Title: Determine manner in which exercise is performed based on Machine Learning Technique

Executive Summary

Computer devices (such as Nike Fueland, Fitbit and even smart phone) the data collection about personal activity relatively inexpensively. Studies (from <http://groupware.les.inf.puc-rio.br/har>) were performed asking 6 participants to lift the dumbbell correctly and incorrectly in 5 different ways (hereafter, known as class). Data were obtained from accelerometers on the belt, forearm, arm, and dumbbell of the 6 participants. Two data set, a large training set and a testing set containing 20 observations were provided.

This paper describe the process used to select the random forest as the machine learning method through different training methods. The random forest method thus obtained accurately predict the exercise class for all the 20 cases.

Data Processing

Raw data analysis and processing

The training set contains 160 columns and 19622 rows. The classe column is a factor variable containing 5 different type as follows.

* A : correctly (per specification)
* B: throwing elbow to front
* C: lifting dumbbell half way
* D: Lowering dumbbell half way
* E: Throwing hip to front.

Steps are required to refine the dataset variable to a more manageable size. Four different types of variables (columns) are of no value in selecting the machine learning methods and will be eliminated. They are

1. Descriptive, non-relevant variables - e.g. user name, timestamps, number of windows.
2. Variables with missing no data – e.g. kurtosis, skewness data
3. Variable with NA data – e.g. standard deviation, variance, average, maximum and minimum data
4. Variable that are highly correlated – (using more of these variables will increase the processing time without improving accuracy). A correlation coefficient of 0.75 or higher is used to screen off highly correlated variable. Examples of correlated data are roll\_belt vs, yaw\_belt, total\_accl\_belt, accel\_belt\_y, accel\_belt\_z, and accel\_arm\_y.

After this screening, a total number of 36 columns (including classe) was obtained. This were written to a file called “corColumn” to enable easy data subsetting later. See Appendix 1: screening and train data Subsetting.

The corColumn file content include (second column is the column name in the training data set, column one indicate column position, i.e. Roll\_belt is the 8th column in the original training data set)

8 roll\_belt

9 pitch\_belt

37 gyros\_belt\_x

38 gyros\_belt\_y

39 gyros\_belt\_z

44 magnet\_belt\_y

46 roll\_arm

47 pitch\_arm

48 yaw\_arm

49 total\_accel\_arm

60 gyros\_arm\_x

62 gyros\_arm\_z

64 accel\_arm\_y

65 accel\_arm\_z

66 magnet\_arm\_x

67 magnet\_arm\_y

84 roll\_dumbbell

85 pitch\_dumbbell

86 yaw\_dumbbell

113 gyros\_dumbbell\_x

114 gyros\_dumbbell\_y

116 accel\_dumbbell\_x

117 accel\_dumbbell\_y

118 accel\_dumbbell\_z

121 magnet\_dumbbell\_z

122 roll\_forearm

123 pitch\_forearm

124 yaw\_forearm

140 total\_accel\_forearm

151 gyros\_forearm\_x

154 accel\_forearm\_x

155 accel\_forearm\_y

156 accel\_forearm\_z

157 magnet\_forearm\_x

159 magnet\_forearm\_z

160 classe

Training method used

Appendix 2 used caret package to subset the training into training and validation dataset using 50%-50% split. Two training methods were used gbm (generalized boosted model) and rf (random forest).

The model is then applied against the validation data set to get the predicted outcome. Figure 1 is an example showing the predicted outcome for each of the 5 classe using the gbm method. The predicted value is cross-validated with the real result in the validation set and tabulated as follows

pred.gbm A B C D E

A 2731 77 4 3 2

B 28 1737 79 7 29

C 11 67 1593 90 40

D 15 14 33 1500 32

E 5 3 2 8 1700

pred.rf A B C D E

A 2788 19 0 0 0

B 1 1868 26 0 0

C 1 11 1676 47 4

D 0 0 9 1560 2

E 0 0 0 1 1797

Cross-validation

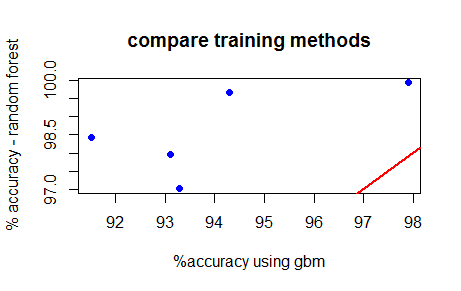
Cursory viewing shows that random forest provides a more accurate prediction. Accuracy and Kappa statistics also show random forest outperforms generalized boosted method.

|  |  |  |
| --- | --- | --- |
|  | Generalized Boosted Method(1) | Random Forest(2) |
| Training time | 13 minutes | 9 minutes |
| Accuracy | 0.936 | 0.983 |
| Kappa | 0.919 | 0.979 |
| Accuracy SD | 0.0044 | 0.0034 |
| Kappa SD | 0.0055 | 0.0043 |

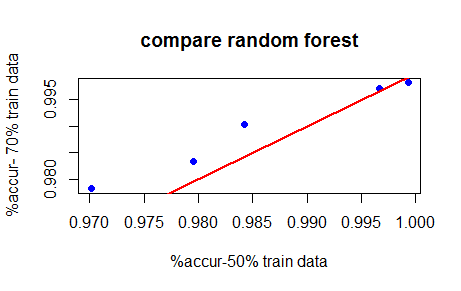
Kappa is a measure of agreement between predicted presences and absences with actual presences and absences corrected for agreement that might be due to chance alone.

1. Quoting statistics of gbm with interaction.depth of 3, n.tree of 150
2. Quoting statistics of rf using mtry = 18

The performance can be visualized easily in the following diagram. Each blue dot represent the prediction accuracy of each classe, in order of A, B, C, D, E. Using the first point as an example, gbm predict with 91% accuracy, whereas random forest predict 98.5%. The red line is a 45 degree slope. Thus any point to the left of it has a high degree of accuracy, i.e. random forest outperform gbm on all fronts.



Random forest is chosen this time using a larger training dataset with 70-30 split for training and validation. Training time increase to 14 minutes. The accuracy comparison is present below graphically. This show the 70-30 split provide marginally better accuracy than the 50-50 split.



Prediction on the provided data set.

Using the same method describe under raw data processing, remove the irrelevant column in the test data set. Add a column called number to identify the observation number. Then apply the prediction model (obtained from random forest with 70-30 split in training data) to each of the observation.