Random Forests and Nearest Shrunken Centroids for the Classification of eNose data

Matteo Pardo*, Giorgio Sberveglieri
Sensor Lab, CNR-INFM & University of Brescia, Brescia, Italy
*pardo@ing.unibs.it

Artificial Olfactory Systems or eNoses are instruments that analyze gaseous mixtures for discriminating between different (but similar) mixtures and, in the case of simple mixtures, quantify the concentration of the constituents. eNoses consist of a gas sampling system (for a reproducible collection of the mixture), an array of chemical sensors, electronic circuitry and data analysis software (Pearce, 2003). Random Forests (RF) and (NSC) are state of the art classification and feature selection methodologies and have never been applied to eNose data.

RFs are ensembles of trees, where each tree is constructed using a different bootstrap sample of the data and each node is split using the best among a subset of features randomly chosen at that node. RF has only two hyper parameters (the number of variables in the random subset at each node and the number of trees in the forest) (Breiman, 2001). NSC classification makes one important modification to standard nearest centroid classification. It "shrinks" each of the class centroids toward the overall centroid (for all classes) by an amount called the threshold (Tibshirani, et al., 2003).

In this paper we compare the classification rate of RF, NSC and Support Vector Machines (SVM) -which we consider as a top level reference method- on three eNose datasets for food quality control applications. Classifiers' parameters are optimized in an inner cross-validation cycle and the error is calculated by outer cross-validation in order to avoid any bias. To carry out computations we used the R package MCRestimate (Ruschhaupt, et al., 2004). MCRestimate is built on top of a number of R packages, e.g. the *randomForest* package (Liaw and Wiener, 2002).

We were interested in three computational aspects:

1. Relative performance of the three classifiers.
2. Since nested cross-validation is computationally expensive we also investigate the dependence of the error on the number of inner and outer folds. We considered a grid of 25 outer folds/ inner folds numbers: outer CV folds: 2, 4, 6, 8, 10; inner CV folds: 2, 4, 6, 8, 10. Altogether this means training e.g. 45050 SVM.
3. Feature rankings produced by RF and NSC.

We find that:

1. SVM and RF perform similarly (each classifier does better on one problem), while NSC consistently performs worse. NSC is by far the simpler (and faster) classifier.
2. There is a slight dependence on the number of external CV folds (particularly in the *fungi* dataset, where four external CV folds produce a consistently –over internal CV folds number- higher classification rate), while the number of inner CV folds seems to be immaterial.
   2x2 nested CV is often enough for a good result. With respect to 10x10 CV, 2x2 CV requires 4% of the training time, so this result may spare quite some time in future computational studies.
3. Of the 30 original features, RF and NSC have the same two top positions. Further, they share other four features in the top ten. Other four features have quite, or even very different rankings. In fact, NSC – differently from RF- ranks features individually and independently in the classifier construction process. In this way, on the one hand it cannot consider the joint discrimination capabilities of features groups and on the other hand it does not exclude correlated features.

Breiman, L. (2001) Random forests, *Machine Learning*, **45**, 5 - 32.
Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *The Newsletter of the R Project*.
Pearce, T.C. (2003) *Handbook of machine olfaction : electronic nose technology*. Wiley-VCH, Weinheim [Germany].
Ruschhaupt, M., Huber, W., Poustka, A. and Mansmann, U. (2004) A Compendium to Ensure Computational Reproducibility in High-Dimensional Classification Tasks, *Statistical Applications in Genetics and Molecular Biology*, **3**, 37.
Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003) Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays, *Statist. Sci.*, **18**, 104-117.