Final Report on Smartphone Activity Competition

Yuchong Zhang, Shun Xi (613-SAFT)

December 1, 2017

1 Overview

The goal of this competition is to predict 6 kinds of human activities based on a 561 dimensional time-series data recorded from a smartphone. We first analyzed the clustering and patterning of the data by dimension reduction methods, and observed that features are highly correlated. A few base models are trained before and after feature selection. Confusion matrix from those base models suggests some classes are easy to predict and some remain problematic. We then feature engineered data by predicting highly misclassified classes based on several key feature. We also used a hierarchical learning approach in which we predicted misclassified activity at later stage. This finally leads us to build an ensemble of base learners in order to reduce the overall variance and improve prediction accuracy. We score 0.94862 in the private board, though we can have some improvement to 0.95511 if we chose another method for submission.

2 Base Learners

We applied multiple base learners to build our predictive models. Their performances in public board and private board are shown in the Table. 1.

The tuning procedure for each method follow the pipeline, as will be in Sec. 4. It is found that among all the multi-classical linear regression learners, group lasso MLR has the best performance since it can deal with the issue of variable correlations and also select features that are of importance. Radial SVM has better performance than linear SVM because the boundary between different classes is non-linear. SVMs and penalized logistic regressions are tunned through a hyper-parameter space grid search.

3 Ensembles

Ensemble methods such as boosting and stacking have the advantage of reducing variance, averaging out bias and being robust from overfitting. We have applied several boosting and bagging methods and also used majority vote for a combination of several base and ensemble learners. The selection criteria of base learners for ensemble is that we select the base learners which are relatively independent to each other so that the variance can be mostly reduced.

Table 1: Performance of different base or ensemble learners

learner	Public Score	Private Score
lasso MLR	0.8865	0.9197
group lasso MLR	0.8959	0.9287
ridge MLR	0.8896	0.9147
linear SVM	0.9151	0.9117
radial SVM	0.9178	0.9207
bagging	0.8850	0.9027
random forest	0.9224	0.9257
adaboosting	0.9104	0.9556
xgboosting	0.9230	0.9551
majority vote	0.9355	0.9486

Random forest, adaptive tree boosting, and xgboost are tuned by stepwise training to avoid overfitting and reduce training time. Xgboost method is considered as the most important for this work due to the nature of this prediction task that half of the data are correlated with the other half from a time series perspective. It is known that xgboost is immune to correlation and high dimension. Though features are correlated in this study, using selected features with base learners such as logistic regression and SVM do not lead to better performance. Number of features sets the upper bound of performance of algorithms. Reducing the dimension might help improve averaged performance but not upper bound.

The performance of these methods on public board and private board are shown in the Table. 1. We find that boosting methods generally yield a high score on the private board and although the majority vote approach does not lead to the best outcome, it is not bad.

4 Model Selection & Assessment

There are quite a few models that are applicable to this multi-classification problem. Usually nested cross-validation is used to select the best method out of all families and further tune the parameters for that particular method. However, since we have relatively plenty opportunities to submit our predictions and see the corresponding performances on public board, we actually devise another pipeline. That is, for a specific method, we use the training data for 5-fold cross-validation and get the optimal parameters according to the accuracy on average test error for the stick-out fold. Then we fit the whole training data using selected parameters from the last step and the model is used to predict for the test data with no label information. The observations belonging to the public board serves as a query data set and the performance on the public board indicates the accuracy of different families of methods. We keep tracking of random number so that the comparison is consistent and the results are reproducible.

Take xgboosting as an example. Since there are quite a few parameters needing to tune, we do the parameter tuning stepwise. We first tune eta and nrounds, since these two parameters are closely related. Then we tune max_depth and

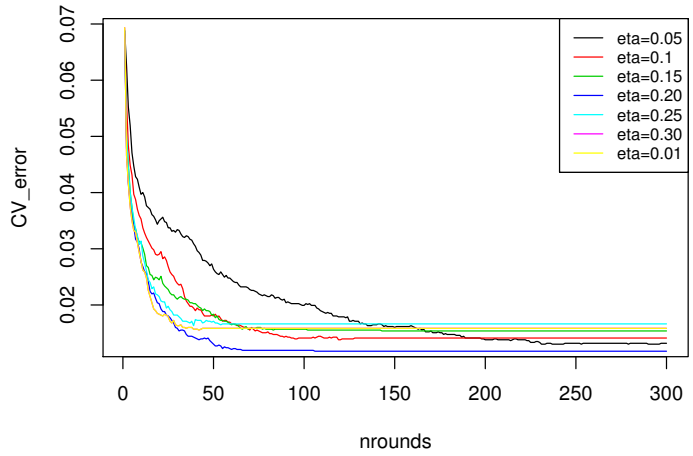

Figure 1: cross-validation on eta and nrounds

min_child_weight. Finally we tune subsample and colsample_bytree. The optimal parameter combination we obtain is eta=0.20, nrounds=100, colsample_bytree=1, max_depth=6, nrounds=100, gamma=0, min_child_weight=1 and subsample=1.

5 Best Scoring Model

Regarding the final models we submitted, Our best scoring model is the xgboosting model. The parameters in our built model are given in Sec. 4. The 5-fold CV error is 0.01177, the prediction accuracy on public and private boards are 0.9230 and 0.9551, respectively.

6 Interesting Findings

What we find interesting is the six activities can be actually divided into two major groups. We implemented Principal Component Analysis on the training data and found that observations labeled as walking, walking downstairs and walking upstairs appear in one group and observation labeled as sitting, standing and laying are in another group and these two groups are well separated. This is reasonable because walking, walking downstairs and walking upstairs all belong to dynamic activities and sitting, standing and laying all belong to static activities. They should have quite large differences for several features, for example, gravitational acceleration. From confusion matrix, we also observe that most misclassification occur within group (dynamic, static) instead of between groups. Furthermore, we find that it can be challenging to distinguish between the sitting class and standing class as in Fig. 2. We also tried a non-linear dimension reduction method to the training data and had a clearer clustering

than PCA method (Fig. 3). This implies a non-liner learning algorithm such as boosting tree and bagging tree are necessary for this study. Therefore, we think about modeling the data in a hierarchical way. That is, we come up with some learner to divide an observation in two major groups. Then with each group, we may use different learners for further classification.

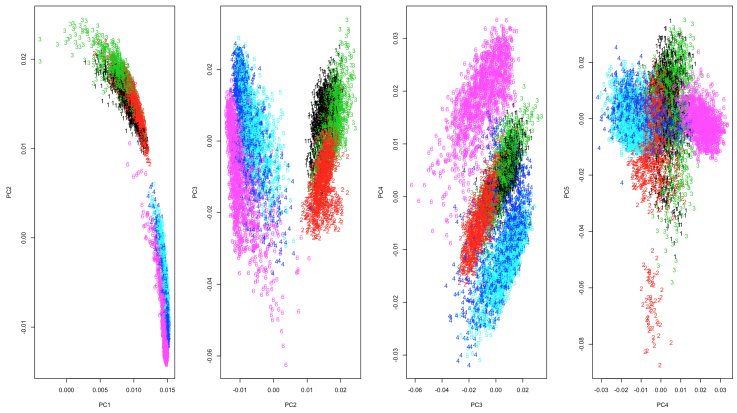

Figure 2: PCA of all training data

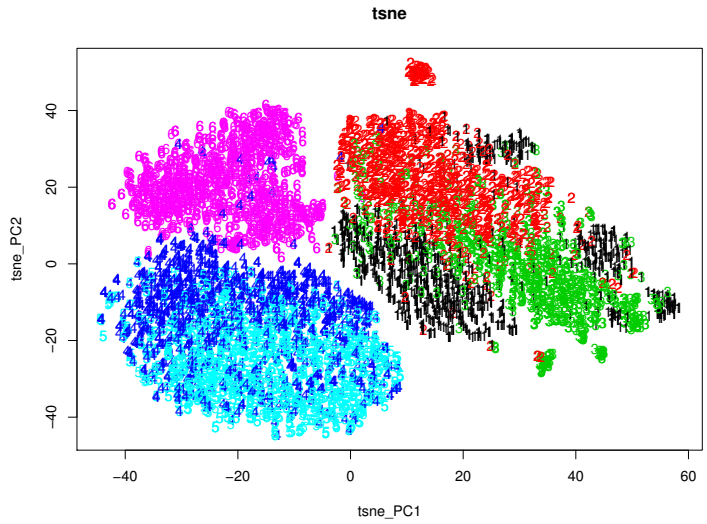

Figure 3: t-SNE for all training data

We implemented PCA on the subset of data labeled as sitting and standing, while still found it hard to separate these two classes, as shown in Fig. 4. We tried several approaches to further classify sitting and standing[1,3]. We first applied directly boosting and bagging decision-tree methods using all features of predetermined sitting and standing data. We also tried to train models with only selected features that are reported to differentiate sitting and standing[2].

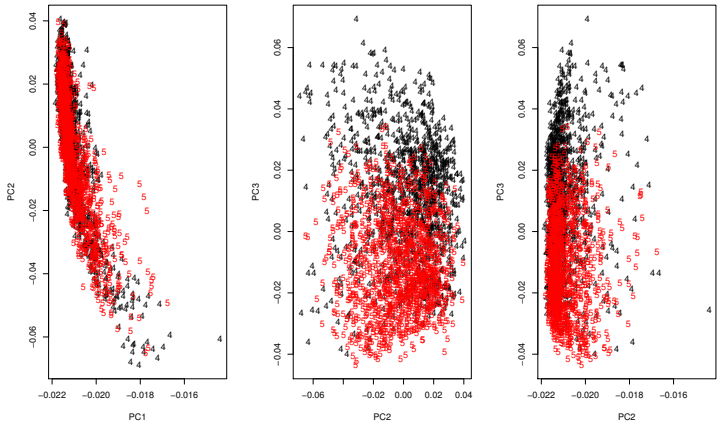

Figure 4: PCA of Sitting and Standing

Due to the normalization of data, applying simple threshold to selected features such as mean, standard deviation and max of gravity acceleration, and difference to gravity acceleration in y direction do not lead to better predictions. Overall no significant accuracy is observed by hierarchical learning from our CV errors so we have not gone further for this method.

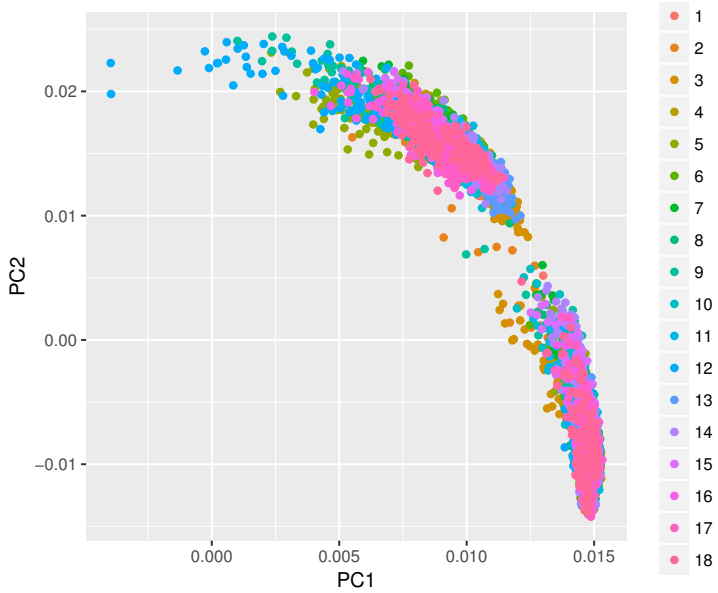

Figure 5: PCA by Subjects

Through some literature study, we also learned that to predict human-related activities, the cross-validation can be implemented by leave-one-person-out cross validation technique: training data is divide into k-fold where k is the same as number of subjects, and each time we cross-validate model against data of a

single subject[4,5]. This makes sense because it is highly possible that certain subjects may have special characteristics that do not follow the general features. If necessary, subjects with suspicious predictive behavior should be removed from the training data. By taking another view of clustering of each subjects in Fig. 5, it is easily to pick outliers from other subjects. Training models with data without outliers might lead to better prediction accuracy. However, we believe it remains uncertain. A more appropriate approach is that given some prior information about the distribution of outliers in population, we weight the contribution of each subject differently during cross-validation. There have been a few study of human activity recognition where different models are built for different groups of subjects at different health situation[2].

From models such as random forest, xgboosting, we can also get the information about which features are important in prediction. For example, from xgboosting, we see that the most important features include tGravityAcc-min()-X, fBodyAccMag-mad() and tGravityAcc-max()-Y.

7 What We Learned

We learned how to practice machine learning methods from this competition. One important lesson we have learned is that we should rely on the cross-validation errors rather than the score in the public board. The way that we treat the public board related data as query set may not be the best because those observations can be skewed and having strong outliers that are not in training data or private data, leading to a misconception of the prediction accuracy of different models.

Another important lesson we learned is it is crucial to look at data by visualization and dimensional reduction techniques to detect correlation and patterning within the data before building model. Know how data might be varied from groups to groups, and how to pick outliers are also important.

Acknowledgement:

Shun Xi and Yuchong Zhang developed and coded up all the models, worked on the presentation slides and reports. Shun Xi gave the final presentation.

References:

1. Capela, Nicole A., Edward D. Lemaire, and Natalie Baddour. 2015 IEEE International Symposium on. IEEE, 2015.
2. Capela, Nicole A., Edward D. Lemaire, and Natalie Baddour. PloS one 10.4 (2015): e0124414.
3. Gjoreski, Hristijan, et al. Proceedings of the Twenty-first European Conference on Artificial Intelligence. IOS Press, 2014.
4. Gjoreski, Hristijan, et al. Applied Soft Computing 37 (2015): 960-970.
5. Venables, William N., and Brian D. Ripley. Modern applied statistics with S-PLUS. Springer Science & Business Media, 2013.