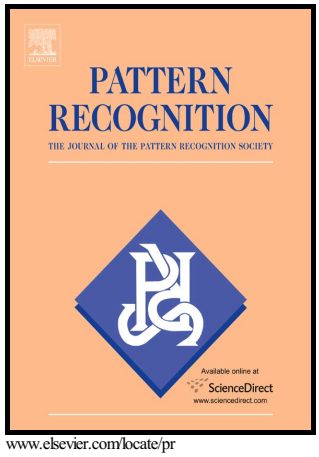

www.elsevier.com/locate/pr

Machine Learning on High Dimensional Shape Data from Subcortical Brain Surfaces: A Comparison of Feature Selection and Classification Methods

Benjamin S.C. Wade, Shantanu H. Joshi, Boris A. Gutman, Paul M. Thompson

PII: S0031-3203(16)30295-3
DOI: http://dx.doi.org/10.1016/j.patcog.2016.09.034
Reference: PR5896

To appear in: *Pattern Recognition*

Received date: 27 January 2016
Revised date: 24 August 2016
Accepted date: 21 September 2016

Cite this article as: Benjamin S.C. Wade, Shantanu H. Joshi, Boris A. Gutmar and Paul M. Thompson, Machine Learning on High Dimensional Shape Dat from Subcortical Brain Surfaces: A Comparison of Feature Selection anc Classification Methods, *Pattern Recognitior* http://dx.doi.org/10.1016/j.patcog.2016.09.034

Machine Learning on High Dimensional Shape Data from Subcortical Brain Surfaces: A Comparison of Feature Selection and Classification Methods

Benjamin S.C. Wade[1*], Shantanu H. Joshi[2], Boris A. Gutman[1], Paul M. Thompson[1]

[1]Imaging Genetics Center, USC, Marina del Rey, CA, USA
[2]Ahmanson-Lovelace Brain Mapping Center, UCLA, Los Angeles, CA, USA

[*]**Corresponding author.** Benjamin Wade, Imaging Genetics Center, University of Southern California, Marina del Rey, CA, USA, 4676 Admiralty Way, Marina del Rey, CA 90292, phone: 323-442-7246, email: ben.wade@loni.usc.edu

Abstract

High-dimensional shape descriptors (HDSD) are useful for modeling subcortical brain surface morphometry. Though HDSD is a useful basis for disease biomarkers, its high dimensionality requires careful treatment in its application to machine learning to mitigate the curse of dimensionality. We explored the use of HDSD feature sets by comparing the performance of two feature selection approaches, Regularized Random Forest (RRF) and LASSO, to no feature selection (NFS). Each feature set was applied to three classifiers: Random Forest (RF), Support Vector Machines (SVM) and Naïve Bayes (NB). Paired feature-selection-classifier approaches were 10-fold cross-validated on two diagnostic contrasts: Alzheimer's disease and mild cognitive impairment, both relative to controls across varying sample sizes to evaluate their robustness. LASSO aided classification efficiency, however, RRF and NFS afforded more