Classification of Daily Activities based on Accelerometer data over Time

Fernando M. Wittmann, 180577
Anderson Rocha

***Abstract*—This report presents the classification of daily activities based on accelerometer data over time. The classification method used was random forest with 50 trees. Five classification results are presented. The first two uses feature vectors with sizes 100 and 200 measures. Third result uses frequency features. Fourth result uses engineered features, but with unbalanced data. And finally, the last result uses both balanced data and engineered features. The results significantly increased from the base case, using time features, to the final result. A relevant result was the average of the main diagonal of the confusion matrix that increased from 52% to 90% in the last case.**

I. Introduction

Section II presents the required activities. Section III some feature's engineering strategies that are used in Sec. IV.C. Next, section IV presents the results, using time, frequency and engineered features. Section V presents the conclusions and finally the Appendix presents some images that were helpful to get insights from data.

II. Activities

The required activities can be summarized in the list below:

1) Create feature vectors with different sizes (number of accelerometer readings).
   Item A and B of Sec. IV provides results using 100 and 200 measures
2) Perform an initial classification based on the feature vectors created using the numbers directly.
   Item A of Sec. IV
3) Perform an initial classification based on the feature fecturs created but transformed to the frequency domain.
   Item B of Sec. IV
4) Try to solve the problem by engineering some features.
   Item A of Sec. III for strategies & Item C of Sec. IV for results
5) Improve the classification model tackling the data imbalance.
   Item B of Sec. III for strategies & Item C of Sec. IV for results
6) All experiments must be done in a 2-fold cross validation procedure.
   Results in Sec. IV

Contact: fernando.wittmann@gmail.com
Contact: anderson.rocha@ic.unicamp.br

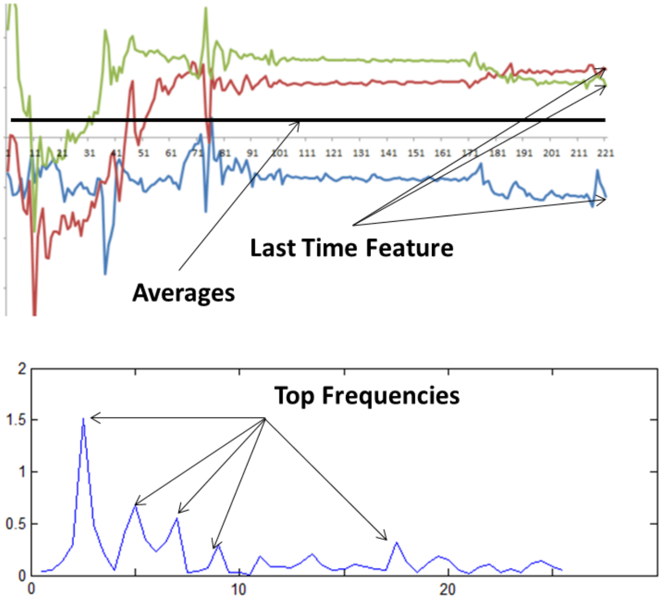


Fig. 1. Informative features that were used for feature engineering. Top frequencies to identity cyclical activities and time attributes to identify not cyclical.

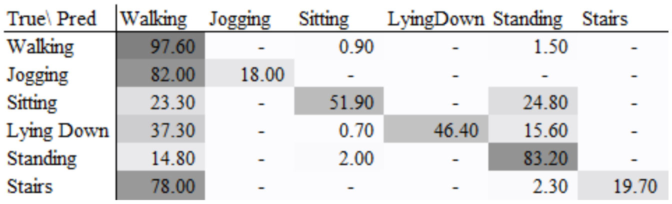
III. Features Engineering

This section presents the strategies used to get better features. The results of these strategies are presented in the Section IV.C.

A. (Act. 4) More Informative Features

After a data analysis, it was observed that frequency could be helpful for identifying cyclical activities. For example the walking activity had a predominant frequency as 1Hz (one step per second), while jogging was almost 2Hz. Also, it was observed that steady activities tends to constant values in the end. A better classifies would be one that combines information to identify cyclical and not cyclical activities. With this intent, three kind of information were taken from the data. Two to identify steady activities and one, in the frequency to identify cyclical. The following list presents these three information:

- Last time feature of X, Y and Z: The time feature of a time series provide useful information to identify not cyclical activities (like standing or sitting) as they always tend to steady values (see Fig. 1).



True\ Pred	Walking	Jogging	Sitting	LyingDown	Standing	Stairs
Walking	97.60	-	0.90	-	1.50	-
Jogging	82.00	18.00	-	-	-	-
Sitting	23.30	-	51.90	-	24.80	-
Lying Down	37.30	-	0.70	46.40	15.60	-
Standing	14.80	-	2.00	-	83.20	-
Stairs	78.00	-	-	-	2.30	19.70

Fig. 2. Confusion matrix for 100 time measures. The second confusion matrix had as maximum changes 3%. Due to the imbalance, mostly activities are being misclassified as walking. Score: 81%. Avg of main diagonal: 52%.

- Average of time features X, Y and Z: The average of time series also is a good predictor as each activity has different averages. Three averages were taken, one for each direction feature (X, Y, Z).
- Top frequencies: Getting the top frequencies (which are dominant) are useful to identify cyclical activities. For this activity it was taken the top 5 frequencies of each transformation.

Fig. 1 summarizes the informative features that were used.

B. (Act. 5) Tackling Imbalance

The strategy adopted to tackle imbalance was to use different offset values for different activities. The test set kept unchanged. The following offset values were used for window size of 100 measures:

- Walking (42.1%): no offset
- Jogging (14.7%): offset of 50 measures
- Stairs (1.9%): offset of 6 measures
- Sitting (22.3%: 50 measures
- Standing (9.7%): 30 measures
- Lying Down (9.3%): 30 measures

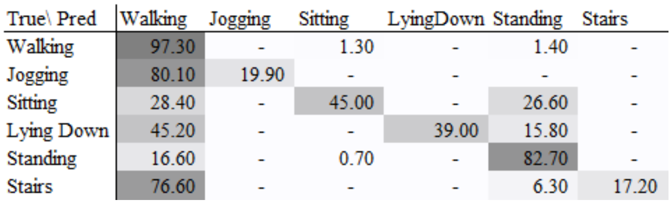
IV. Classification Results

The classifier used in this activity was random forest in Python platform. As parameters, it was used 50 threes in the forest. The validation method was 2-fold cross-validation. Only the first confusion matrix is presented and some comments about the second.

A. (Act. 1 & 2) Time based

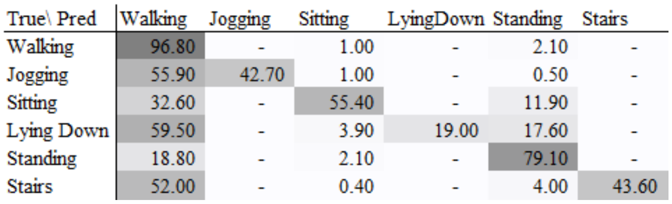
Using a window of 100 measures and overlap of 50 measures, the best score obtained in random forest was 81%. Score is a harsh metric since it is required for each sample that each label set be correctly predicted. A better metric is the confusion matrix, presented in Fig. 2. From the confusion matrix is is possible to observe that mostly classes are being classified as Walking. The average of the main diagonal is 52% which is a low accuracy value.

Using 200 measures the results did not change very much. The score was 80%. The confusion matrix is in the Fig. 3 with average of the main diagonal as 50.2%.



True\ Pred	Walking	Jogging	Sitting	LyingDown	Standing	Stairs
Walking	97.30	-	1.30	-	1.40	-
Jogging	80.10	19.90	-	-	-	-
Sitting	28.40	-	45.00	-	26.60	-
Lying Down	45.20	-	-	39.00	15.80	-
Standing	16.60	-	0.70	-	82.70	-
Stairs	76.60	-	-	-	6.30	17.20

Fig. 3. Confusion matrix for 200 time measures. Results are not very different from 100 measures. Score: 80%. Avg of main diagonal: 50.2%.



True\ Pred	Walking	Jogging	Sitting	LyingDown	Standing	Stairs
Walking	96.80	-	1.00	-	2.10	-
Jogging	55.90	42.70	1.00	-	0.50	-
Sitting	32.60	-	55.40	-	11.90	-
Lying Down	59.50	-	3.90	19.00	17.60	-
Standing	18.80	-	2.10	-	79.10	-
Stairs	52.00	-	0.40	-	4.00	43.60

Fig. 4. Confusion matrix using frequency features for 100 readings. Score: 81.8%. Avg of diagonal: 56%. Results significantly improved for some cyclical activities (jogging and stairs) and worsened for Lying Down (not cyclical). The maximum change in the second confusion matrix was 3.3% and average 0.5%.

B. (Act. 1 & 3) Frequency based

Slightly better results were observed using frequency features. Frequency features are better to identify cyclical activities. Fig. 4 shows one of the confusion matrices. Still it is possible to see the effect of unbalance in the walking activity. Comparing cyclical activities from previous result, we can see that increased jogging from 20% to 43% and Stairs from 17% to 44%. However the activity Lying Down, which is not cyclical, worsened from about 42% to 19%. The score slightly improved from 79% to 82%. The average of the main diagonal slightly improved to 56.1%.

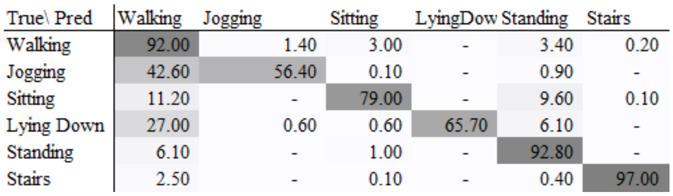
C. (Act. 4 & 5) With Engineered Features

- Informative Features

As discussed in the section III.B, it was taken three kind of informative features in order to improve the classification. One of them from frequency spectrum and two from time spectrum. The frequency spectrum was used to identify cyclical activities and the time for not-cyclical activities. Fig. 5 shows the confusion matrix. The data still was not balanced for this test. The prediction of Stairs increased to 97% which is a very significant improvement (previous result is 44%). Score is 87.9% and the diagonal average is 80.48%. The second confusion matrix was very similar with maximum difference of 4% and score of 87%.

- With balanced Data

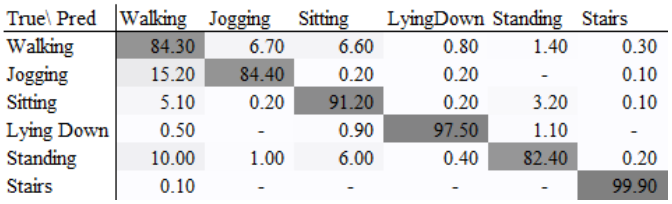
Fig. 6 shows one of the confusion matrices using balanced data. When comparing with the previous solution, it is possible to observe that many of the features that were previously classified as Walking (the most frequent set), now are being correctly classified. As a side effect, the correctly classification of Walking activity decreased from more that 90% to 84%.



True\ Pred	Walking	Jogging	Sitting	LyingDow	Standing	Stairs
Walking	92.00	1.40	3.00	-	3.40	0.20
Jogging	42.60	56.40	0.10	-	0.90	-
Sitting	11.20	-	79.00	-	9.60	0.10
Lying Down	27.00	0.60	0.60	65.70	6.10	-
Standing	6.10	-	1.00	-	92.80	-
Stairs	2.50	-	0.10	-	0.40	97.00

Fig. 5. Confusion matrix using more informative features but with unbalanced data. Score: 88%. Avg of main diagonal: 80%. Results highly increased for Stairs, Sitting and Lying Down.

The best score was 89.5% and the average of main diagonal is 89.95%. When comparing with the second confusion matrix, maximum difference is 2.5% and average difference is 0.3%.



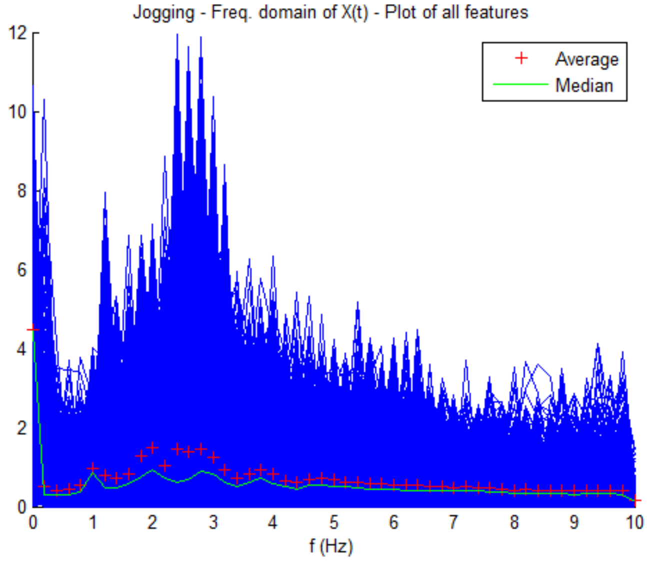
True\ Pred	Walking	Jogging	Sitting	LyingDown	Standing	Stairs
Walking	84.30	6.70	6.60	0.80	1.40	0.30
Jogging	15.20	84.40	0.20	0.20	-	0.10
Sitting	5.10	0.20	91.20	0.20	3.20	0.10
Lying Down	0.50	-	0.90	97.50	1.10	-
Standing	10.00	1.00	6.00	0.40	82.40	0.20
Stairs	0.10	-	-	-	-	99.90

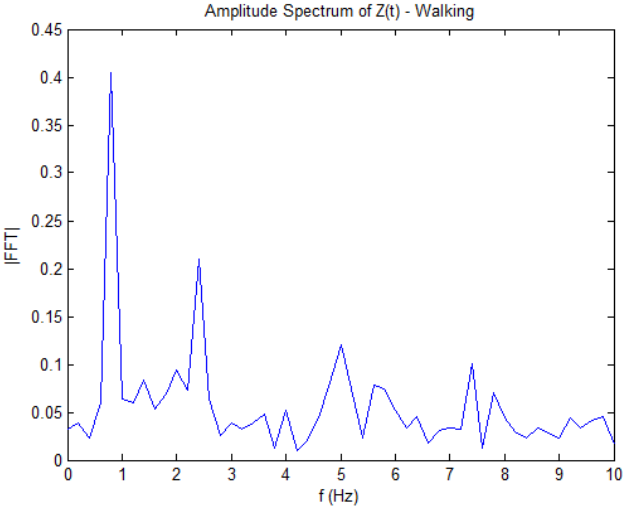
Fig. 6. Confusion matrix for balanced data and engineered features. Score: 89.5%. Avg of main diagonal: 90%. With balanced data, the overall results improved. As a side effect, the accuracy for Walking activity decreased.

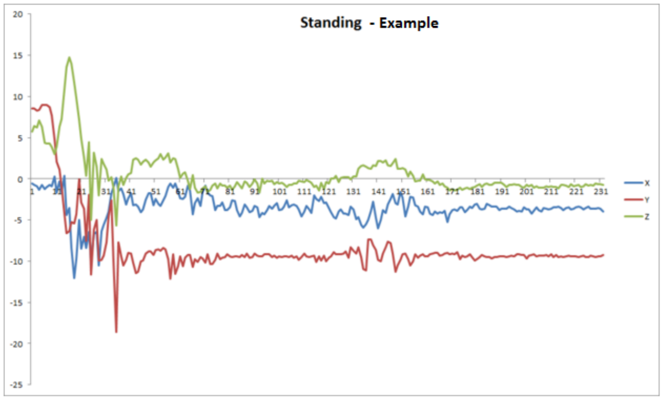
V. Conclusions

This report presented the use of different strategies to classify daily activities based on mobile accelerometer's data. The best results were got when using balanced data with informative features. The best case used different offsets for balancing and three kind of information. The average of the main diagonal of the confusion matrix increased from 52% to 90 %. The two confusion matrices have as maximum 3% of difference which imply the accuracy won't change very much.

VI. Appendix - Additional Images







References

[1] MO 444 Class Notes

[2] "Random Forest for Astronomy", Jake VanderPlas, ESAC Data Analysis and Statistics Workshop 2014

[3] ”Classifier Comparison”, Scikit-learn 0.16.1 Documentation. Web. 04 Nov. 2015.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.