Intelligent Environments 2018
I. Chatzigiannakis et al. (Eds.)

doi:10.3233/978-1-61499-874-7-129

Dealing with Imbalanced Data Sets for Human Activity Recognition Using Mobile Phone Sensors

Ky Trung NGUYEN [a], François PORTET [a] and Catherine GARBAY [a]
[a] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France*

Abstract. In the recent years, the wide spreading of smart-phones that are daily carried by humans and fit with tens of sensors triggered an intense research activity in human activity recognition (HAR). HAR in smartphones is seen as essential not only to better understand human behavior in daily life but also for context provision to other applications in the smartphone. Many statistical and logical based models for on-line or off-line HAR have been designed, however, the current trend is to use deep-learning with neural network. These models need a high amount of data and, as most discriminative models, they are very sensitive to the imbalanced class problem. In this paper, we study different ways to deal with imbalanced data sets to improve accuracy of HAR with neural networks and introduce a new over-sampling method, called Border Limited Link SMOTE (BLL SMOTE) that improves the classification accuracy of Multi-Layer Perceptron (MLP) performances.

Keywords. human activity recognition, smartphone, over-sampling, class imbalance problem, context-aware computing, Ambient Intelligence

1. Introduction

Human Activity Recognition from wearable sensors and in particular smartphones has been subject of an intense research and industrial activity this last decade [1]. Many learning algorithms have been used to classify physical human activities such as *Running, Walking, etc.* as well as interactive and social activities (chatting, talking, playing, etc.). HAR is useful for health monitoring, senior care and personal fitness training as well as for providing context to smartphone applications. Physical human activities are generally classified from recorded sensor data (e.g. accelerometers, GPS, audio, etc.) which are embedded into wearable devices (e.g. smartphones and smart watches).

HAR systems performances are highly dependent on the classification model (Decision Tree, Support Vector Machine, Multi-Layer Perceptron, etc.), the feature used, the number of classes and the size of the datasets available for training [1]. However, another aspect that plays an important role in this domain is the lack of an uniform collection of different activities. In fact, this is the case for most smartphones datasets (e.g. *Running* = 4% and *Walking* =40% distribution). This is called the Class Imbalance Problem that is known to have a serious influence on the performance of learning algorithms, because most standard algorithms expect balanced class distributions [2].

In the past, research on HAR based on wearable sensor did not systematically handle the class imbalance problem. Therefore, in this paper we introduce a generic framework that integrates active learning with over-sampling method based on MLP to overcome this problem. We also introduce a new over-sampling method, called BLL SMOTE - an extension of SMOTE [3] - which can apply to non-convex spaces.

Contributions. Our contributions are summarized as follows. (i) A framework integrating MLP and active learning with over-sampling. (ii) A new over-sampling method, BLL SMOTE. (iii) Experiments with 2 available datasets that show the impact of taking the class imbalance problem into account in the learning.

The paper is organized as follows. Section 2 presents a summary of the state of the art in HAR and in learning techniques with imbalanced data. The overall framework and the BLL SMOTE method are detailed in Section 3. Several experiments are reported in Section 4. The paper ends with a short discussion and an outlook of future work.

2. Related Work

Human Activity Recognition from wearable sensors data is a very rich domain of research. We restrict here in presenting the main work regarding the classification models being used, the available datasets and the techniques to deal with imbalanced class distribution in data.

Regarding the classification models, there have been many approaches to deal with HAR from wearable sensors. Over the last decade, the most common approach is to process windows of data streams to extract a vector of features which will in turn be fed to a classifier. Many instance-based classifiers have been used in the field, such as Bayesian Network [4], Decision Trees [4,5], Random Forest [5], Artificial Neural Network (ANN) [4, 6], Support Vector Machines (SVM) [4, 7], etc. Since human activities can be seen as a sequence of smaller sub-activities, sequential models such as Conditional Random Fields [5], Hidden Markov Model [8] or Markov Logic Network [9] have also been applied. However, since the advent of Deep Learning, ANN have become of the most popular model in HAR from wearable sensors [10, 11].

Machine learning is highly dependent on datasets. It is even more the case with Deep Learning. The survey by [1] presents a large number of datasets acquired from a smartphone. However, it also shows the lack of uniformity in tasks, sensors, protocol, time windows, etc. It is worth to notice that most of the datasets are restricted to inertial sensors such as accelerometers. The audio sensors are largely ignored while being among the only ones that are always found on a smartphone. It is also worth noticing that some are very imbalanced since the distribution among classes are very different. For instance, in the ExtraSensory Dataset [12], sitting represents 44.2% of the data while running only 0.3%. In this case, the learning approach should consider the class imbalance problem.

Imbalanced data has a serious influence on the performance of learning algorithms, because most standard algorithms expect balanced class distributions, as reported in [2]. Hence, datasets exhibiting imbalanced class distribution make these algorithms fail to correctly represent the distributive characteristics of the data. As a consequence, it would produce mis-classification of minority classes higher than mis-classification of majority classes, and leads to a decrease in the overall accuracy of learning algorithms. In fact, in HAR, a few studies coped explicitly with this problem such as [13] who proposed Weighted Support Vector Machines (WSVM) to improve learning of minority classes. However, the approach is based on a scheme that put more weight on the errors on the minority classes than on the majority classes. Therefore, this approach is highly dependent on the instances of the minority classes.

In general, in order to deal with imbalanced data, several other approaches were introduced in [2] such as over-sampling and active learning. For the former approach, some methods were proposed such as SMOTE [3] or Borderline SMOTE [14] which works by generating new synthetic instances of minority classes. Their studies showed that over-sampling techniques succeeded to enhance the classification accuracy for imbalanced datasets. For the latter approach, [15] introduced a SVM-based active learning framework in which SVM starts to train on a given training dataset, then selects the most informative instances from a pool of training samples, afterward adds the newly selected instances to the training set and finally trains SVM again. This approach has been pursued in the VIRTUAL framework [16]. The study showed that active learning can efficiently handle the class imbalance problem. However, all above-mentioned studies did not combine together in order to settle the imbalance data problem. Therefore, in this paper we introduce a generic framework to cope with this overall issue. More details will be provided in Section 3.

3. Oversampling and active learning framework for HAR

Our objective is to improve the learning of HAR model in case of imbalanced datasets. The problem can be defined as follows : Let $A = \{a_1, ..., a_k\}$ be the set of all activities, given a set $T = \{t_1, ..., t_m\}$ of m equally sized time windows, and a set of sensors $S_i = \{S_{i,1}, ..., S_{i,q}\}$. Given a feature space $X \in R^n$, an instance $x \in X$ extracted from sensors S_i at time frame t_j is to be classified, e.g. attached an activity label from A.

In this paper, we focus on the classification problem. Our goals are (1) to find a learning algorithm $f : X \to A$ returning a label $f(x) = a^*$ as close as possible to the actual activity performed during $t_i \in T$, (2) to enhance the classification task using active learning, and (3) to improve the recognition task by over-sampling to balance the imbalanced training set.

In this section, we present the general framework to reach these objectives, and then detail each of its components in the subsequent sections.

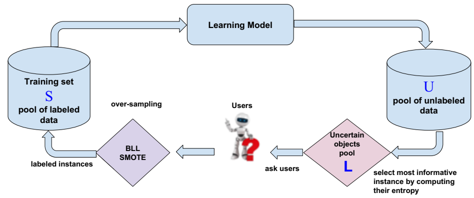

Figure 1. The active over-sampling framework

3.1. Proposed Framework

The framework, as shown in Figure 1, is an extension of VIRTUAL [16] to integrate active learning with over-sampling method to overcome the imbalance problem.

First of all, the learning is initiated with a pool of training instances S, which is used to learn a classification model. Then, to choose the most relevant sample to add in the training set, the entropy of each instance from pool of unlabeled data U is computed using the classifier output by using Equation (1). From this, a small pool of uncertain samples L is created by grouping the instances that maximized the Shannon entropy. After that, the small pool L is removed from U and user is queried for its labels.

Secondly, once L is annotated, our specific over-sampling method, called BLL SMOTE (cf. Section 3.4), looks for minority instances inside the pool L and generates new artificial instances of these minority classes. The original pool L plus the generated instances of the minority classes are added to the training set S and the training restarts. This means that at each iteration, the training set is bigger but less and less imbalanced. Each part of this framework is detailed in the following sections.

3.2. Classification model: Multi-Layer Perceptron

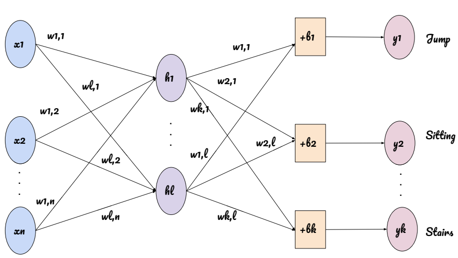

Figure 2. Multilayer Perceptron

For activity classification, many existing techniques such as SVM [7] or Random Forest [5] can be used. Among them, MLP is one of the most common methods used in HAR systems. In the rest of this paper, MLP is used as classification model. This choice is, on the one hand, justified by the fact that it demonstrates high performances on the task as well as high promises of improvement and, on the second hand, by the incremental learning strategy that fits well with active learning.

MLP can be seen as a class of feed-forward neural network composed of at least three layers of nodes, namely input, hidden and output layer. MLP can be learned using the back-propagation method, an efficient optimization method that operates iteratively. More precisely, our MLP network is designed as follows. Given inputs $X = \{x_1, ..., x_n\}$ are the features extracted from the sensors, the hidden layer nodes $H = \{h_1, ..., h_l\}$ and the output layer nodes $Y = \{y_1, ..., y_k\}$ are computed as follows:

$$\begin{bmatrix} h_1 \\ . \\ . \\ . \\ h_l \end{bmatrix} = \begin{bmatrix} w_{1,1} * x_1 & +... & +w_{1,n} * x_n \\ . & . & . \\ . & . & . \\ . & . & . \\ w_{l,1} * x_1 & +... & +w_{l,n} * x_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_k \end{bmatrix} = \begin{bmatrix} w_{1,1} * h_1 & +... & +w_{1,l} * h_l \\ . & . & . \\ . & . & . \\ . & . & . \\ w_{k,1} * h_1 & +... & +w_{k,l} * h_l \end{bmatrix} + \begin{bmatrix} b_1 \\ . \\ . \\ . \\ b_k \end{bmatrix}$$

This is shown in Figure 2, where $\{w_{1,1}, ..., w_{l,n}\}$, $\{w_{1,1}, ..., w_{k,l}\}$ and $\{b_1, ..., b_k\}$ represent weights and bias.

3.3. Active Learning

The principle of Active Learning (AL) is to learn to label unknown instances by selecting (querying) some specific instances and ask an external system (e.g., a human operator) to label them. It has become an emerging research topic with applications in many fields such as image segmentation [17], data clustering [18] and interactive data analysis [19]. Applying AL in HAR is thus an interesting approach since it can further boost up the accuracy by involving humans in the classification task, especially for hard to classify activities. Moreover, its scheme provides a natural way to cope with data imbalance by exploring some most uncertain data spaces, as pointed out in [15].

Typically, an active learning algorithm chooses objects that their labels are among the most uncertain ones to query users for. Uncertain instances can be chosen in many different ways [20]. Our technique is built upon the uncertainty sampling technique [20] whose principle is that the most relevant instances to be selected for annotation are the ones for which the estimates are the less certain. Thus, after MLP training, we predict the labels of U using the training output Y of MLP. Y can be seen as a vector of probability of labels. Then the instances in U are ranked according to their decreasing Shannon Entropy, because the higher entropy of an instance is, the more uncertainty there is on its class. Therefore, the most uncertainty instance can be picked up by maximized Shannon entropy [21] using Equation (1):

$$x_H^* = arg \max_x - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x) \tag{1}$$

where x is an instance, $P_\theta(y_i|x)$ is the probability of all possible labels on the instance.

3.4. Over-sampling Border Limited Link SMOTE Method

While classical active learning methods only add in the training set the uncertainty instances, that were labeled by a user, our method also performs over-sampling on queried data. This makes it possible to put new information into the training set and tackle the class imbalance problem.

Over-sampling consists in adding new sample to a training set, whether they are synthetic or real. For instance, SMOTE [3] generates a new synthetic instance, using Equation (2):

$$x_{new} = x_i + (x_i^\theta - x_i) * \lambda \tag{2}$$

where x_{new} is the new sample generated from $x_i \in S_{min}$, with S_{min} is the samples of minority class, x_i^θ is one of the k-nearest neighbors of x_i: $x_i^\theta \in S_{min}$, and $\lambda \in [0,1]$ is the random number, which allows to randomly generate the new synthetic instance x_{new} along the line between x_i and x_i^θ.

However, this method is not relevant in case of non-convex spaces. For instance, imagine a space as represented Figure 3. If x_i and x_i^θ are two samples of the green (circle) class, a direct application of Eq. (2) would produce a new sample x_{new} which would not be in the right space.

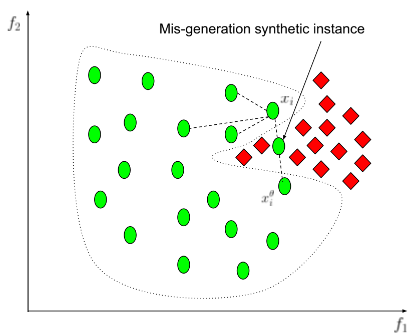

Figure 3. Example of mis-generation synthetic instance in non-convex dataset

To avoid the mis-generation of synthetic instances in the case of non-convex dataset, we introduce the BLL SMOTE method described as follows. The method uses Eq. (2) but calculates the distance from x_{new} to each of the k-nearest neighbors of x_i, denoted as $d_j = d(x_{new}, x_i^\theta),\ j = 1, ..., k$, where d is the Euclidean distance. Then, the distance of the artificial instance x_{new} with its nearest instance $x_{diff} \notin S_{min}$ such that $x_{diff} \in S$, denoted as $d_{diff} = d(x_{new}, x_{diff})$ is computed. Finally, each d_j is compared to d_{diff}. If any d_j is greater than d_{diff}, then this artificial instance x_{new} is not accepted to be generated. Otherwise, x_{new} is accepted.

An advantage of BLL SMOTE is to avoid the mis-generated new synthetic instance in non-convex datasets

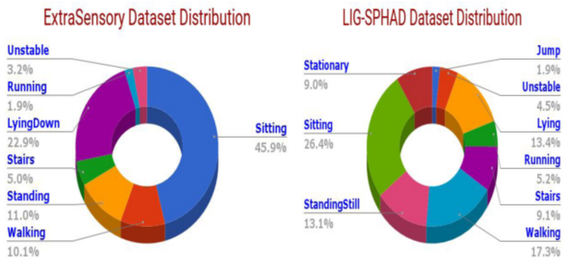

Figure 4. Distribution of the activity labels over the datasets

4. Experimental Evaluation

4.1. Dataset

To perform HAR, we restricted ourself to comparable datasets that contain at least audio data and accelerometer data which are the only sensors that are guaranteed to be found on any smartphone. We selected the LIG Smart Phone Human Activity Dataset (LIG-SPHAD) [5] and the ExtraSensory Dataset [12] which are both publicly available, do contain continuous audio and accelerometer data and are annotated using physical human activity labels. As it can be seen from Figure 4, the two datasets are imbalanced. For instance, in the LIG-SPHAD, *jump* is the smallest distribution class (1.9%) compared to the highest one *sitting* (26.4%). For the ExtraSensory Dataset, *running* (1.9%) is a minority class while *walking* (45.91%) is by far the most frequent one. The datasets were randomly split into training set and test set for classification task.

4.2. Baseline results with the MLP

The MLP we implemented is composed of three layers as described in Section 3.2. The TensorFlow library was used to implement MLP. The experiment conducted on a workstation with 3.2Ghz CPU and 16GB RAM.

The learning results are presented in Figures 5 for the two datasets. The blue line correspond to the F1 score on the test set. On LIG-SPHAD the overall score is 68% F1 while it is about 65% on the ExtraSensory Dataset. In the beginning of the learning phase, the F1 score of minority classes are very low while the F1 score of majority classes are high. At the end of learning, the F1 of majority classes still have a high score while the F1 score of minority classes steadily increase but stay below the overall score. It shows that, as every discriminative learning that does not naturally take the imbalance class problem into account, the learning favors majority classes.

4.3. MLP learning with BLL SMOTE

The MLP was then learned using the BLL SMOTE method. In this experiment, BLL SMOTE is parametrized using a query budget limitation σ of 950, a query size of $\alpha = 50$ and a neighborhood size k of 6. Different learning tasks were carried out using either: (i) A random AL: the instances are picked up randomly from U

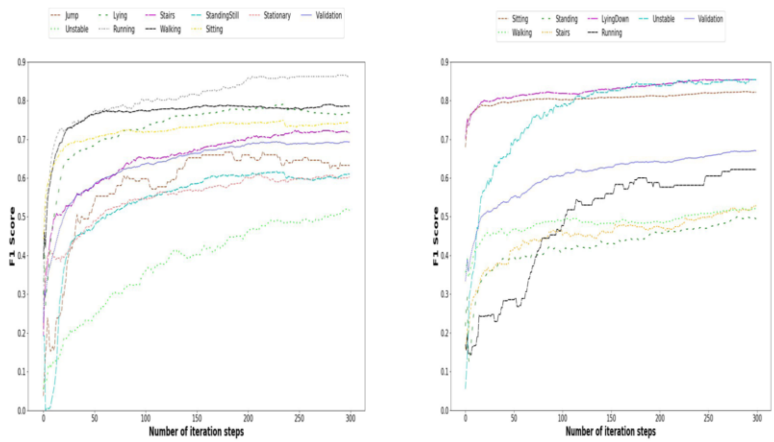

Figure 5. Baseline learning curve of the MLP on LIG-SPHAD (left) and ExtraSensory Dataset (right).

and added in S; (ii) AL without over-sampling: only the most uncertain instances with largest entropy are chosen and added to S; (iii) AL with SMOTE: the most uncertain instances are chosen, then new instances of the minority classes are created without taking (non-)convexity of the instances space into account. Then they are added to S; (iv) AL with BLL SMOTE: our method. Unless specified otherwise, the parameter values are the same for all methods.

The Figure 6 shows the F1 score curves of the four methods on the test set of the two datasets. For LIG-SPHAD on the left side of Figure 6, BLL SMOTE gave the best performances reaching 80% far better than the original 68%. For the ExtraSensory dataset, BLL SMOTE also gave the best result reaching 76%, which is better than the previous performance of 65%. However the difference wrt the other methods is less pronounced. In any case, these results show that AL and over-sampling greatly improve global performances in case of imbalanced data.

BLL SMOTE also has an effect on the classification performance of each class. On Figure 7, the MLP performance on LIG-SPHAD at the last step of active learning with over-sampling BLL SMOTE, demonstrates that the minority classes such as *Jump, Unstable* can achieve nearly 0.7 and 0.65 F1 score respectively, that is much higher than the MLP performance in the Figure 5 where F1 score are 0.6 and 0.5 respectively. On the ExtraSensor dataset, the right side of Figure 7 also illustrates that minority classes such as *Running, Stairs* can reach an F1 score of 69% and 65% respectively higher than in the right side of Figure 5, where the same classes achieved 60% and 50% F1 score respectively. Hence, BLL SMOTE makes it possible to increase minority class performance in a discriminative setting.

5. Conclusion

In this paper we introduced a generic framework which integrates active learning with over-sampling method based on MLP to overcome the class imbalance

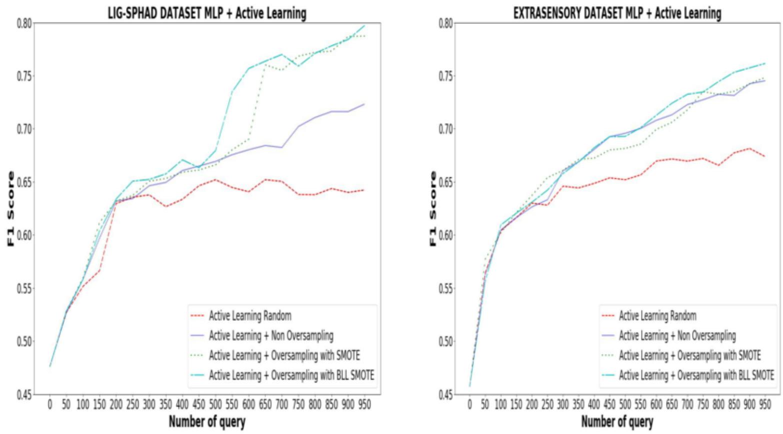

Figure 6. MLP + Active Learning on LIG-SPHAD (left) and ExtraSensory Dataset (right).

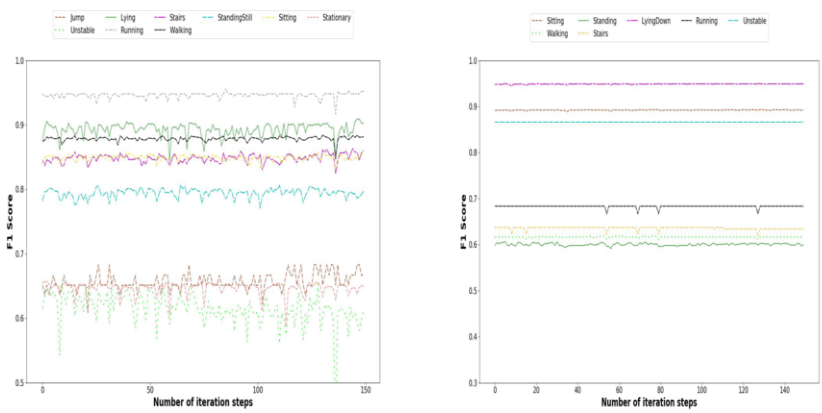

Figure 7. Last step of Multilayer Perceptron after last query of Active Learning and Over-sampling BLL SMOTE on LIG-SPHAD (left) and ExtraSensory Dataset (right)

problem. We also introduce a new over-sampling method, called BLL SMOTE - an extension of SMOTE [3] - which can be applied to non-convex spaces.

The experiments carried out on two different datasets demonstrated that using active learning with over-sampling to tackle the imbalance distribution of class can increase the global F1 score of the two datasets by about 15% absolute over the baselines. In each case BLL SMOTE shows slightly higher performances than using SMOTE plus Active learning. In addition, BLL SMOTE is able to increase the classification performance of minority classes. Another important point of this study is the fact that our method prevents the mis-generation of synthetic sample, thanks to its capacity to manage non-convex datasets.

These results show two advantages over classical approaches: the method makes it possible to improve overall and local performances and does not require extra external data. This last advantage is important in a domain such as smartphone HAR where data collection is costly and available datasets might differ too much in term of target, features or time resolution.

Acknowledgment

The authors would like to thank Dr. Mai Thai Son for valuable discussion on active learning.

References

[1] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," *IEEE Communications Surveys Tutorials*, pp. 1192–1209, 2013.

[2] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowl. and Data Eng.*, pp. 1263–1284, 2009.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Int. Res.*, pp. 321–357, 2002.

[4] O. D. Lara and M. A. Labrador, "A mobile platform for real-time human activity recognition," in *IEEE CCNC*, pp. 667–671, 2012.

[5] D. Blachon, D. Coskun, and F. Portet, "On-line context aware physical activity recognition from the accelerometer and audio sensors of smartphones," in *AmI*, pp. 205–220, 2014.

[6] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explor. Newsl.*, pp. 74–82, 2011.

[7] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," IWAAL, pp. 216–223, 2012.

[8] C. Zhu and W. Sheng, "Multi-sensor fusion for human daily activity recognition in robot-assisted living," in *HRI*, pp. 303–304, 2009.

[9] 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," *Journal of ambient intelligence and smart environments*, vol. 8, no. 4, pp. 399–422, 2016.

[10] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Procedia Computer Science*, vol. 34, pp. 450 – 457, 2014.

[11] D. Arifoglu and A. Bouchachia, "Activity recognition and abnormal behaviour detection with recurrent neural networks," in *MobiSPC*, pp. 86–93, 2017.

[12] Y. Vaizman, K. Ellis, and G. R. G. Lanckriet, "Recognizing detailed human context in the wild from smartphones and smartwatches," *IEEE Pervasive Computing*, pp. 62–74, 2017.

[13] M. B. Abidine and B. Fergani, "A new multi-class WSVM classification to imbalanced human activity dataset," *JCP*, pp. 1560–1565, 2014.

[14] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *Proceedings of the 2005 International Conference on Advances in Intelligent Computing*, pp. 878–887, 2005.

[15] S. Ertekin, J. Huang, L. Bottou, and L. Giles, "Learning on the border: Active learning in imbalanced data classification," in *CIKM*, pp. 127–136, 2007.

[16] S. Ertekin, "Adaptive oversampling for imbalanced data classification," in *ISCIS*, pp. 261–269, 2013.

[17] A. Biswas and D. W. Jacobs, "Active image clustering: Seeking constraints from humans to complement algorithms," in *CVPR*, pp. 2152–2159, 2012.

[18] S. T. Mai, I. Assent, and M. Storgaard, "AnyDBC: An Efficient Anytime Density-based Clustering Algorithm for Very Large Complex Datasets," in *KDD*, pp. 1025–1034, 2016.

[19] S. T. Mai, S. Amer-Yahia, and A. D. Chouakria, "Scable Active Temporal Constrained Clustering," in *EDBT*, 2018.

[20] B. Settles, "Active learning literature survey," tech. rep., 2010.

[21] C. E. Shannon, "A mathematical theory of communication," *Mobile Computing and Communications Review*, vol. 5, pp. 3–55, 2001.