ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, Vol 12(2), DOI: 10.17485/ijst/2019/v12i2/141453, January 2019 ISSN (Online) : 0974-5645

A Robust Sampling Technique to Reduce Classification Time for Human Activity Recognition

Ahsan Memon*

Department of Computer Science, SZABIST Hyderabad, Pakistan; ahsan.memon@hyd.szabist.edu.pk

Abstract

Objectives: This study is an endeavor to provide quick, on-the-go classification of a human activity dataset with an aim to improve on the classification time of a machine learning algorithm for Human Activity Recognition (HAR) datasets. **Methods/Statistical analysis**: It proposes the use of a customized sampler called the Normal On-The-Go (Normal OTG) sampler to reduce the classification time. Concocted using a combination of stratified, random and normal sampling, the Normal OTG sampler was tested on HAR datasets and was found to significantly reduce the training time of the most commonly used machine learning algorithms. Three datasets, ShoaibSA, ShoaibPA and USC-HAD were used to conduct the experiments. **Findings:** It was found that using as little as 5% samples from the training dataset sampled by the Normal OTG sampler, sufficiently reliable accuracy was obtained from most of the 9 classifiers that were used. The results indicated that almost 96% of time was saved in the training process in the case of USC-HAD, and 62% and 83% time was saved in the case of ShoaibPA and ShoaibSA respectively. It was also found that the results were consistent among the three datasets. **Application/Improvements:** The study helps training of data in human activity recognition a faster process and thereof, making algorithm selection a less tedious procedure.

Keywords: Classification Time, Human Activity Recognition, Robust, Sampling Technique

1. Introduction

Human activity recognition has been an essential part of contemporary research owing to its importance in assisted living and ubiquitous computing. With its promising applications in the Internet of Things paradigm[1] and its public acceptability, it has been approached with statistical, probabilistic, logical reasoning and machine learning where the state-of-the-art activity recognition techniques have largely been attributed to machine learning techniques with continuous streaming as its target.

The process of machine learning activity recognition involves five steps as mentioned by[1]: Data Acquisition, Preprocessing, Feature Selection and Extraction, Training and Testing. State-of-the-art uses continuous streams of data for acquisition; creation of features based on discriminative models; training the models with adaptive and personalized approaches; and testing with hybrid classifiers. Author[3] provides an in-depth survey on the preprocessing, adaptive sensor selection and resource consumption of smartphone based activity recognition. Author[4] provides a detailed survey on feature selection and classifier evaluation of these sensors. Author[5] provides evaluation of a broad range of classifiers, used for activity recognition. Author[6] provides a comprehensive survey on challenges faced by live data streams.

According to the survey by[6], the biggest challenges that the research community faces in activity recognition are: scarcity of labelled data, recognition on evolved of activities, and lack of literature on adaptation/refinement of the classifier models. All of these challenges are an inevitable aftermath to the increasing amount of the overall data population.

With the growing population of data, it has become increasingly difficult to find a universal algorithm that accounts for all the diversity produced by varying activities and non-standard hardware. The change in available activities or appearance of new ones makes it imperative to assimilate them into existing model to achieve better real-life recognition[7]. The changes in real-time data also creates concept drifts that Bayes rule defines as the change in the prior and/or the likelihood. Therefore, model that performs optimally for all users in activity recognition are difficult to create[6]. The training and testing of data in

*Author for correspondence

a continuous sensor stream is hence, regularly repeated with a variety of machine learning classifiers to find the most appropriate classifiers for personalized models.

However, the perpetual increase in data population makes it a very expensive task to re-evaluated models every once in a while, therefore, rendering personalization models such as[8] and[9] computationally costly, making it one of the lingering challenges in the activity recognition domain. In such cases, there is a desperate need for schemes that take samples from large datasets with the goal of taking the least number of samples that yield similarly accurate predictions.

This paper aims to address the issue of re-training models by improving on the classification time by many folds. We do this by proposing a sampling scheme, the Normal On-The-Go Sampling, and evaluating its efficacy in the activity recognition setting based on sensory inputs.

The rest of this paper is organized as: Section II presents the Design of our Experimental System, Section III shows the implementation of proposed algorithm, Section IV presents the results and Section V provides the conclusions.

2. System Design

This section describes implementation of the five steps involved in activity recognition mentioned earlier.

A minimalistic implementation setting was chosen to distinctly observe the effect of sampling on a population for activity recognition. To demarcate the necessary processes involved in the design, we further divide this section into Datasets, Preprocessing, Feature Extraction and Selection and Training and Testing.

2.1 Datasets

Since the testing of the sampling algorithm on streaming data was not feasible because we needed pre-stored data for evaluating the running time and dataset distribution, the data was analyzed from datasets instead. For the purpose, we used three well-known datasets ShoaibPA, ShoaibSA and the USC-HAD for activity recognition. ShoaibPA[10] and ShoaibSA[11] appeared in their names in[12] and USC-HAD was published as[13]. All of these datasets used an accelerometer and a gyroscope for measurements.

Using these datasets helped achieve two goals: 1) the classifier models were tested over same samples, hence providing conclusive and standardized decisions on the effectiveness of the results, and 2) allowed for implementation of a relatively simple mechanism that helped clearly identify the trends in results.

2.2 Preprocessing, Feature Extraction, and Selection

Preprocessing of data involved steps of windowing and construction of the feature set. For windowing, we used the fixed width sliding window of 2 seconds with an overlap of 50% as this window size and overlap appeared to yield better results over other schemes[14–16]. The number of training samples obtained after application of a fixed length sliding window of the three datasets is given in Table 1.

Table 1. Training samples in datasets

	Sample Distribution (Percentage)		
	USC-HAD	ShoaibPA	ShoaibSA
Downstairs	12.8	14.4	20.4
Sitting	21.2	22.6	20.4
Standing	19.1	22.6	20.4
Upstairs	16	16.8	18.4
Walking	30.9	23.6	20.4
Total Samples	24686	2655	8820

On the other hand, a survey of most commonly used features provided in[4] suggests that four features, mean, power, standard deviation and interquartile range, were most commonly used human activity recognition. Hence, two of these features, mean and standard deviation were chosen for construction of the training dataset. Each feature was computed on the magnitude function of the three axis to allow for independence in device orientation that was mentioned in[16].

2.3 Training and Testing

The use of appropriate classifiers is paramount to accurate activity recognition. However, our choice of classifier was inspired more from our goal of developing a sampling mechanism that was more representative of the variety of classifiers. Hence, we used the set of classifiers enlisted in Table 2. A survey on the most commonly used classifiers can however, be found in[4]. A cross validation of 10 folds was also used for train-test splitting and evaluation as suggested in the survey by[17].

Table 2. Classifiers

Algorithm	Abbreviation
Logistic Regression	LR
Support Vector Classifier	SVC
K Nearest Neighbors	KNN
Gaussian Naïve Bayes	GNB

Perceptron	P
Linear Support Vector Classifier	LSVC
Stochastic Gradient Descent Classifier	SGDC
Decision Tree	DT
Random Forest	RF

3. Implementation of Normal OTG Sampler

The proposed OTG Sampler sampled from the three datasets while maintaining the Probability Distribution Functions (PDF) of features chosen for sampling. In Figure 1,

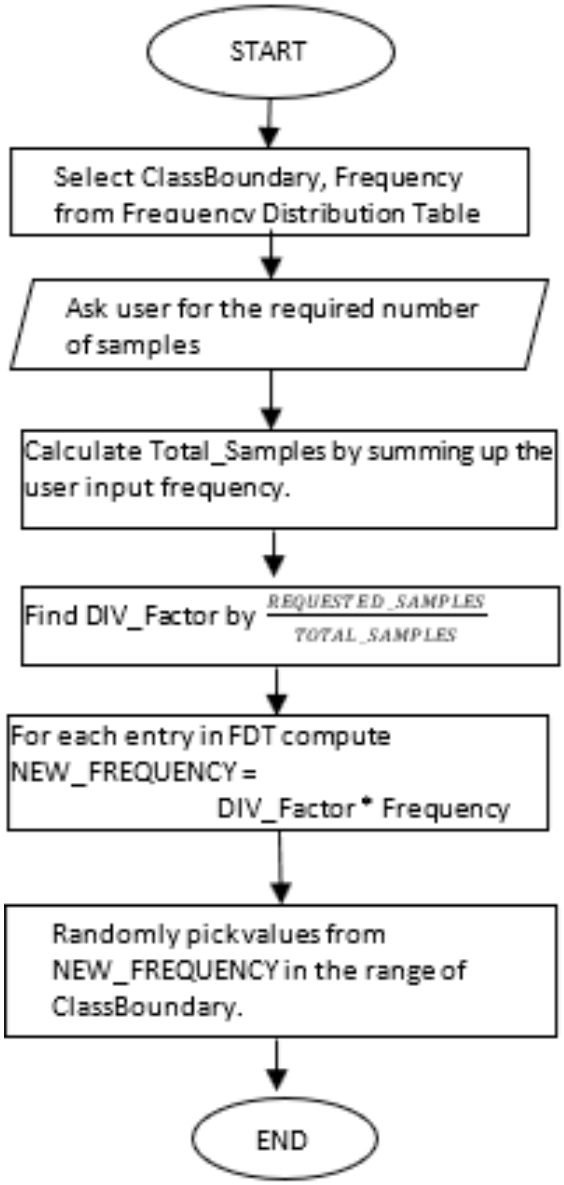

Figure 1. Normal OTG sampler algorithm.

Normal OTG sampler algorithm shows algorithm that was used for sampling. It was implemented on datasets with sample size ranging from 5% till 95% and the bin size chosen for the Frequency Division Table was 50. In Figure 2, Normal OTG sampler probability distribution functions show the distributions of the resulting datasets at each sample size. A Pearson constant of correlation was also formulated to statistically compare similarity in the resulting distributions. Figure 3 - Correlation of PDFs shows the comparison.

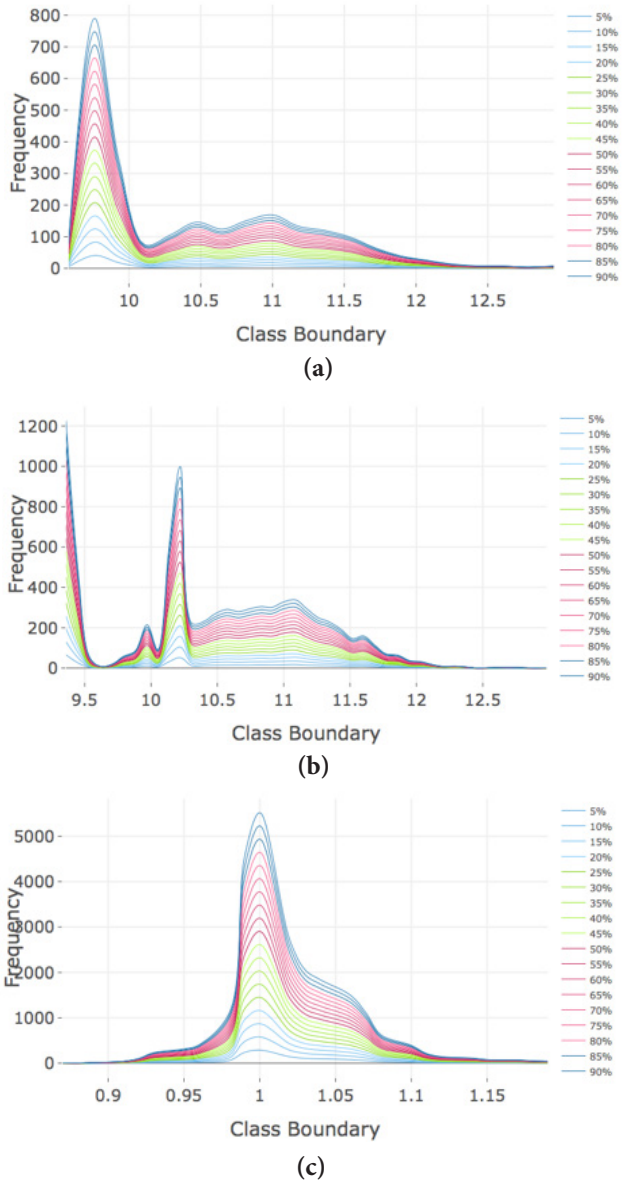

Figure 2. Normal OTG Sampled Probability Distribution Functions of (a) ShoaibPA, (b) ShoaibSA and (c) USC.

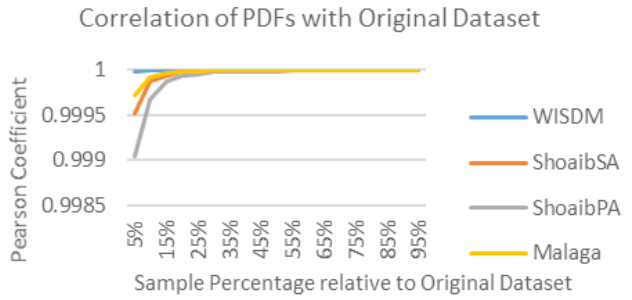

Figure 3. Correlation of PDFs.

4. Results and Discussion

At the outset, our results have been divided broadly into three parts: 1) a comparison of classification accuracies of nine classifiers that have been commonly used with HAR datasets, 2) a comparison of classifier results of the sampled datasets and 3) the impact of Normal OTG sampling on classification time.

The first set of results attempts in comparing the activity prediction accuracies by LR, SVM, KNN, GNB, P, LSVC, SGDC, DT and RF classifier on each one of the datasets. From the results shown in Figure 4, Classifier results with random OTG sampling, it was found that the RF, KNN and SVM produces the more accurate results than the rest. On the other hand, the figure also indicates that Normal OTG Sampling produced consistent results across the board.

The second set of results was obtained from the Random Forest classifier. In Figure 5, random forest classifier accuracy vs sample percentage and duration

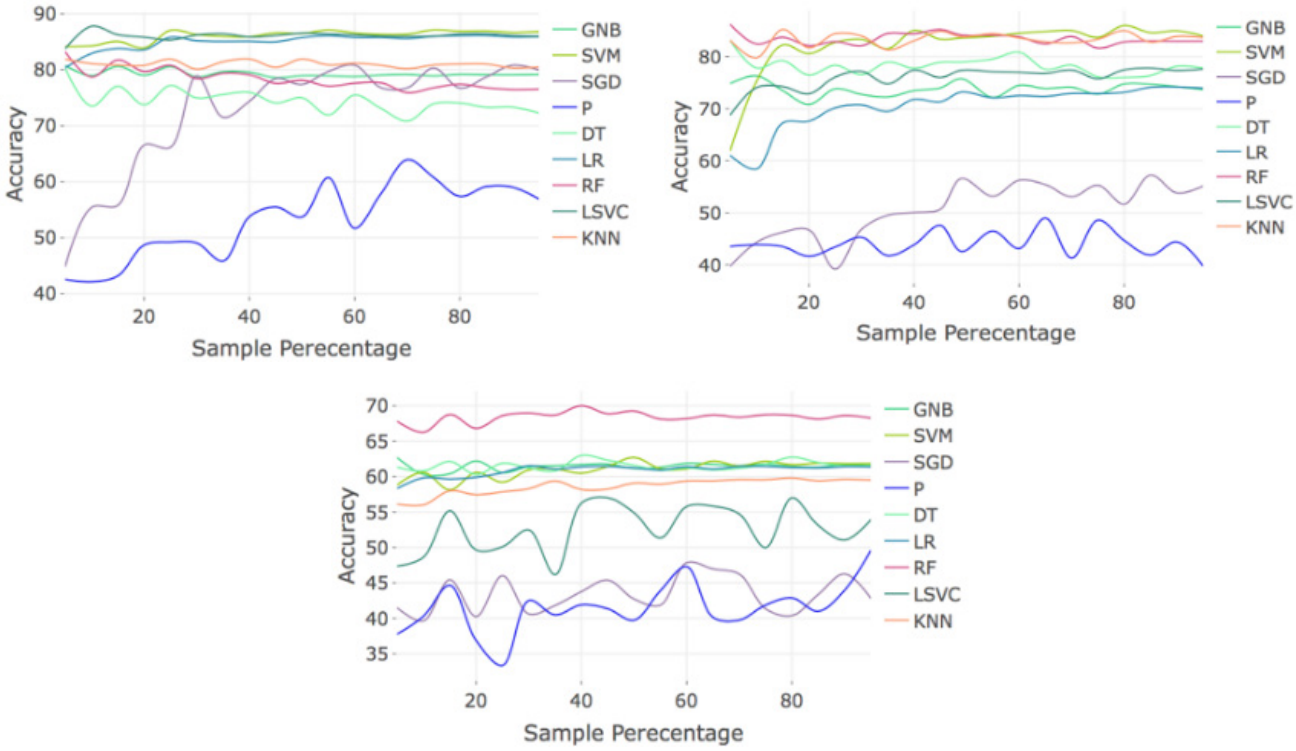

Figure 4. Classifier results with random OTG sampling.

compares the classifier accuracy with sample percentage along with time taken by the classifier to yield the results. The results evidently show that despite increase in duration taken by the classifier, the accuracy of the classifier accuracy remained relatively constant. The accuracy of 87% received from 5% samples in the ShoaibPA dataset was in fact, representative of accuracy results from unsampled dataset. The USC-HAD and ShoaibSA dataset also yielded similar results where the accuracy remained relatively constant. With greater time taken to compute each set of classifier results, we observe that 5% samples should be sufficient for sampling a dataset with the proposed sampling mechanism for testing a classifier out on data.

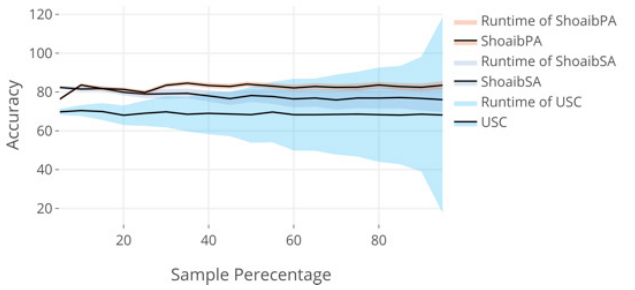

Figure 5. Random forest classifier accuracy vs sample percentage and duration.

The third set of results provides insights into the efficacy of Normal OTG sampler. In Figure 6, computational cost of normal OTG sampling shows the computational cost of Normal OTG sampler measured in Time/Sample. Whereas, in Figure 7, savings in classification time shows the relative time in percent that was saved by using the Normal OTG classifier. The time is calculated from the total time that it would have taken to compute the results with full dataset length for similar accuracy results. While the time itself is dependent on the computational machinery used and therefore cannot be generalized, the trend line indicates that the computational cost does not increase abruptly over increase in the sample size.

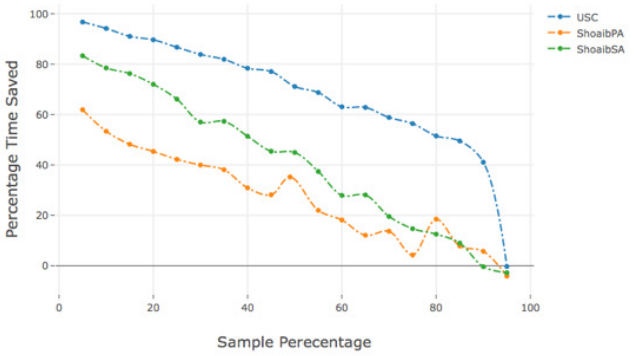

Figure 6. Computational cost of normal OTG sampling.

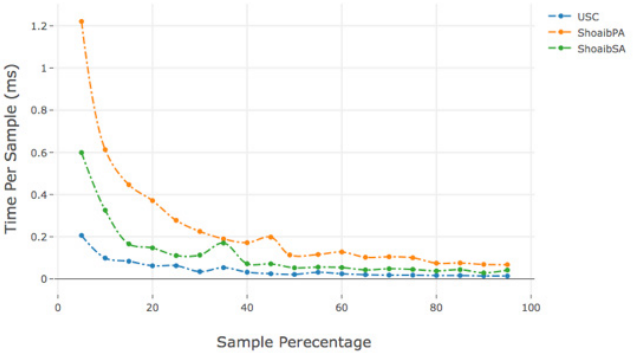

Figure 7. Savings in classification time.

Lastly, Figure 7, savings in classification time shows that larger datasets like USC-HAD save 96% of the total time they take to complete the classification process with just 5% samples from their dataset and similar accuracy results. However, smaller datasets such as ShoaibPA or ShoaibSA, that only take a few seconds to run, save up 62% and 83% of their running time.

5. Conclusions

In this paper, we proposed a dataset sampling mechanism that reduces the number of samples that are used to classify datasets with machine learning classifiers. Our proposed method involved sampling the dataset while keeping the same probability distribution function. We implemented the mechanism on three well-known Human Activity Recognition datasets: ShoaibSA, ShoaibPA and USC-HAD and discovered that a sample size of 5% was adequate for a reasonably accurate classification with the nine classifiers that we tested. It was observed that despite the increase in sample size from 5% of the total dataset size, up till 95%, there was no significant change in classification accuracy. Therefore, a small sample size should be sufficient for the classification process. With experiments, we found out that a larger dataset such as the USC-HAD, saved 96% time compared to what it would have spent if we would have trained all of its samples, while achieving a similar prediction accuracy with the Random Forest classifier. We have also studied the time per sample that it takes for the Normal OTG sampler to complete its sampling process and found that the proposed mechanism is not computationally heavy and can thus, be implemented on even larger datasets

6. Future Work

In future, we would like to investigate the effect of Normal OTG sampling with more features in the training dataset and examine its effect on duration and the classifier accuracy. We would also like to explore the use of this technique on sensor-based datasets that are not limited to HAR and investigate the effect of bin size on activity recognition.

7. References

1. Perera, C, Zaslavsky A, Christen P, Georgakopoulos D. Context aware computing for The Internet of Things: A Survey. IEEE Communications Surveys and Tutorials. 2014; 16:414–54. https://doi.org/10.1109/SURV.2013.042313.00197
2. Wannenburg J, Malekian R. Physical activity recognition from smartphone accelerometer data for user context awareness sensing. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2016; 47:1–8.
3. Shoaib M, Bosch S, Incel O, Scholten H, Havinga P. A survey of online activity recognition using mobile phones. Sensors. 2015; 15:2059–85. https://doi.org/10.3390/s150102059. PMid:25608213. PMCid:PMC4327117
4. Morales J, Akopian D. Physical activity recognition by smartphones, a survey. Biocybernetics and Biomedical Engineering. 2017; 37:388–400. https://doi.org/10.1016/j.bbe.2017.04.004
5. Peterek T, Penhaker M, Gajdoš P, Dohnálek P. Comparison of classification algorithms for physical activity recognition BT - innovations in bio-inspired computing and applications, Abraham A, Krömer P, Snášel V, ediotrs, Springer International Publishing; 2014. p. 123–31.
6. Abdallah ZS, Srinivasan B. Activity recognition with evolving data streams: A review. ACM Computing Surveys; 2018. p. 51.
7. Yang J, Lu H, Liu Z, Boda PP. Physical activity recognition with mobile phones: Challenges, methods, and applications BT - Multimedia interaction and intelligent user interfaces: Principles, methods and applications. Shao L, Shan C, Luo J, Etoh M, editor, Springer London; 2010. p. 185–213. https://doi.org/10.1007/978-1-84996-507-1_8
8. Gomes JB, Krishnaswamy S, Gaber MM, Sousa PAC, Menasalvas E. Mobile activity recognition using ubiquitous data stream mining. Proceedings of the 14th International Conference on Data Warehousing and Knowledge Discovery; 2012. p. 130–41
9. Weiss GM, Lockhart JF, Pulickal TT, McHugh PT, Ronan IH, Tim JL et al. Actitracker: A smartphone-based activity recognition system for improving health and well-being. 2016 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA); 2016. p. 682–8.
10. Shoaib M, Scholten H, Havinga PJM. Towards physical activity recognition using smartphone sensors. 2013 IEEE 10th International Conference on Autonomic and Trusted Computing; 2013. p. 80–7. https://doi.org/10.1109/UIC-ATC.2013.43
11. Shoaib M, Bosch S, Incel OD, Scholten H, Havinga PJM. Fusion of smartphone motion sensors for physical activity recognition. Sensors. 2014; 14.
12. Micucci D, Mobilio M, Napoletano P. UniMiB SHAR: A new dataset for human activity recognition using acceleration data from smartphones. 2016. https://doi.org/10.3390/app7101101
13. Zhang M, Sawchuk AA. USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. Proceedings of the 2012 ACM Conference on Ubiquitous Computing; 2012. https://doi.org/10.1145/2370216.2370438
14. Attal F, Mohammed S, Dedabrishvili M, Chamroukhi F, Oukhellou L, Amirat Y. Physical human activity recognition using wearable sensors. 2015; 15(12):31314–38. https://doi.org/10.3390/s151229858
15. Stisen A, Blunck H, Bhattacharya S, Prentow TS, Kjærgaard MB et al. Smart devices are different: assessing and mitigating mobile sensing heterogeneities for activity recognition with real world HAR dataset. Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems; 2015. p. 127–40. https://doi.org/10.1145/2809695.2809718
16. Jain A, Kanhangad, V. Human activity classification in smartphones using accelerometer and gyroscope sensors. IEEE Sensors Journal. 2018; 18(3): 1169–77.
17. Janidarmian M, Fekr AR, Radecka K, Zilic Z. A comprehensive analysis on wearable acceleration sensors in human activity recognition. Sensors. 2017; 17(3):529. https://doi.org/10.3390/s17030529