Classification of connectivity data.

Bla, bla, bla, blabla...

Andrea Insabato
Universitat Pompeu Fabra
Theoretical and Computational Neuroscience
Center for Brain and Cognition
Roc Boronat, 138
08018 Barcelona, Spain,

Matthieu Gilson
Universitat Pompeu Fabra
Theoretical and Computational Neuroscience
Center for Brain and Cognition
Roc Boronat, 138
08018 Barcelona, Spain

Vicente Pallarés Picazo
Universitat Pompeu Fabra
Theoretical and Computational Neuroscience
Center for Brain and Cognition
Roc Boronat, 138
08018 Barcelona, Spain,

Keywords: machine-learning; classification; effective connectivity; functional connectivity; features selection; dimensionality reduction
Running title: Classification of connectivity data.
March 12, 2018

Abstract

This is the abstract.

Contents

1 Introduction

2 Methods

3 Results

3.1 Cross-validation to assess generalization performance

With high dimensionality datasets, in particular when p¿n, it is paramount to use cross-validation in order to avoid overfitting. Indeed, given N points there will always be an N-1 dimensional hyperplane that separates all possible arrangement of labels in a binary classification problem (Cover). **We can show that the separating hyperplane calculated over different subsets of training samples (of size 1 samples per subject) have a wide variability and are indeed not different than those estimated from random arrangement of lables. Notably the misclassification rate on the training set is always 0 (both for real lables and random ones). This indicates the need to control for overfitting since the retrived boundary might not generalize well to new samples.** However we note that the arrangement of real labels is not random, thereby mitigating this effect. Indeed the accuracy on the test set for real labels was significantly higher than for random labels.

Many studies used limited datasets that do not allow a comprehensive assessment of generalization performance. We used a dataset that allows to use multiple sessions to test the generalization of the learned model. Here we show the distribution of generalization score as a function of the number of samples in the test set.

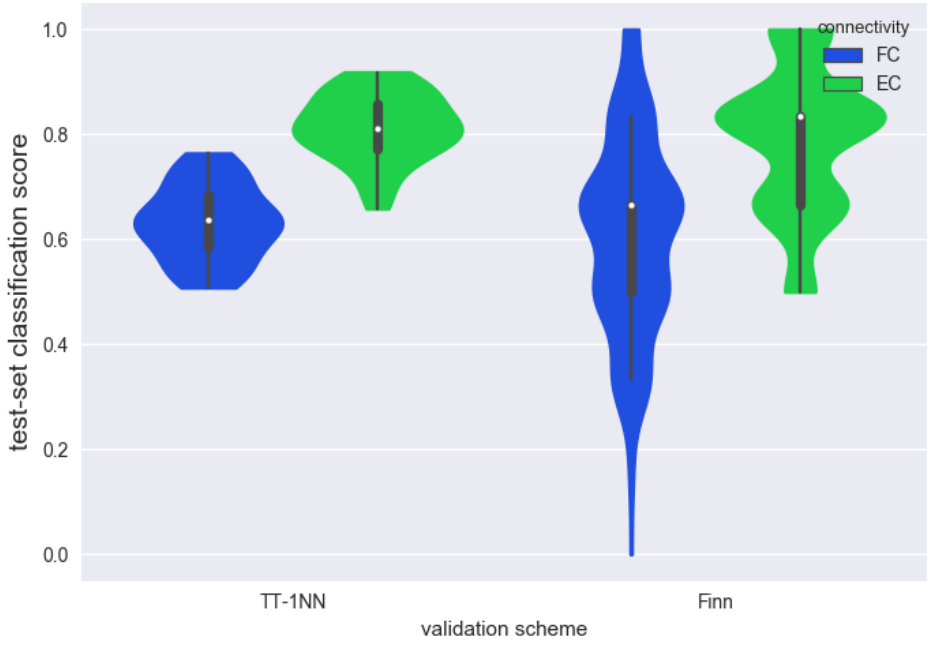

Figure 1: Distribution of test-set generalization score for large and small test set size.

3.2 Comparison of classification pipelines for subjects' identity classification

3.2.1 z-score

Feature-wise

Sample-wise

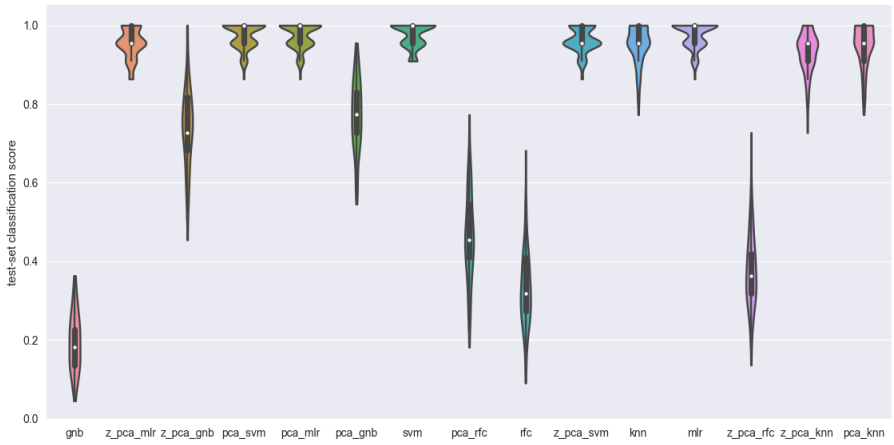

Figure 2: Comparison of classifiers.

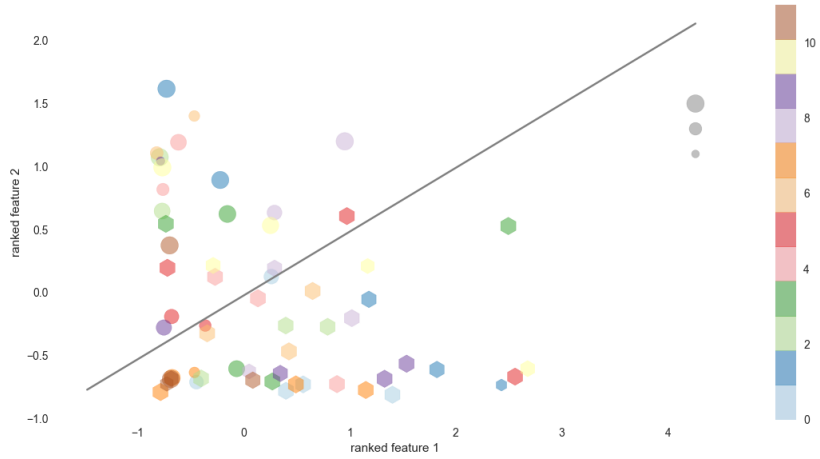

Figure 3: Probability of class assignment.

discuss generality, dependence on parameters, etc.

discuss other possible approaches (other models, methods, etc.)

possible applications to other fields, themes, etc.

other collateral themes

discuss future directions

Acknowledgments

We thank...

References