Classification of exercise acivity using the weight lifting dataset from

zchen

February 27, 2016

Summary

The goal of this analysis is to model and predict the quality of execution of weight lifting exercises using data from accelerometers attached to the human subject's body and dumbbell. First the dataset was filtered by by removing the variables consisting mostly of missing values and/or with zero or near zero variance. Then the training set was splited into internal training and testing sets (70:30 ratio). The internal training set was used to train and built a list of models, including linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), tree based model (RPART), random forest (RF), tree based boosting (GBM), model based boosting (MBOOST) using the Caret package with 5 fold cross validation method (no repeats). The performance of these models was evaluated using the internal testing dataset. It shows that the random forest model is the best performing model, giving a prediction accuracy of over 99%, thus this modle was used to predict on the external testing data with 20 observations.

Background and Dataset

Human Activity Recognition (HAR) has gained increasing attention by the computing research community in recent years, due to the development of context-aware systems. There are many potential applications for HAR, e.g., elderly monitoring, life log systems for monitoring energy expenditure and for supporting weight-loss programs, and digital assistants for weight lifting exercises, etc.

The data for this project come from this source: http://groupware.les.inf.puc-rio.br/har. Six young health participants were asked to perform one set of 10 repetitions of the Unilateral Dumbbell Biceps Curl in five different fashions:

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

In this project, the goal is to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants, to predict the five classes of activities.

The training data and test data for this project are downloaded from:
https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv
https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

More information about the data can be found at: http://groupware.les.inf.puc-rio.br/har.

Load data and exploratory analysis

The training and testing datasets were downloaded from above links. The training data set has 19622 observations and 160 columns, and the testing data set has 20 obersvations and same number of columns.

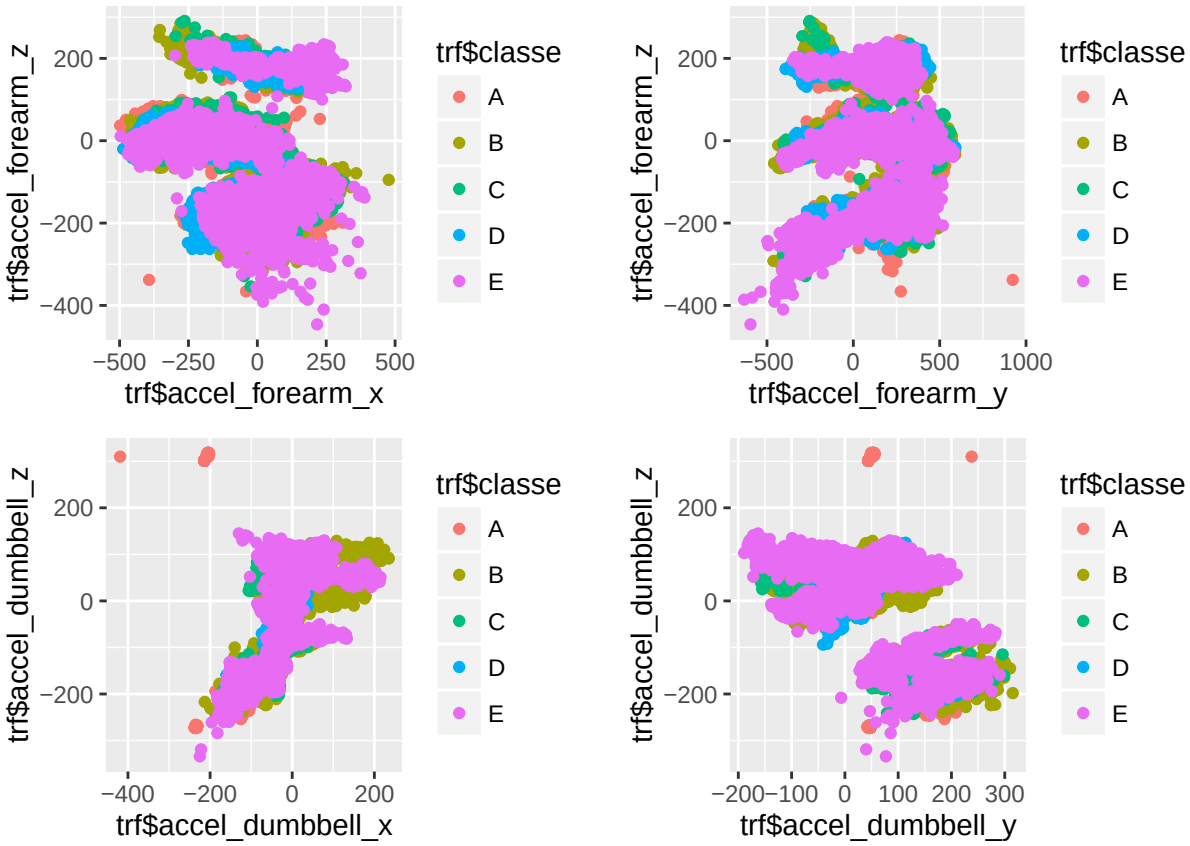
```
## [1] 19622   160
```

```
## [1]  20 160
```

First, the training set was filtered by dropping zero or near zero variance variables using nearZeroVar function from the Caret package, and columns with largely missing values (NA) are also removed. After filtering, the remaining training set has 19622 observations and 53 variables.

```
## [1] 19622    53
```

Scatter plots between several measurements, colored based on different classes



Data preprocessing and modelling

The training dataset was splitted into an interanl training set and internal testing set (ratio 70:30), so we can build the models using the same internal training set and measure the performance of all the models using same internal testing set, thus we can have an unbiased way to evaluate the models.

We tested a variety of models, including linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), tree based model (RPART), random forest (RF), tree based boosting (GBM), model based boosting (MBOOST) using the Caret package with 5 fold cross validation method (no repeats). The performance of these models was evaluated using the same internal testing set.

Model performance evaluation

To estimate the accuracy and the out of sample error of the models, the internal testing dataset with was used. The observations in this set were not used in model generation and all have a known classe assignment, which was used for comparison with the predicted one.

```
##                      LDA        QDA       RPART        GBM      RF
## Accuracy       7.0739e-01 8.9346e-01  4.8972e-01 9.6194e-01 0.99405
## Kappa          6.2980e-01 8.6551e-01  3.3306e-01 9.5186e-01 0.99248
## AccuracyLower  6.9558e-01 8.8529e-01  4.7687e-01 9.5673e-01 0.99174
## AccuracyUpper  7.1900e-01 9.0123e-01  5.0258e-01 9.6668e-01 0.99585
## AccuracyNull   2.8445e-01 2.8445e-01  2.8445e-01 2.8445e-01 0.28445
## AccuracyPValue 0.0000e+00 0.0000e+00 1.6016e-241 0.0000e+00 0.00000
## McnemarPValue  2.3582e-59 1.0386e-50         NaN 7.6393e-06     NaN
```

The result show that RF model is the best performing model.

predict the external testing set with 20 observations using RF model

```
##  [1] B A B A A E D B A A B C B A E E A B B B
## Levels: A B C D E
```

Conclusions

A list of models were built using the Caret package to predict different classies of activities, the results show that the best performing model is random forest model, which achieves over 99% of accuracy with internal testing set and 100% accuracy on the external testing data set.