [13], researchers used different types of Hidden Markov Models (HHM) to train classifier for given data composed of accelerometer, GPS and magnetometer sensory inputs. They obtained 99.1% accuracy with cHMM-based (continuous emissions HMM) sequential classifier and 92.2% accuracy with single-frame GMM (Gaussian Mixture Model) classifier.

Random forest classifier were employed in [14] to recognize everyday life activities with a new set of features extracted from wearable data by authors of the paper. They obtained up to 94% accuracy with their proposed method on the dataset they had.

Logistic regression [15], K-nearest neighbor (KNN) [16], decision tree [17] and neural network [18] classifiers were also employed for HAR purposes and they all managed to provide high accuracies for the specific datasets used in the mentioned papers.

Finally, ensemble learning approaches were employed to further improve the results obtained by the classifiers mentioned above like ensemble learning using KNN as a weak classifier [19] or ensemble learning based on random forest as a weak classifier [20]. A special type of ensemble learning method called Extreme Gradient Boosting (XGBoost) was used to classify human actions by researchers in [21]. They compared XGBoost performance with other commonly used machine learning methods, such as SVM, Naive Bayes (NB), KNN, Random Forest (RF) etc and found out that XGBoost can achieve better results in activity classification based on inertial sensors. In this project, we also focused on XGBoost method since the dataset we have is composed of inertial sensors data.

III. Background

A. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) algorithm has many other different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines. In this paper, XGBoost name was preferred and the paper [21] and the talk given by Tong He[1] (author of R Package XGBoost) were followed to implement the algorithm.

Extreme gradient boosting algorithm is an ensemble learning algorithm which means it creates a final model based on a collection of individual models. Decision trees are used

[1]https://www.youtube.com/watch?v=ufHo8vbk6g4

as a weak individual model. It is a supervised learning algorithm capable of predicting both classification and regression problems. Like gradient descent, gradients are used to minimize a loss function. Nonetheless, it uses more accurate approximations to find the best tree model by computing second order partial derivatives of the loss function (similar to Newton's method). In this way, it provides more information about the direction of gradients than gradient descent method. To improve model generalization by avoiding overfitting, it employs advanced regularization techniques [22].

Model specification of XGBoost is as follows. Suppose we have K trees, the model will become collection of all decision trees:

$$\sum_{k=1}^{K} f_k \tag{1}$$

The prediction will be done by:

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right) \tag{2}$$

where x_i is the feature vector for the i^{th}sample

Multi logarithm loss (mlogloss) function L is used for multi-class classification with regularization term Ω:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log\left(p_{i,j}\right) \text{ and } \Omega = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$

where T is # of leaves, and w_j^2 is the score on the j^{th} leaf (3)

After combining loss function and regularization term, the objective function becomes:

$$Obj = L + \Omega \tag{4}$$

To train the algorithm, gradient descent is used to optimize objective function. For this purpose, objective function is redefined for an iterative solution:

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{N} L\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t}\Omega\left(f_i\right) \\ &= \sum_{i=1}^{N} L\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(x_i\right)\right) + \sum_{i=1}^{t}\Omega\left(f_i\right) \end{aligned} \tag{5}$$

At this step, unlike standard gradient decent, both first and second order gradient are used to optimize redefined objective function. To achieve this, second-order Taylor series approximation of objective function is calculated:

$$Obj^{(t)} \simeq \sum_{i=1}^{N}\left[L\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2}h_i f_t^2\left(x_i\right)\right] + \sum_{i=1}^{t}\Omega\left(f_i\right)$$

where $g_i = \partial_{\hat{y}^{(t-1)}} L\left(y_i, \hat{y}^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{\gamma}^{(t-1)}} L\left(y_i, \hat{y}^{(t-1)}\right)$ (6)

After removing the constant terms, the objective function is simplified to:

$$Obj^{(t)} = \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right) + \frac{1}{2}h_i f_t^2\left(x_i\right)\right] + \Omega\left(f_t\right) \tag{7}$$

This simplified objective function represents the objective at the t^th step. It is inserted to decision tree building algorithm with boosting approach (similar to Adaboost approach) to find an optimal f_t so that predictions will be improved along the gradient in the tree bulding algorithm.

Lastly, it is good to explain some of the hyper-parameters of XGBoost algorithm related to this project[2]:

- **eta**: Step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative.
- **max_depth**: Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
- **objective**: Specifies the learning task and the corresponding learning objective.
- **booster**: Specifies which booster (weak classifier) to use.
- **n_estimators**: Number of gradient boosted trees. Equivalent to number of boosting rounds.

IV. Methodology

A. Dataset

The dataset "Human Activitiy Recognition Using Smartphones" were employed for this project [11]. The data was obtained from 30 subjects doing six daily activities WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING and LAYING. The signals were collected using accelerometer and gyroscope on waist-mounted Samsung Galaxy S II. 70% of the data was randomly divided into training set and the rest 30% was used as a test set.

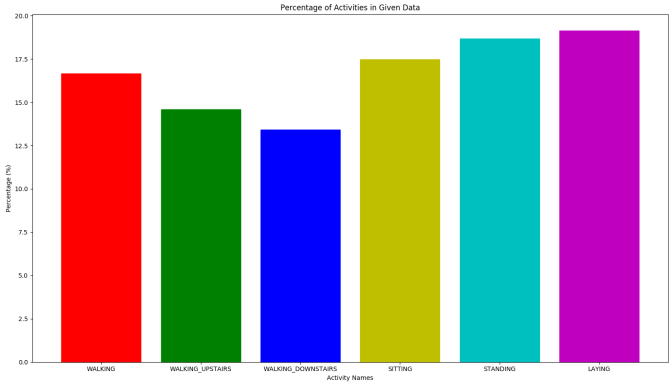
B. Why XGBoost?

There are some reasons that XGBoost was preferred over other machine learning methods for the specific dataset. Firstly, XGBoost classifier gives the best accuracy for activitiy recognition with inertial sensors [21] among traditional ML methods, and the specific dataset used in this project were obtained using inertial sensors. Secondly, the library XGBoost was written in an efficient way so that it pushes hardware limits for better performance [23]. This reduces training time for this classifier. Lastly, XGBoost is by far the best algorithm to win machine learning competitions like Kaggle competition since it takes bias-variance tradeoff into consideration during model fitting [24].

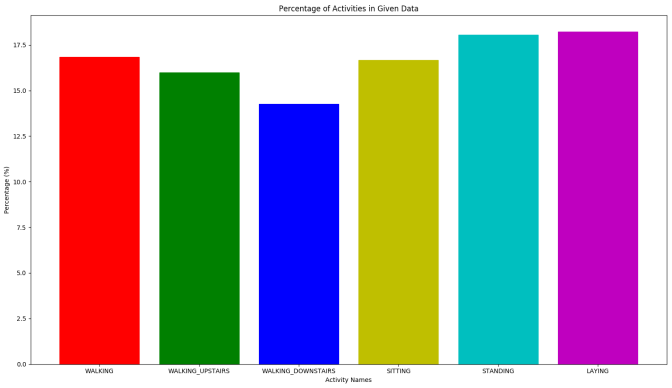
C. Procedure

To use XGBoost in practice, the dataset was analyzed to gain deeper understanding about the given dataset. Firstly, number of samples per activity were calculated to see whether we have a balanced training and test set since imbalanced datasets usually require different approaches than balanced ones [25]. As one can see in Figure 1, given dataset is balanced, i.e. the number of samples per class is similar to each other.

[2]https://xgboost.readthedocs.io/en/latest/parameter.html



(a) Training Set



(b) Test Set

Fig. 1. Percentage Wise Class Distribution for the Dataset

Fig. 2. Correleation Heatmap of Features

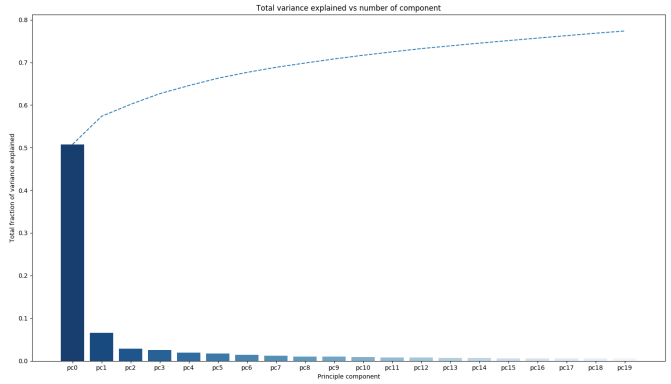


Fig. 3. First 20 Principal Components in PCA

Secondly, since the original data has so many features, the correlation between features was examined and as it can be seen in Figure 2, most of the features are correlated to each other. After this observation, Principal Component Analysis (PCA) was employed for feature extraction. To use Principal Component Analysis (PCA) for properly, training and test set features were standardized to have zero mean and one variance. Then, all the principal components corresponding to 95% variance retaining (102 components) were selected. To better understand this step, first 20 principal components were plotted like in Figure 3.

After feature extraction step, XGBoost classifiers were trained with different parameters and parameters that give the best accuracy after 10-times-10-fold cross validation were selected, i.e. fine-tuning step. In the end, fine tune parameters were used to calculate accuracy and confusion matrix.

V. Results

Here are the best parameters obtained after fine-tuning with 10-times-10-fold cross validation:

```
learning_rate = 0.6
n_estimators = 800
max_depth = 2
objective="reg:logistic"
# Here is the function to call the classifier
clf_xgbo = XGBClassifier(learning_rate=learning_rate
    , n_estimators=n_estimators,max_depth=max_depth,
    objective="reg:logistic")
```

Here are the results of 10-times-10-fold cross validation with the best parameters:

```
Best Accuracy:  99.66101694915255
Mean Accuracy:  95.3866020984665
Var Accuracy:  9.814400670324089
```

Here is the calculated confusion matrix and accuracy for this setting (not the best case scenario that gives the best accuracy above):

```
Extreme Gradient Boosting Accuracy: 94.26
Confusion Matrix
[[479   1  16   0   0   0]
 [  5 455  11   0   0   0]
 [ 10  33 377   0   0   0]
 [  0   1   0 438  47   5]
 [  0   0   0  40 492   0]
 [  0   0   0   0   0 537]]
```

One can infer from these results that, selected classifier is actually quite successful for classification of given dataset with mean accuracy of 95.38% and the best accuracy of 99.66% despite of relatively high variance between cross validation scores. Calculated confusion matrix offers an interesting conclusion. The most miss-classified activities are activity 4 (SITTING) and activity 5 (STANDING). Trained classifier had difficult times to distinguish these two actions comparing to other actions. The second worst case is about the distinction between activity 2 (WALKING UPSTAIRS)

and activity 3 (WALKING DOWNSTAIRS). The interesting observation about these miss-classifications is that classifier is not very good to differentiate activities which are reverse of each other (walking upstairs vs downstairs or sitting vs standing). If one can thing about the nature of these activities, it is easy to pinpoint that these activities has many features in common only directions are reversed. Lastly, classifier were able to distinguish activity 6 (LAYING) perfect, which might mean that feature values are quite different for laying activity than other activities.

VI. Conclusion

In this paper, XGBoost classifier were trained for HAR and 99.66% accuracy were obtained after feature extraction with PCA.

VII. Future Work

In the future, the trained XGBoost model will be improved with further tuning of parameters related to XGBoost. Moreover, all the features will be used to get higher accuracy with better hardware. Furthermore, as a feature work, other ensemble learning methods will be employed with the current method such as bagging or boosting, and the trained XGBoost classifier will be used as a weak classifier in ensemble learning.

References

[1] Y. Hbali, S. Hbali, L. Ballihi, and M. Sadgal, "Skeleton-based human activity recognition for elderly monitoring systems," *IET Computer Vision*, vol. 12, no. 1, pp. 16–26, 2018.

[2] E. Velloso, A. Bulling, H. Gellersen, W. Ugulino, and H. Fuks, "Qualitative activity recognition of weight lifting exercises," in *Proceedings of the 4th Augmented Human International Conference*, 2013, pp. 116–123.

[3] D. Sánchez, M. Tentori, and J. Favela, "Activity recognition for the smart hospital," *IEEE intelligent systems*, vol. 23, no. 2, pp. 50–57, 2008.

[4] A. Mukherjee, S. Misra, P. Mangrulkar, M. Rajarajan, and Y. Rahulamathavan, "Smartarm: a smartphone-based group activity recognition and monitoring scheme for military applications," in *2017 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*. IEEE, 2017, pp. 1–6.

[5] A. Subasi, M. Radhwan, R. Kurdi, and K. Khateeb, "Iot based mobile healthcare system for human activity recognition," in *2018 15th Learning and Technology Conference (L&T)*. IEEE, 2018, pp. 29–34.

[6] A. Roitberg, A. Perzylo, N. Somani, M. Giuliani, M. Rickert, and A. Knoll, "Human activity recognition in the context of industrial human-robot interaction," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, 2014, pp. 1–10.

[7] P. Chahuara, A. Fleury, F. Portet, and M. Vacher, "On-line human activity recognition from audio and home automation sensors: Comparison of sequential and non-sequential models in realistic smart homes 1," *Journal of ambient intelligence and smart environments*, vol. 8, no. 4, pp. 399–422, 2016.

[8] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, 2016.

[9] F. Shahmohammadi, A. Hosseini, C. E. King, and M. Sarrafzadeh, "Smartwatch based activity recognition using active learning," in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2017, pp. 321–329.

[10] R. San-Segundo, H. Blunck, J. Moreno-Pimentel, A. Stisen, and M. Gil-Martín, "Robust human activity recognition using smartwatches and smartphones," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 190–202, 2018.

[11] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," 01 2013.

[12] A. Mannini, M. Rosenberger, A. Sabatini, and S. Intille, "Activity recognition in youth using single accelerometer placed at wrist or ankle," *Medicine and science in sports and exercise*, vol. 49, pp. 801–812, 04 2017.

[13] A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, 2010.

[14] P. Casale, O. Pujol, and P. Radeva, "Human activity recognition from accelerometer data using a wearable device," in *Pattern Recognition and Image Analysis*, J. Vitrià, J. M. Sanches, and M. Hernández, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 289–296.

[15] N. C. Krishnan and S. Panchanathan, "Analysis of low resolution accelerometer data for continuous human activity recognition," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 3337–3340.

[16] S. Kaghyan and H. Sarukhanyan, "Activity recognition using k-nearest neighbor algorithm on smartphone with tri-axial accelerometer," *International Journal of Informatics Models and Analysis (IJIMA), ITHEA International Scientific Society, Bulgaria*, vol. 1, pp. 146–156, 2012.

[17] J. Pärkkä, L. Cluitmans, and M. Ermes, "Personalization algorithm for real-time activity recognition using pda, wireless motion bands, and binary decision tree," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 5, pp. 1211–1215, 2010.

[18] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2015, pp. 1488–1492.

[19] A. Jurek, C. Nugent, Y. Bi, and S. Wu, "Clustering-based ensemble learning for activity recognition in smart homes," *Sensors*, vol. 14, no. 7, pp. 12 285–12 304, 2014.

[20] Z. Feng, L. Mo, and M. Li, "A random forest-based ensemble method for activity recognition," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 5074–5077.

[21] Y. Zhang and Z. Peng, "Human motion classification based on inertial sensors with extreme gradient boosting," in *2018 International Conference on Image and Video Processing, and Artificial Intelligence*, vol. 10836. International Society for Optics and Photonics, 2018, p. 108361R.

[22] S. Elsinghorst, "Machine learning basics - gradient boosting xgboost," Nov 2018. [Online]. Available: https://www.shirin-glander.de/2018/11/ml_basics_gbm/

[23] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[24] D. Nielsen, "Tree boosting with xgboost-why does xgboost win" every" machine learning competition?" Master's thesis, NTNU, 2016.

[25] M. M. Rahman and D. Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing*, vol. 3, no. 2, p. 224, 2013.