Summary of A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies

1 Background

* Problem:
  + Question raised from recent comparison studies about neutrality, unbiasedness, utility and ways of performed comparisons as there is little consensus between the findings of such studies
  + Existence of tendency: Demonstrate superiority of specific method using datasets favourable for the chosen method + multi-rule bias
* Neutral comparisons:
  + Limitation: Comparison mainly based on real datasets 🡪 problem with sampling error or noise in the estimated performance measures
  + Best performance in one or a few instances does not imply so on average or on population level
* Goal of study in the paper:
  + Extensive simulation experiment to compare classification performance of LDA (Linear Discriminant Analysis), SVM, Random Forest (RF) and kNN
  + Evaluate and compare generalisation errors (leave-one-out cross validation (CV) errors) for large number of combinations of seven factors:
    - number of variables (p),
    - training sample size (n),
    - biological (or, between-subjects) variation (σb),
    - within-subject variation (σe),
    - effect size (fold-change, Θ),
    - replication (r) and correlation (ρ) between variables
  + Create guideline “which method performs better in what circumstances”
  + Evaluation of findings with simulated data on real life experimental datasets

2 Methods

* Optimizing tuning parameters:
  + For every method on every simulated dataset using grid search over supplied parameter space (including software default!)
  + RF: quite robust to variation of its tuning parameters, only mtry important (= # variables to be used as candidates at each node) + in paper: nodesize (= minimum size of terminal nodes) + ntree (= # trees to grow)
  + SVM: kernel = radial basis function, since less biased and more general, depending on *cost* vlb (= controls margin of SV, smaller value relaxes penalty on margin error 🡪 increases margin of classification) + *gamma* (= controls curvature of the decision boundary – higher values 🡪 more flexible (non-linear) decision boundary)
  + kNN: parameter *k* (= number of nearest neighbours)
* Performance estimators:
  + Classification error, sensitivity, specificity
  + Leave-one-our CV (chosen out of 4 possible estimators) 🡪 less biased, more variable for smaller samples (applicable to data of any size)
* Example datasets:
  + Real datasets from: bipolar gene-expression data from DNA microarrays, MRI brain imaging data (from Alzheimer disease Neuroimaging Initiative database) and measured brain activity derived from EEG system
* Simulation:
  + Simulation model to generate realistic gene expression data (vary: variability, effect size, correlation, etc.) 🡪 investigate effects of such data characteristics on the performance of classification algorithms
  + Simulation based on real microarray data:
    - Set of base expressions (µ), proportional to true mRNA abundance signal of the corresponding marker due to normalised and averaged over many individuals
    - With random effects model: between-subject (biological) and within-subject (technical) variation
    - For given training sample size (n) simulate data for each marker independently with:



Generate independent or uncorrelated gene expression data with different amount of stochastic noise

* + -  = simulated log2 expression value for marker in the jth replicate (1…r) of the ith subject (1…n), each  independently (uncorrelated) 🡪 univariat
    - = random effect for the ith biological subject
    -  = random experimental noise
    - Problem: Independence rarely a realistic assumption!
    - In gene expression data: genes working together forming network relationships to perform certain biological functions (variables within such groups highly correlated)
    - Use block-diagonal correlation matrix to model such network relationships

3 Results

* For main simulation:
  + Average leave-one-out cross-validation estimate of classification error over the 500 replications of simulated datasets for all the four methods (RF, SVM, LDA and kNN) plotted against the values of the various data characteristics