Human Activity Recognition Analysis

Chiara Todaro

12 March 2019

Synopsis

In this report, a Conditional Inference Tree Classification algorithm is performed on data recorded from accelerometers worn by participants performing physical activity. The in-sample accuracy is 94%, while the estimated out-sample accuracy is ~91%. The probability of correct prediction on 20 new samples is 0.95 on average, with a standard deviation of 0.13.

Analysis

The Conditional Inference Tree Classification algorithm has been chosen because it reaches a good trade-off between accuracy and computational cost (for a complete description of the method see Hothorn, Hornik and Zeileis (2006)[1]). Briefly, this algorithm works like a classification tree in which the variable selection and the subsequent splitting procedure is ruled by χ^2 hypothesis tests between a nominal response and one of the covariates: the one with the highest association is selected for splitting. The entire analysis has been performed in R, and the implementation of the Conditional Inference Tree Classification algorithm is taken from the package *party* (https://cran.r-project.org/web/packages/party/party.pdf).

Data loading and cleaning

Data are taken from the Human Activity Recognition database (http://groupware.les.inf.puc-rio.br/har) and consist in the recordings by accelerometers worn by six participants performing 10 repetitions of the Unilateral Dumbbell Biceps Curl in five different fashions:

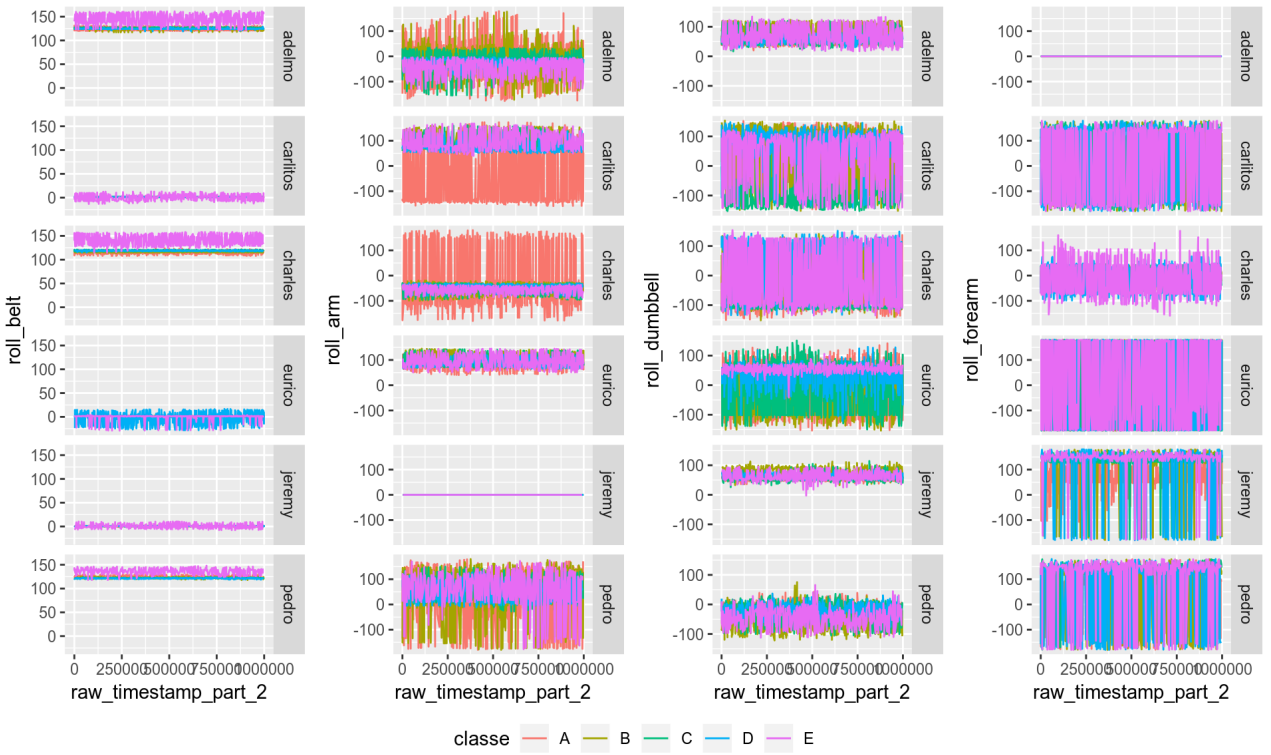
- *A* exactly according to the specification,
- *B* throwing the elbows to the front,
- *C* lifting the dumbbell only halfway,
- *D* lowering the dumbbell only halfway
- *E* throwing the hips to the front.

The goal of the experiment is to predict the way in which a participant performed the exercise. For more information on data acquisition and original analysis see Ugolino et. al (2012)[2].

Since 61.2068087% of the total elements are *NA*, only 60 out of 160 variables are kept, for both the training and testing set. The training and testing set have 19622 and 20 observations, respectively.

Exploratory analysis

The remaining variables consist of measures, e.g. roll, pitch, yaw, total acceleration, acceleration, gyros, magnet, taken from different body parts, e.g. belt, arm, dumbbell, and forearm. As an example, Figure 1 shows the roll of all body parts for each participant: the recorded signals are coloured accordingly to the variable "*classe*" which represent the type of movement to predict. For example, it can be seen that for the belt the signals related to class *A* are more discernable from the other type of exercise.



Raw data: roll signals

Cross-validation

Cross-validation is performed by splitting into k-folds (k=15) the training set. Subsets of features are choosen combining variables relative to different body parts. In the table shown below the presence of features relative to a specific body part is represented by "1", while the absence is represented by "0". For each selection the in-sample accuracy (*accuracyTrain*) and out-sample accuracy (*accuracyTest*) are reported.

```
##    belt arm dumbell forearm accuracyTrain accuracyTest
## 1     0   0       0       1     0.8695462    0.8143621
## 2     0   0       1       0     0.8396396    0.8071920
## 3     0   0       1       1     0.9014907    0.8586707
## 4     0   1       0       0     0.8044998    0.7442748
## 5     0   1       0       1     0.8954352    0.8188073
## 6     0   1       1       0     0.8869655    0.8456837
## 7     0   1       1       1     0.9188643    0.8645754
## 8     1   0       0       0     0.8723451    0.8247896
## 9     1   0       0       1     0.9158567    0.8662080
## 10    1   0       1       0     0.9136773    0.8446825
## 11    1   0       1       1     0.9420084    0.9014515
## 12    1   1       0       0     0.8809173    0.8339709
## 13    1   1       0       1     0.9218588    0.8747135
## 14    1   1       1       0     0.9296751    0.8959449
## 15    1   1       1       1     0.9417900    0.9083270
```

The most accurate prediction is obtained combining all variables (row 15) with in-sample and out-sample accuracy of 94.1789985% and 90.8326967%, respectively. Since in the cross-validation procedure the out-sample accuracy is 3.3463018% lower than the in-sample accuracy, similar performances are expected in the testing set.

Training set accuracy

The percentages of predictions (rows) vs the actual *classe* (columns) are shown below.

```
##
## predTREE           A           B           C           D           E
##        A 27.61696055  0.55549893  0.11721537  0.25991234  0.07644481
##        B  0.47905412 17.81164000  0.63704006  0.30068291  0.31597187
##        C  0.09173377  0.44847620 16.30312914  0.58607685  0.20385282
##        D  0.19875650  0.29049027  0.25991234 15.09530119  0.20385282
##        E  0.05096320  0.24462338  0.12231169  0.14779329 17.58230558
```

The in-sample accuracy of the whole training set is 94.4093365% and the average probability of correct prediction is 0.9440934 with standard deviation of 0.1208337. The expected out-sample accuracy is around 90.6023588%.

Prediction of testing set

The predictions on the test data set are

B, A, B, A, A, E, C, E, A, A, B, C, B, A, E, E, A, B, B, B

with probability of correct prediction of

0.9183673, 1, 0.9756098, 0.9879518, 1, 0.9850746, 0.5208333, 0.9375, 1, 1, 0.625, 0.9684543, 1, 1,

1, 0.9952153, 0.996129, 1, 0.9473684, 1

As expected, only 90% of the predictions are likely to be correct.

---

1. Torsten Hothorn, Kurt Hornik and Achim Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651–674. Preprint.pdf (http://statmath.wu-wien.ac.at/~zeileis/papers/Hothorn+Hornik+Zeileis-2006.pdf)↩
2. Ugulino, W.; Cardador, D.; Vega, K.; Velloso, E.; Milidiu, R.; Fuks, H. Wearable Computing: Accelerometers' Data Classification of Body Postures and Movements. Proceedings of 21st Brazilian Symposium on Artificial Intelligence. Advances in Artificial Intelligence - SBIA 2012. In: Lecture Notes in Computer Science. , pp. 52-61. Curitiba, PR: Springer Berlin / Heidelberg, 2012. ISBN 978-3-642-34458-9. DOI: 10.1007/978-3-642-34459-6_6.↩