Enhancement of Activity Recognition (AR) Performance over New and Rare Activities

Mohsen Nabian

Northeastern University
Dept. of Mechanical Engineering,
Boston,MA,USA

ABSTRACT

Activity Recognition (AR) algorithms are machine learning algorithms developed for cellphones and smart watches applications to recognize real-time human activities such as walking, standing, siting, running and biking. This paper applied several conventional classification models such as SVM on the AR dataset provided by UCI to generate reliable algorithms for real-time Activity Recognitions. However, these data are highly unbalanced toward one activity (Standing Still). The motivation is to understand how the performance of the conventional classification models (SVM, KNN, Artificial Neural Network, Decision Tree, Random Forest, and Naive Base) would be affected by these training data and whether balancing the training data by sampling will improve the performance of the classification models. Finally, the results will be compared with the newly introduced FE-AT method. It is shown that by training 'Random Forest' model with balanced dataset, more accurate prediction will be obtained comparing to FE-AT.

Key words:

Activity Recognition; Attribute-based Learning; Insufficient Data Problem; Imbalanced Data Problem.

INTRODUCTION

Wearable-based activity recognition (AR) systems are typically built to recognize a predefined set of common activities such as sitting, walking, and running [1]. However, to adapt to the needs of individuals and application scenarios, these AR systems often need to be extended to recognize new activities of interest. For example, people working out at a gym need the AR system to correctly distinguish between individual types of exercises, whereas applications helping users quit smoking depend on the system's ability to recognize smoking activities.

The amount of Data available for the new activities (like biking) are significantly lower than the common activities such as "Standing still". Activities like 'Standing still' are: 1) Very common and are the dominant daily activity of human 2) Easy to generate data in this position. 3) being measured since a while ago and decent amount of data is available for these activities as opposed to rare and new activities that have just received attention.

To learn new activities of interest, AR systems can ask users to label additional training data. However, it is impractical to assume that users will provide a large amount of annotations, since labeling activities is a time-consuming and laborsome process [3-7]. Therefore, being able to learn new activ- ities with a limited amount of training data is highly desirable for practical AR systems.

Training machine learning models over these unbalanced data will cause inaccurate predication of the test data. For example since the 'standing still' data are 10 times more than the 'biking' data in our training dataset, the machine learning model will likely predict the real 'biking' activity as 'standing still' activity.

This paper is intended to obtain sufficiently accurate classifier for AR problem that is comparable

to the newly introduced FE-AT model. The following nonlinear classifiers have been trained for the AR prediction: 1) SVM 2) KNN 3) Artificial Neural Network 4) Decision Tree 5) Random Forest, and 6) Naive Base. These models have been trained with two sets of training data: 1) **Method I**: the **original datasets** provided by UCI**. 2) Method II:** Reduced version of the unbalanced original data for more uniform training dataset. It is shown that some models learned by method II are not only more predictive than models obtained from method I, but also they demonstrate more accurate prediction over the FE-AR [2] method for AR problems.

For example, instead of having training data, composed of 90% allocated to 'standing still', it is given 70% standing still data thus we have increased the role of other rare activities in our model. It is shown that, this method will improve the overall accuracy of the models comparing to the models obtained from the biased training data.

1.PROBLEM EXPLANATION

Highly Non-uniform and biased training data may result in in-accurate and biased prediction models. Activity Recognition datasets are mostly non-uniform and biased toward the common activities that both comprises high percentage of the daily activities and are more convenient for data collection. The main motivation of this paper is to achieve more accurate models for non-uniform AR datasets.

The idea of this paper are as followings:

1) study the robustness of different machine learning algorithms against the highly non-uniform AR training data.

2) Modification of the training set based on the performance of the classifiers: The ratio of the dimension of the data for the dominant activity (standing still) will be decreased to minimize a pre-defined cost function $J(\alpha)$ that will be defined later in this paper. For instance, the original dataset contains 90% of very common activities and only 10% of the rest of the activities. By modifying this ratio (a), (like 70% common activities), the trained model may be significantly more accurate toward the rare activities while not loosing much of its accuracy toward common activities and finally will reduced the cost function $J(\alpha)$.

3) The proposed learning models will be compared in terms of accuracy to the current AR algorithms.

2. DATA DESCRIPTION:

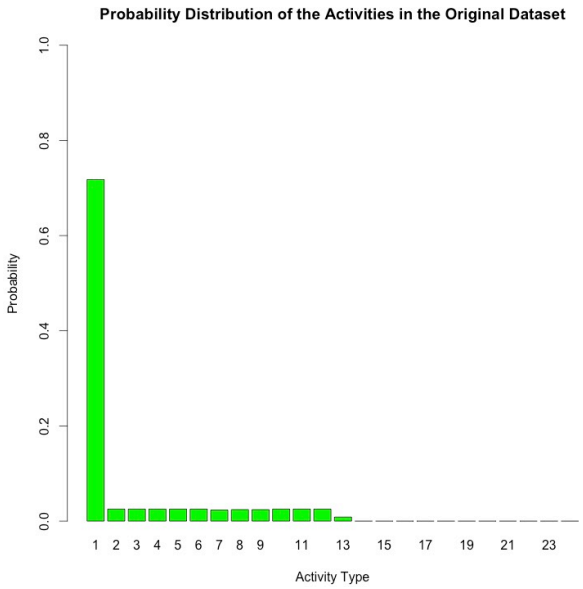
The dataset comprises body motion and vital signs recordings for ten volunteers of diverse profile while performing several physical activities. Sensors placed on the subject's chest, right wrist and left ankle are used to measure the motion experienced by diverse body parts, namely, acceleration, rate of turn and magnetic field orientation. The sensor positioned on the chest also provides 2-lead ECG measurements, which can be potentially used for basic heart monitoring, checking for various arrhythmias or looking at the effects of exercise on the ECG.

The meaning of each attribute is detailed below:

Column 1: acceleration from the chest sensor (X axis)
Column 2: acceleration from the chest sensor (Y axis)
Column 3: acceleration from the chest sensor (Z axis)
Column 4: electrocardiogram signal (lead 1)
Column 5: electrocardiogram signal (lead 2)
Column 6: acceleration from the left-ankle sensor (X axis)
Column 7: acceleration from the left-ankle sensor (Y axis)
Column 8: acceleration from the left-ankle sensor (Z axis)
Column 9: gyro from the left-ankle sensor (X axis)
Column 10: gyro from the left-ankle sensor (Y axis)
Column 11: gyro from the left-ankle sensor (Z axis)
Column 13: magnetometer from the left-ankle sensor (X axis)
Column 13: magnetometer from the left-ankle sensor (Y axis)
Column 14: magnetometer from the left-ankle sensor (Z axis)
Column 15: acceleration from the right-lower-arm sensor (X axis)
Column 16: acceleration from the right-lower-arm sensor (Y axis)
Column 17: acceleration from the right-lower-arm sensor (Z axis)
Column 18: gyro from the right-lower-arm sensor (X axis)

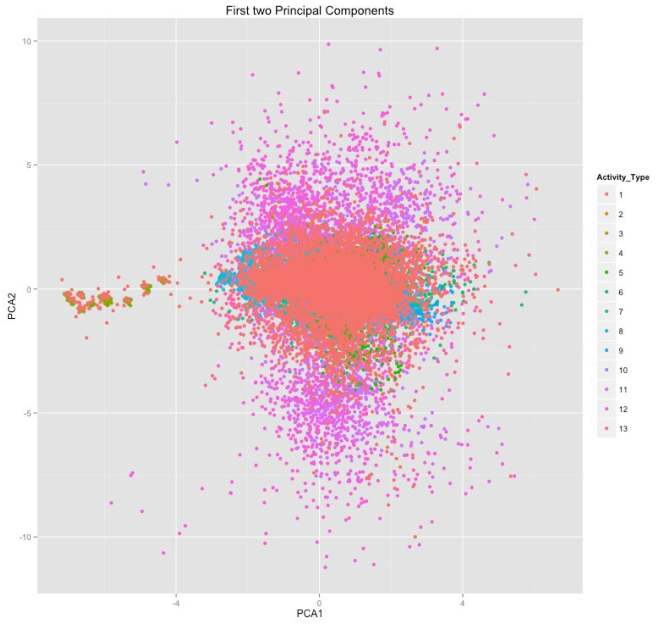
Column 19: gyro from the right-lower-arm sensor (Y axis)
Column 20: gyro from the right-lower-arm sensor (Z axis)
Column 21: magnetometer from the right-lower-arm sensor (X axis)
Column 22: magnetometer from the right-lower-arm sensor (Y axis)
Column 23: magnetometer from the right-lower-arm sensor (Z axis)
Column 24: Label (0 for the null class)
The activity set is listed in the following:
L1: Standing still (1 min)
L2: Sitting and relaxing (1 min)
L3: Lying down (1 min)
L4: Walking (1 min)
L5: Climbing stairs (1 min)
L6: Waist bends forward (20x)
L7: Frontal elevation of arms (20x)
L8: Knees bending (crouching) (20x)
L9: Cycling (1 min)
L10: Jogging (1 min)
L11: Running (1 min)
L12: Jump front & back (20x)
NOTE: In brackets are the number of repetitions (Nx) or the duration of the exercises (min).

The existing training data is highly balanced toward 'standing still' activity composing 70% of all the data alone. 25% of the data composed almost uniformly of 12 other activities. The remaining 5% of the data is almost uniformly distributed among the 12 remaining activities. Therefore, the data is highly non-uniform. The following figure demonstrate the distribution of the data with respect to the activity type:



3.DATA EXPLANATORY ANALYSIS:

In order to obtain more clear and deeper understanding of the data, the data is visualized using 'principal component analysis' (PCA). The following figure demonstrates the data labeled in the plane of the the first two principal components.

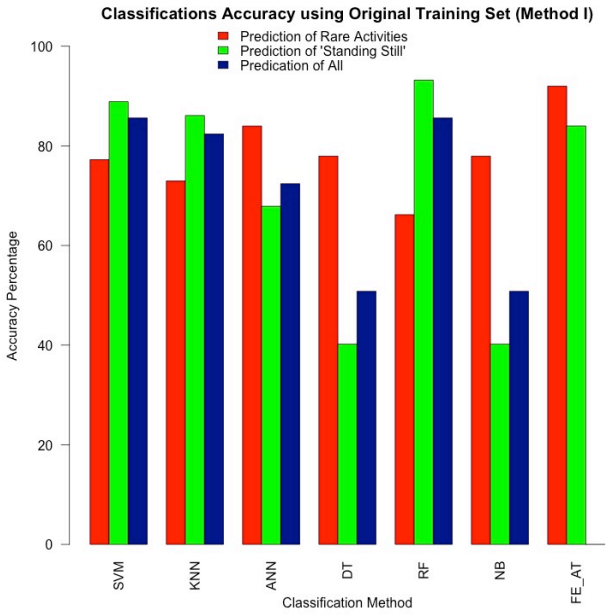


This figure indicates how the 'standing still' activity is accumulated around the center of the plane of the first two principal components.

4.CLASSIFACATION

METHOD I

In the 1st method, the unbiased data provided by UCI is used as the training set for the pre-mentioned conventional classification models. The following figure demonstrates the prediction result of the models. The red bars represent the prediction of the rare activities comprising 23 activities while containing only 30% of the training data. The green bars show the prediction accuracy of 'Standing Still' data which has 70% of the training data allocated to it. Finally, the black bars demonstrate the predication of the whole test data.



The figure implies that the models are unable to predict the rare activities sufficiently accurate. While SVM and Artificial Neural Network (ANN) have prediction accuracy over rare activities around 75%, Random Forest(RF) shows significantly lower accuracy.

Overall, method I is not satisfactory for prediction of rare activities, however, theses models presents high accuracy over the 'Standing Still' and even higher than FE_AT. The reasons is that the models are highly trained for the 'Standing Still' activity and less trained for other activities.

METHOD II

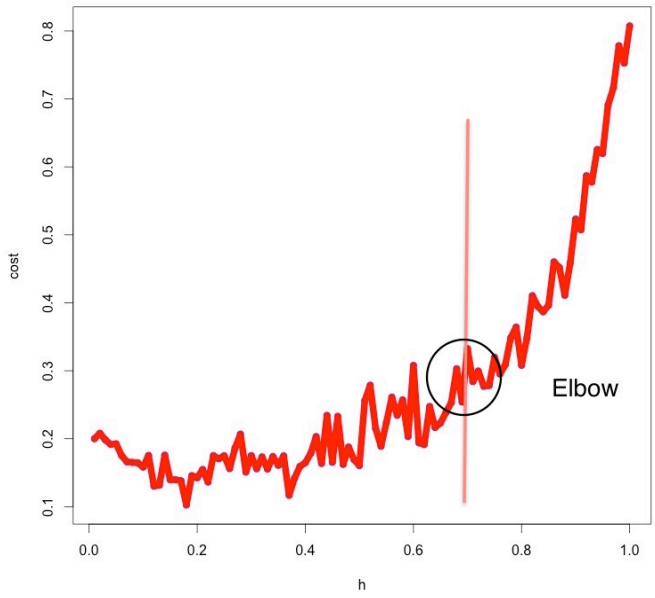
Opting the size ratio (1-h) of the training data allocated to rare activities needed for acceptable learned model is controversial.

Here, a cost function J is defined to rationally penalize the inaccuracy prediction of both 'Standing Still' (dominant) activity and the the rare activities.

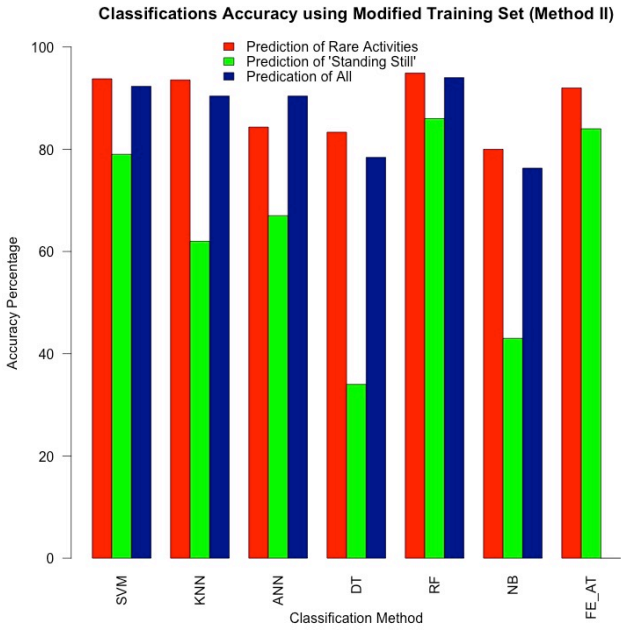
$$J = 0.8 * (False, Dominant) + 0.2 * (False, Rare)$$

The defined cost function, penalizes the inaccuracy predication of dominant activity 4 times more than the rare activities. The reason is that the rare activity is indeed occurs most of the time. Therefore, it is quite important to modify the training set without so much affecting the prediction accuracy of the dominant activity. The objective is to opt h as higher as possible while minimizing the cost function J.

The next figure, demonstrates that there is an elbow shape increase of cost function starting around h=0.7. As a result, h=0.7 is chosen for the modification of the training set in method II.



The next step is to feed the leaning models with the method II of the training set. The results are shown in the following figure.



In method II , the predictions of the rare activities have been significantly improved. This is due to the fact that higher portion of training data have been assigned to the learning of rare activities. However, there exists a contained decrease in the accuracy of the dominant activity comparing the models achieved in method I.

Support Vector Machine has proved a great classifier in method II, however, Random Forest is the presenting the highest accuracy in both Rare activities and dominant activity. It is also shown that Random Forest has provided more predictive model than the FE-AT model.

5.SUMMARY AND CONCLUSIONS

In this research the Activity Recognition (AR) classification problem with the highly non-uniform training set has been investigated. The conventional classifier models including SVM, KNN, Artificial Neural Network, Decision Tree, Random Forest, and Naive Base have shown relatively low accuracy in the prediction of rare activities given the non-uniform training data set. However, by modifying the training dataset (balancing the data toward the rare activities) it is shown a significant improvement in all classifier models. Moreover, Random Forest model has shown even more accurate prediction than the top notch FE-AT model that is proposed recently.

6.SIMILAR WORKS

FE-AT [2] (Feature-based and Attribute-based learning) approach for Activity Recognition problems, has been introduced which leverages the relationship between existing and new activities by adding semantic analysis of the activities to compensate for the shortage of the labeled data. FE-AT on three public datasets and has demonstrated that it outperforms traditional AR approaches in recognizing new activities, especially when only a few training instances are available. This research has shown that Random Forest with modified data (Method II) can be as much and slightly better in performance as FE-AT.

ACKNOWLEDGMENTS

We would like to acknowledge Prof. Nik Bear Brown for his support and encouragement for this research.

REFERENCES

1. Lara, O. D., and Labrador, M. A. A survey on human activity recognition using wearable sensors. Communications Surveys & Tutorials, IEEE 15, 3 (2013), 1192–1209.
2. Nguyen, L. T., Zeng, M., Tague, P., Zhang, J. (2015). Recognizing New Activities with Limited Training Data. In IEEE International Symposium on Wearable Computers (ISWC).
3. Stikic, M., Van Laerhoven, K., and Schiele, B. Exploring semi-supervised and active learning for activity recognition. In IEEE International Symposium on Wearable Computers, IEEE (2008), 81–88.
4. Stikic, M., and Schiele, B. Activity recognition from sparsely labeled data using

multi-instance learning. In Location and Context Awareness. Springer, 2009, 156–173.

5. Nguyen, L. T., Zeng, M., Tague, P., and Zhang, J. Superad: Supervised activity discovery. In International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, ACM (2015).
6. Banos, O., Garcia, R., Holgado, J. A., Damas, M., Pomares, H., Rojas, I., Saez, A., Villalonga, C. mHealthDroid: a novel framework for agile development of mobile health applications. Proceedings of the 6th International Work-conference on Ambient Assisted Living an Active Ageing (IWAAL 2014), Belfast, Northern Ireland, December 2-5, (2014).

7. Banos, O., Villalonga, C., Garcia, R., Saez, A., Damas, M., Holgado, J. A., Lee, S., Pomares, H., Rojas, I. Design, implementation and validation of a novel open framework for agile development of mobile health applications. BioMedical Engineering OnLine, vol. 14, no. S2:S6, pp. 1-20 (2015).