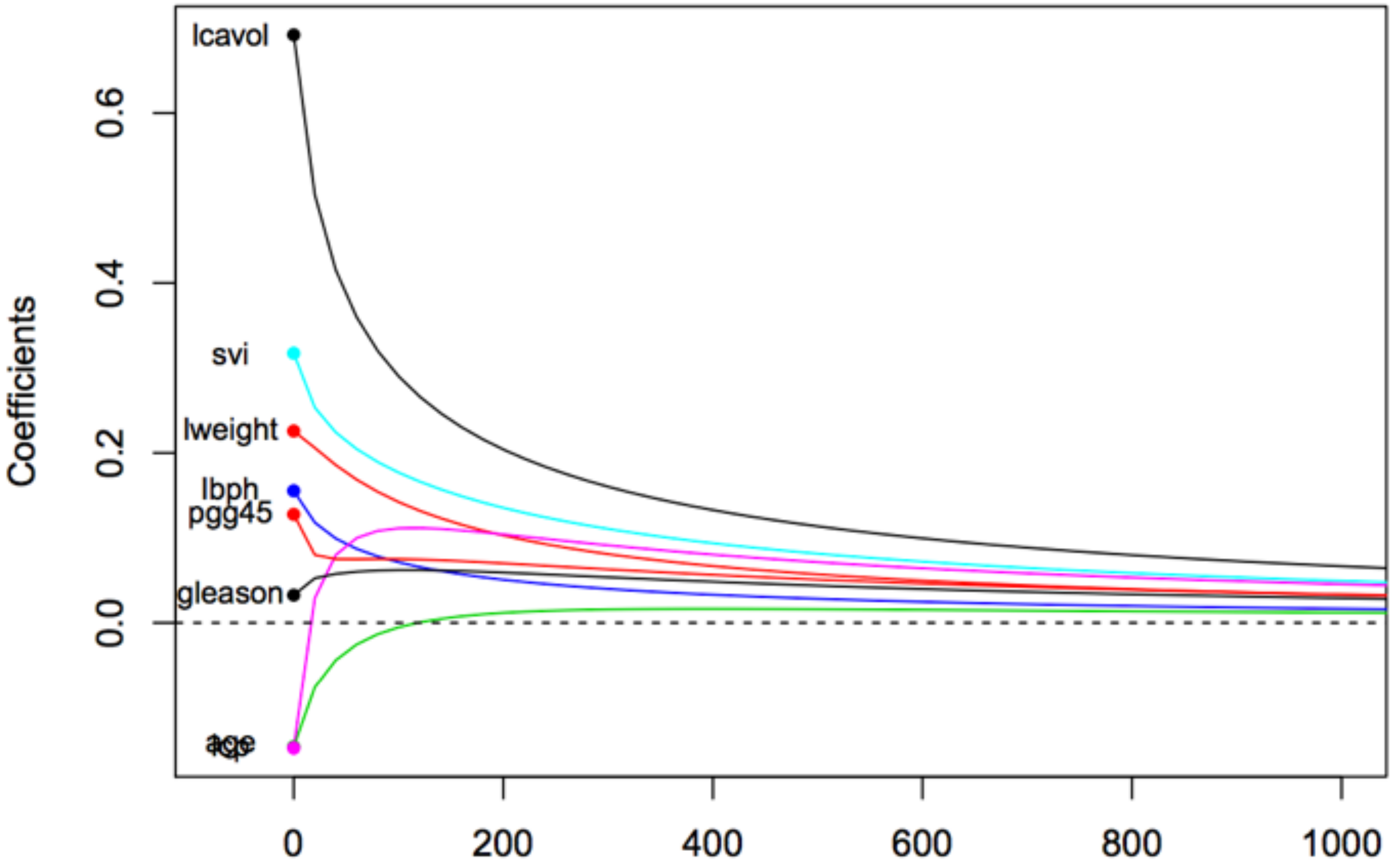


Feature Selection

Two-day workshop

Duke University
Adam Chekroud

Motivation for feature selection

Feature selection (aka variable selection) is the process of selecting a subset of (relevant) features for model construction

Motivation

Some of our data is *redundant* or *irrelevant*

- Identify and remove it

Result?

1. Simplifies models to improve interpretability
2. Alleviates computational demand
3. Improves generalizability by avoiding *overfitting*
    - "reduction of variance"

Methods for feature selection

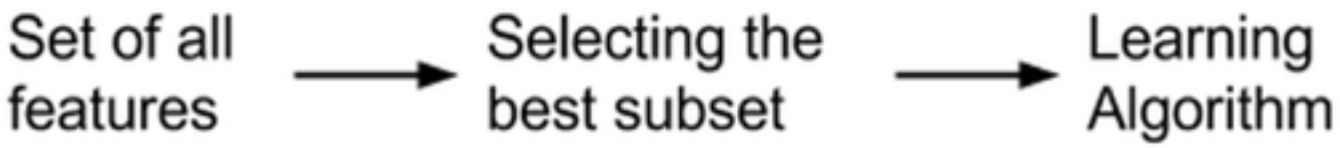
Choosing features that give us as good or better performance with less data

How?

Three general classes:

- Filter method
  - Minimum correlation coefficient!
- Wrapper method
  - Stepwise, or recursive feature elimination
- Embedded method
  - Least angle regression, regularization

Filter methods



Apply a statistical measure to “rate” each variable

Use these ratings to decide which variables to keep

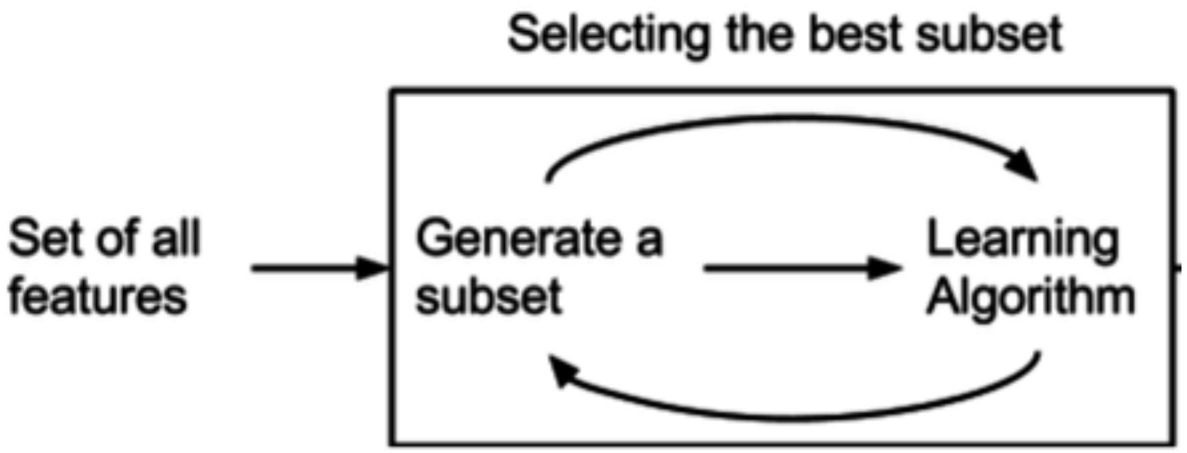
Which measures?

- Typically univariate (although decreasingly so)
    - e.g. p-val from GLM contrast in fMRI
- Can consider feature independently
    - e.g. variance of predictor
- Or with regard to the target
    - e.g. correlation coefficient, chi-squared

Pros and cons?

- Pro: *really* quick, quite effective at eliminating uninteresting variables, doesn’t usually overfit
- Con: can select redundant variables (keep two good, but correlated variables)

Wrapper methods



- Evaluate subsets of features in combination
- Use performance of model to decide which *group* of variables to keep

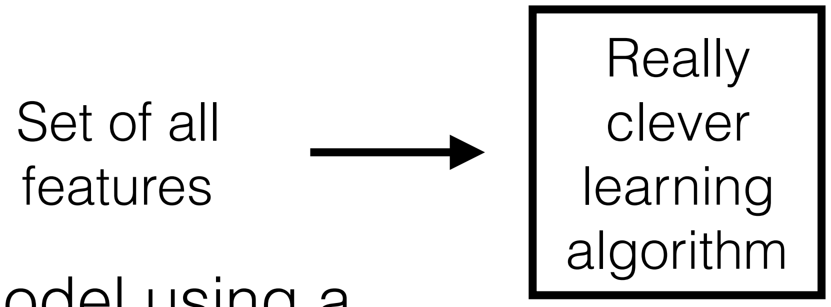
Search procedures

- Forward/Backward stepwise selection
- Recursive feature elimination
- Advanced: genetic algorithms, simulated annealing

Pros and cons?

- Pro: can detect interactions, usually give good performance, advanced methods can technically find “optimal” solution
- Con: *really* slow. Serious risk of overfitting.

Embedded methods



Set of all features

Really clever learning algorithm

Definition: a machine learning algorithm that returns a model using a limited number of features

- Variable selection is built in ("embedded") to the learning algorithm
- Typically this is done using "regularization" methods

Examples:

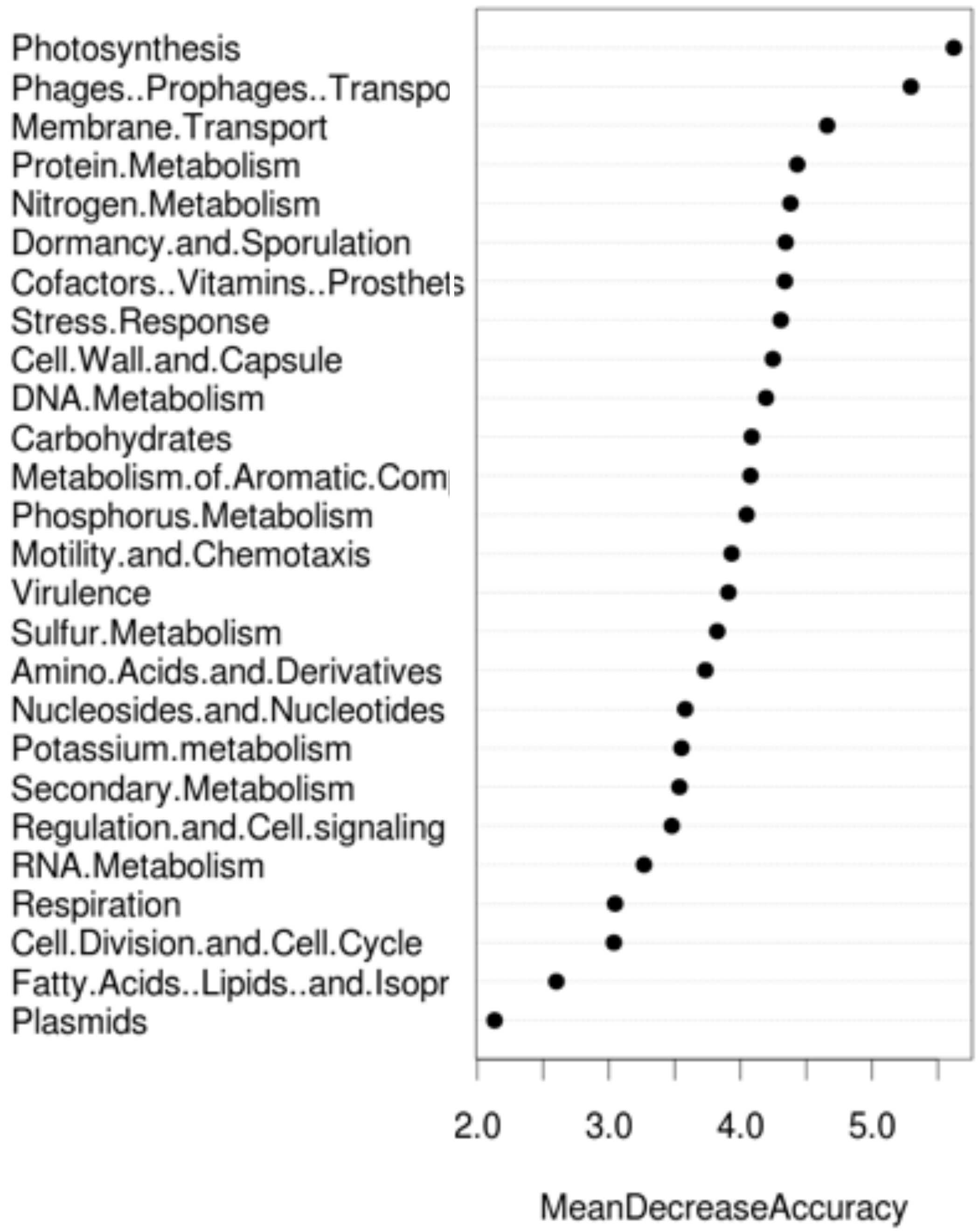
- Decision trees!
- Penalized regression
    - L1-norm: LASSO, least angle regression
    - L2-norm: Ridge regression
    - Blend: Elastic net regression
- Fancier adaptations of some models: e.g. support vector machines

Pros and cons?

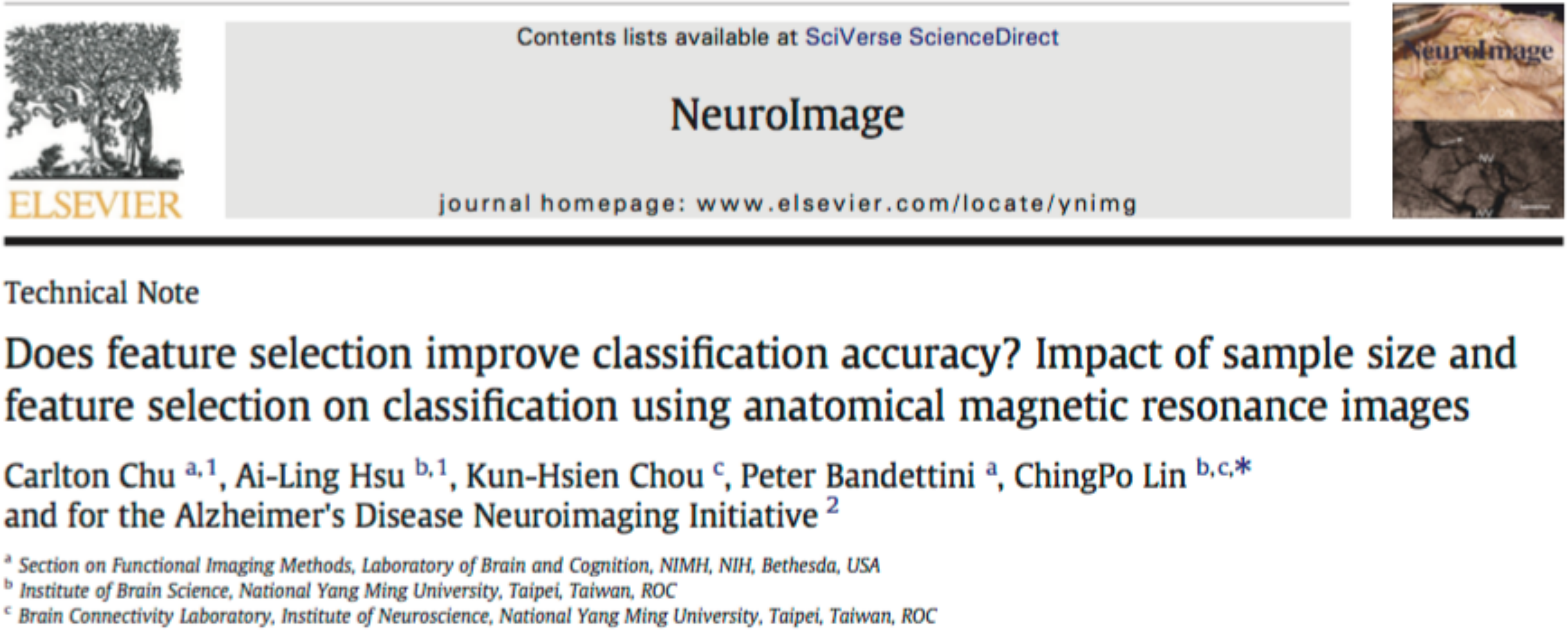
- Pro: can detect interactions, don't fit many models, gives good performance
- Con: Limited selection of algorithms, technically more challenging.

Variable Importance Plot

- FS outcomes usually represented with variable importance plot



- Visual ranking of predictor utility
- Various ranking metrics available
  - r value
  - t-stat
  - reduction in RSS
  - permutation based measures



Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Technical Note

Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images

Carlton Chu [a,1], Ai-Ling Hsu [b,1], Kun-Hsien Chou [c], Peter Bandettini [a], ChingPo Lin [b,c,*] and for the Alzheimer's Disease Neuroimaging Initiative [2]

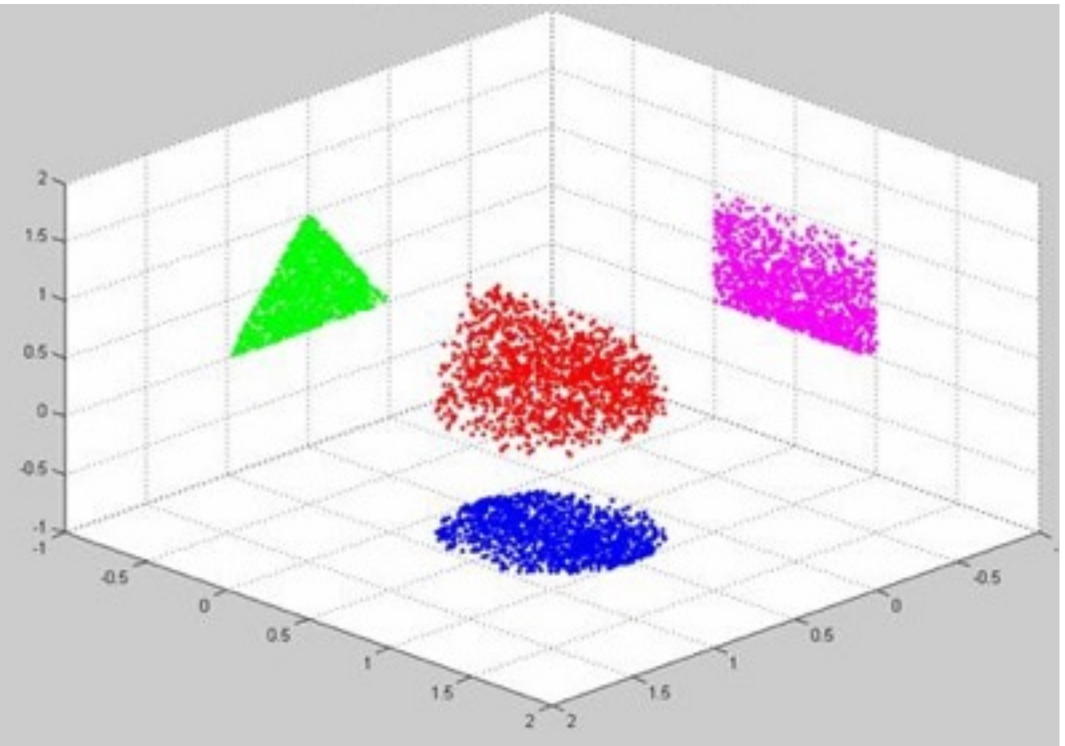
[a] *Section on Functional Imaging Methods, Laboratory of Brain and Cognition, NIMH, NIH, Bethesda, USA*
[b] *Institute of Brain Science, National Yang Ming University, Taipei, Taiwan, ROC*
[c] *Brain Connectivity Laboratory, Institute of Neuroscience, National Yang Ming University, Taipei, Taiwan, ROC*

- NB: “feature selection” based on relevant *a priori* ROIs outperformed traditional automated feature selection

Feature Selection - Warnings

Where things can go wrong

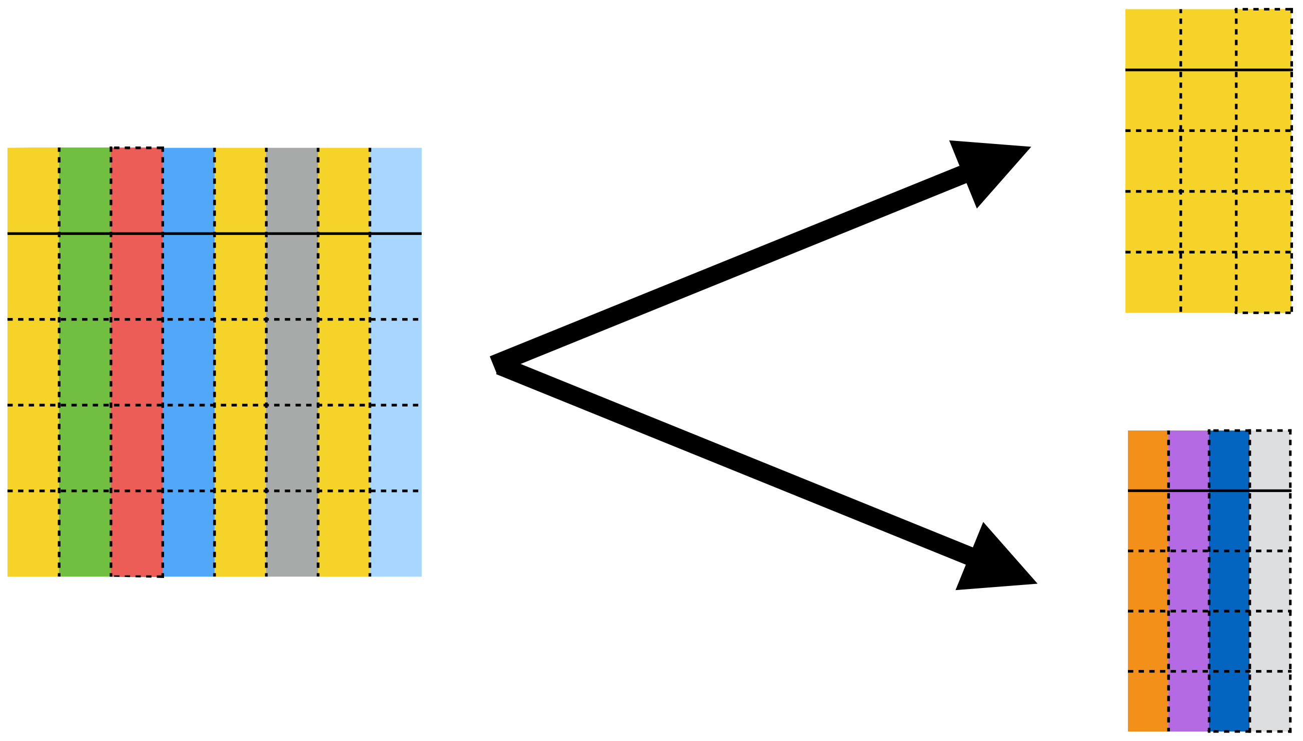
- Feature subset misrepresents reality! >
- Crash your computer



- Overfitting
- Multiple model testing

- Using same data to pick features and examine model performance
- Equivalent to test set peeking!

Relationship with dimensionality reduction

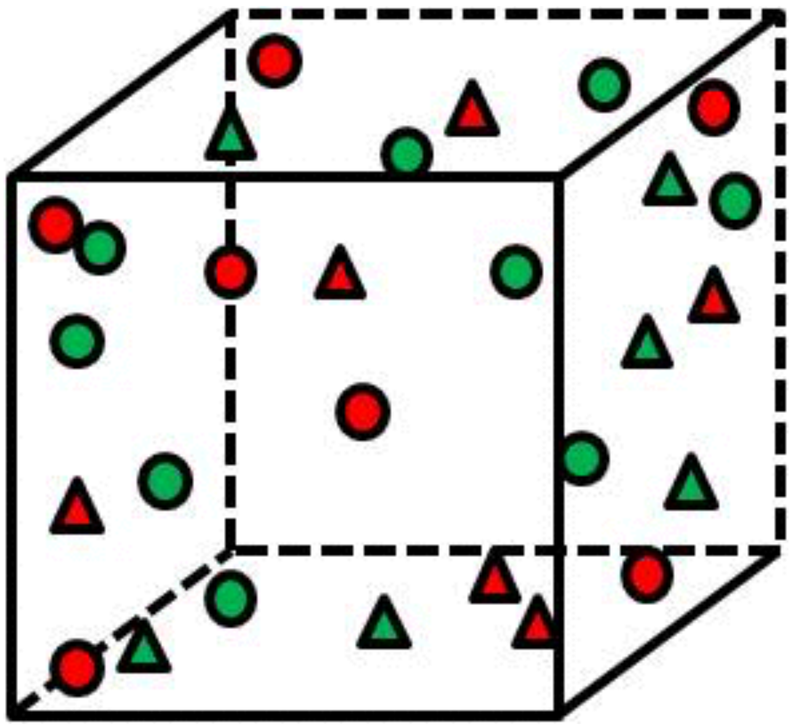


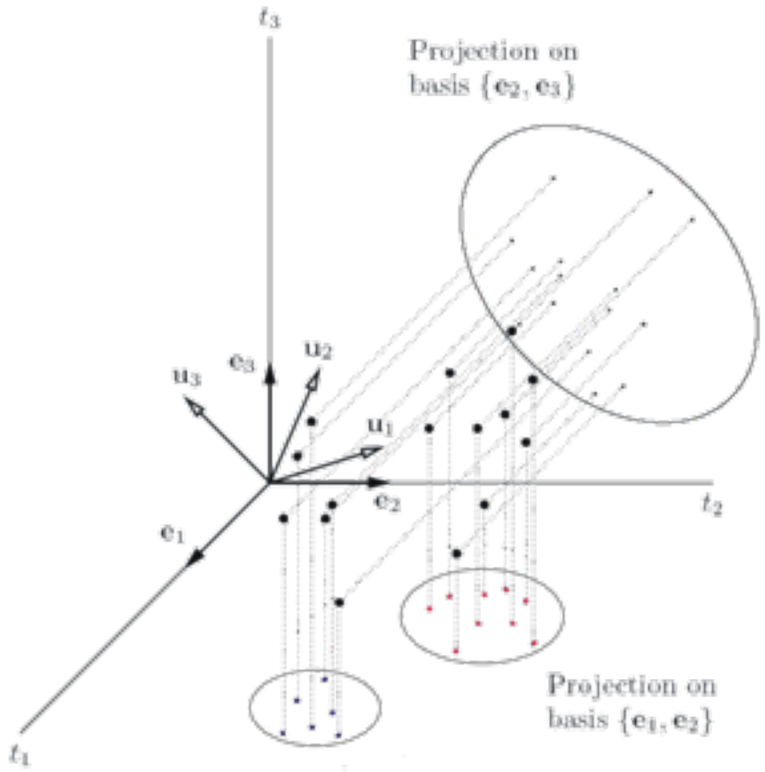
FS: Identify three best “raw” variables

DR: Combine all variables into smaller number

- Both seek to reduce the number of attributes in the data
    - Dimensionality reduction methods create combinations of features
    - Feature selection methods include or exclude attributes that already exist, without changing them

“The curse of dimensionality”




t_3
Projection on basis $\{\mathbf{e}_2, \mathbf{e}_3\}$
$\mathbf{u}_3$
$\mathbf{e}_3$
$\mathbf{u}_2$
$\mathbf{u}_1$
$\mathbf{e}_2$
t_2
$\mathbf{e}_1$
t_1
Projection on basis $\{\mathbf{e}_1, \mathbf{e}_2\}$

Data (usually) exist in a high-dimensional space

- Points may be close on one dimension, but very sparse in complete high-dimension space

Learning about the structure of this space is difficult!

- Requires a huge amount of data to ensure several samples with each combination of variables

“The curse of dimensionality”

Can we preserve informative structure in a lower-dimensional space?

We can try:

- PCA/ICA, non-Negative Matrix Factorization (nNMF), locally linear embedding

Not easy:

- Reducing dimensionality reduces information available for prediction
- Do it wrong, destroy your data

Knowledge discovery by accuracy maximization

Stefano Cacciatore[a,b,c], Claudio Luchinat[a,d,1], and Leonardo Tenori[d]

[a]Magnetic Resonance Center (CERM) and Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy; [b]Department of Medical Oncology, Dana–Farber Cancer Institute, Harvard Medical School, Boston, MA 02115; [c]Metabolomics Platform, Spanish Biomedical Research Center in Diabetes and Associated Metabolic Disorders, Rovira i Virgili University, 43007 Tarragona, Spain; and [d]FiorGen Foundation, 50019 Sesto Fiorentino, Italy

(Unofficial) guidelines/suggestions

Do you have domain knowledge? If yes, construct a better set of custom features

Are your features on comparable scales? If no, consider normalizing them.

Do you suspect interdependence of features? If yes, expand your feature set by constructing interaction or conjuctive features, as much as your computer resources allow you.

Do you need to prune the input variables (e.g. for cost, speed or data understanding reasons)? If no, construct weighted sums of feature

Do you need to assess features individually (e.g. to understand their influence on the system or because their number is so large that you need to do a first filtering)? If yes, use a variable ranking method; else, do it anyway to get baseline results.

Do you suspect your data is "dirty" (has a few meaningless input patterns and/or noisy outputs or wrong class labels)? If yes, detect the outlier examples using the top ranking variables obtained in step 5 as representation; check and/or discard them.

Do you know what to try first? If no, use a linear predictor. Use a forward selection method. Can you match or improve performance with a smaller subset? If yes, try a non-linear method with that subset.

Do you have new ideas, time, computational resources, and enough examples? If yes, compare several feature selection methods, including your new idea, correlation coefficients, backward selection and embedded methods. Use linear and non-linear predictors. Select the best approach with model selection

Do you want a stable-er solution (to improve performance and/or understanding)? If yes, subsample your data and redo your analysis for several "bootstrap".

Adapted from Guyon and Elisseeif (2003) *J. ML research*
http://jmlr.csail.mit.edu/papers/volume3/guyon03a/guyon03a.pdf