Biomarker Identification for Early-Stage Diabetes Diagnosis in Mice Liver Cells

Andrea Agazzi[1], Vincent Deo[2]

[1] Theoretical Physics Department, Université de Genève, Switzerland
[2] Electrical Engineering Department, Stanford University, CA, U.S.A.
contact: *vdeo, agazzian@stanford.edu*

Abstract

Diabetes Mellitus is one of today's most common chronic diseases, affecting approximately 382M people worldwide in 2013. Type 2 diabetes, accounting for ~90% of cases, is a gradually developing disease, and can be prevented if diagnosed early.

The liver controls sugar metabolism and is expected to be influenced by diabetes in its earliest stage. Given the intrinsically metabolic nature of this disease, it is an appealing challenge to identify incipient stage biomarkers in liver metabolic profiles through the application of different machine learning algorithms for classification and feature selection.

Dataset

112 samples of metabolic profiles: concentrations of 756 metabolites measured by mass spectrometry, from liver cells of mice with very similar genetic profiles. Half of the mice received a genetic knock-out (KO) inducing diabetes, the others were unaltered (CTRL). The distribution of the dataset is:

Type	Week 4	Week 5	Week 6	Week 10 (post-symptoms)
CTRL	16	14	18	16
KO	14	10	12	12

Classification - Breadth Approach

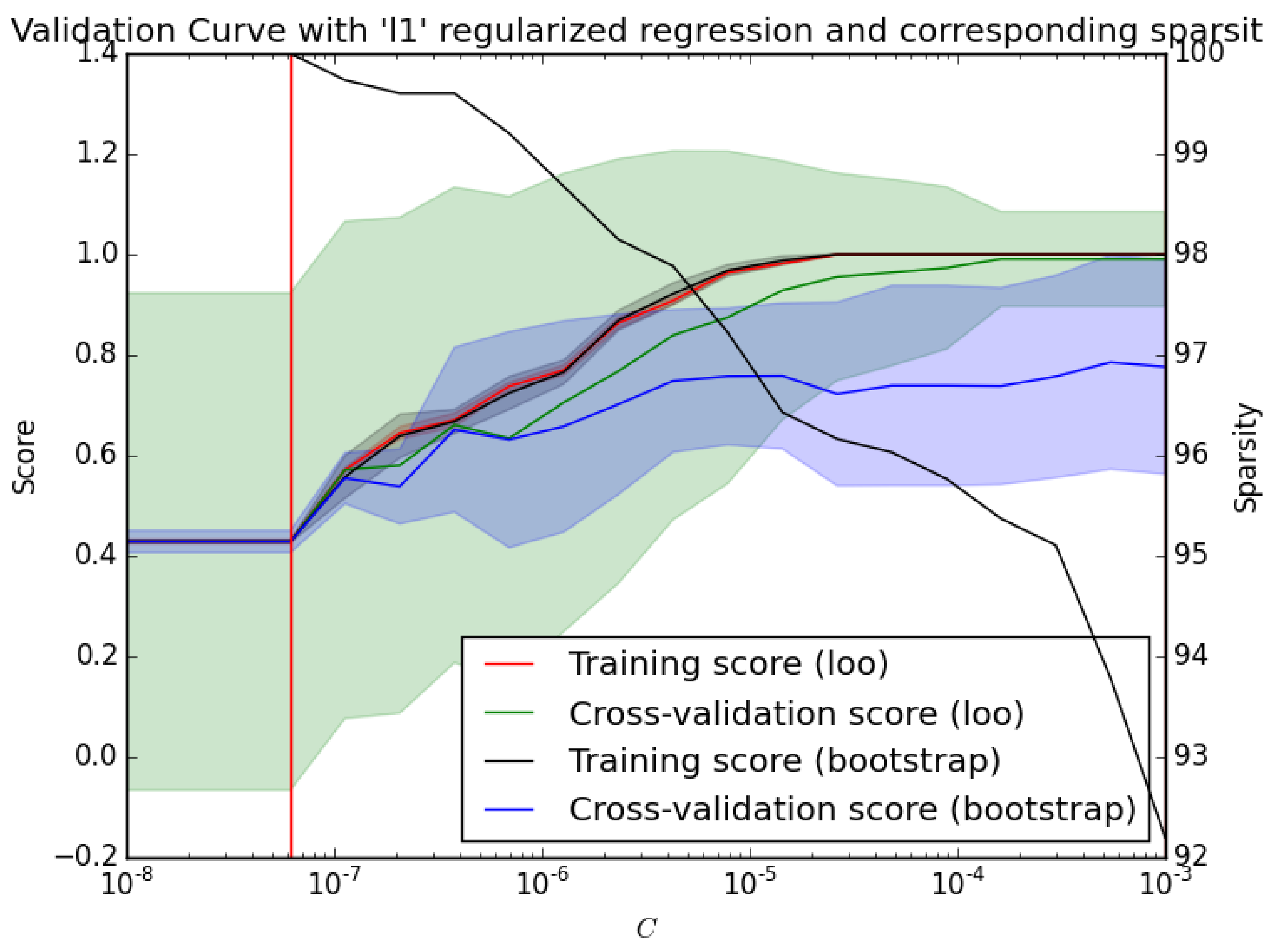
Several standard classifiers were attempted to solve this problem, as listed below with their best cross-validation scores:

Method	Preproc.	CV score	Sparsity
Linear Discriminant (LDA)	PCA	97.4%	60.3%
Linear SVM	PCA	96.1%	0%
Quadratic SVM	PCA	92.7%	0%
L_1-penalized Logistic Reg.	·	90.8%	95.6%
L_2-penalized Logistic Reg.	PCA	94.1%	0%
Random Forests (RF)	·	97.4%	0%

We selected the LDA, L_1-classifier and Random Forest approaches as a core onto which build technical refinements. The L_1-classifier was selected for natural feature extraction properties.

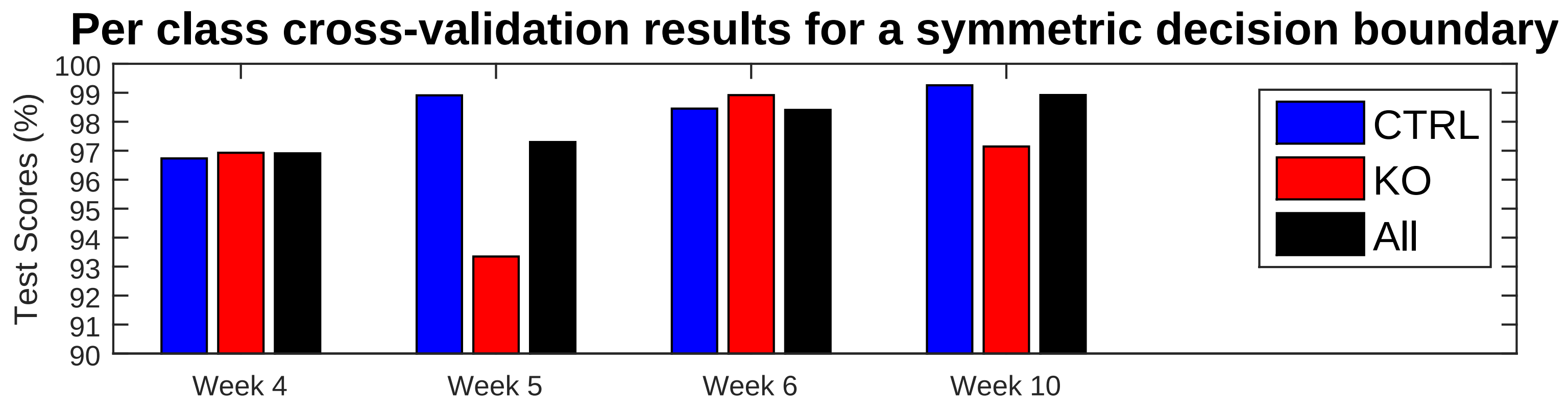
Classification - Details

L_1-regularization: encourages sparsified features by imposing an exponential prior on the distribution of the data. This particular procedure induces sparse features and the results are of straightforward interpretation. Below, the train and test cross-validation scores:

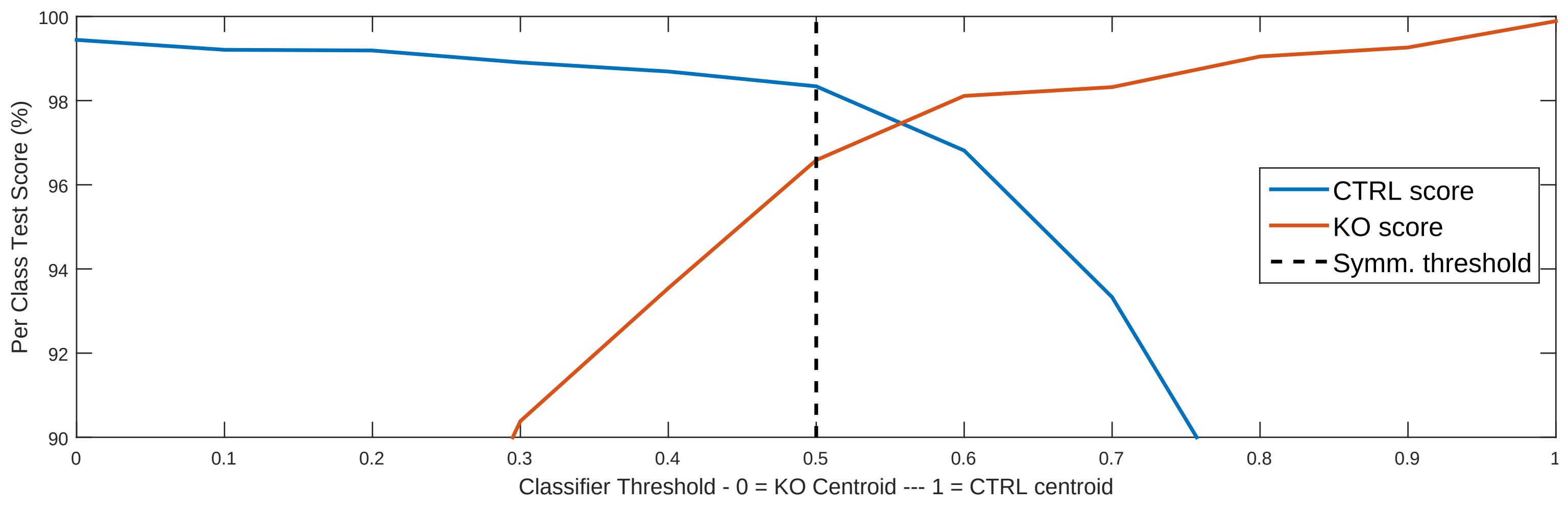


Random Forests: are an ensemble learning approach that compensates for correlations known to unstabilize decision trees. A set of trees is trained on bootstrapped training sets, and the trained trees vote for the best classification of a test sample.

Linear Discriminant (i) - Select ~300 features of maximum relative variance. (ii) - Center the data and run 60 dimensional PCA projection. (iii) - Find the projection vector maximizing the separation between classes. The cross-validation below is run by bootstrapping 20 vs. 92 (18%) partitions. We analyze the repartition of errors depending of the labeling of the test samples.



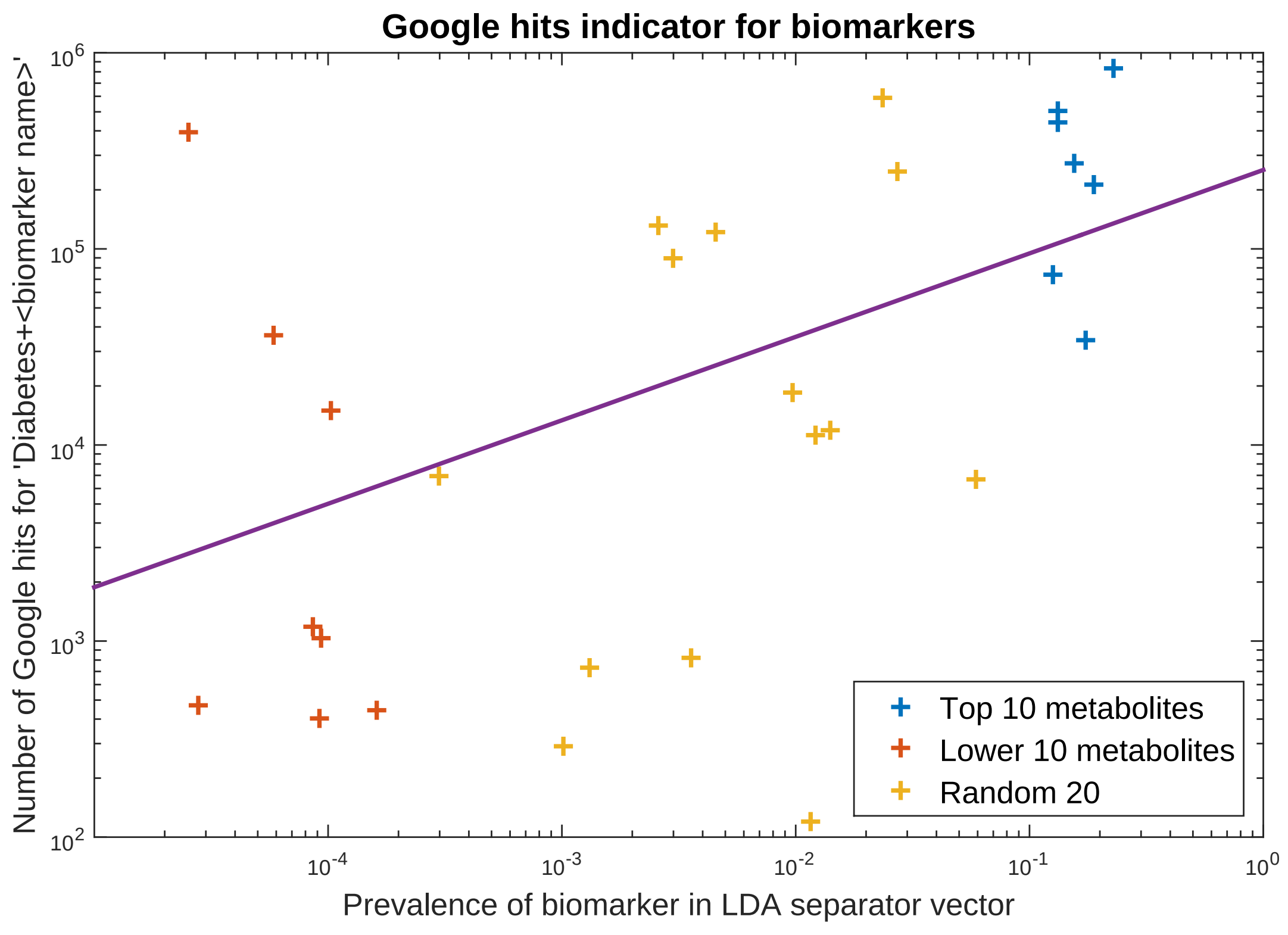
Threshold Biasing The common LDA yields a symmetric decision boundary. In the medical domain, it is interesting to bias this threshold to reduce false negatives at the cost of more false positives. The analysis charted below indicates trade-off results. A better pick than 0.5: 0.7 yielding 98% for KO and 94% for CTRL.



Biological Feature Extraction & Validation

Compound \ Rank	L_1	RF	LDA
L-Rhamnose	*	1	82
UDP-glucosamine	2	9	2
Cholic Acid	3	36	48
Cortisol	*	3	49
UDP-glucose	4	12	36
Ajugalactone	1	380	1
Uridine	5	85	3

As a pre-expertise validity check of the biomarkers found, we compare the importance of biomarkers in a classifier (LDA) against the number of google hits for "<Biomarker name>+diabetes".



Conclusions

The breadth-to-depth approach among known classification models has allowed us to extract valuable information. We isolated biomarkers of significance to establish reliably the incipience of type 2 diabetes.

The authors hope that this work will drive interest for further analytics in the biomedical scientific community, in validation, continuance, and in building denser datasets that will help shed light on the statistical significance of this work and future ones.