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.

agazzian, vdeo@stanford.edu

December 11th, 2015

ABSTRACT

This projects investigates automated diagnosis possibilities from metabolite micro-array measurements in liver cells, at an incipient stage of the disease. L_1-regularized logistic regression, Fisher Discriminant Analysis and Random Forests classifiers are studied, yielding promising results and hindsight on biomarker significance.

1. INTRODUCTION

Diabetes Mellitus is one of today's most common chronic diseases, affecting approximately 382M of people worldwide in 2013 [1]. The disease, resulting in serious cardiovascular complications, is divided into two categories. Type 2 diabetes, accounting for approximately 90% of the cases, is a gradually developing form of the disease, and can be prevented if diagnosed at an early stage. Motivated by the inherently metabolic character of the disease, the Maechler Lab at the University Hospital of Geneva is investigating a novel form of early stage diagnosis, searching for quantitative biological markers in a patient's liver cells. The liver is responsible for carbohydrate metabolism and is therefore expected to be affected first by incipient diabetes.

This project aims to analyze the data conveyed from the Maechler Lab and tackle the issues of early-stage diagnosis and biomarker identification. Specifically, the algorithm input is the result of micro-array experiments of metabolite concentrations in both sick and healthy mice liver cells. We apply different combinations of preprocessing and supervised classification algorithms, for the purpose of automated diagnosis of new data points. Furthermore, we aim to extract the most significant features indicating the presence of the disease through inspection of the best performing classifiers.

2. RELATED WORK

Micro-array and metabolite profiling experiments have been an object of interest in the machine learning literature ever since they appeared in cell biology. The data is high-dimensional, and therefore heavily regularized approaches often become the methods of choice for these problems [2]. For feature extraction and preprocessing of micro-array data, the most recurring algorithms in the literature are PCA [3], information gain, correlation analysis, and False Discovery Rate thresholding [4].

For the classification, widely applied methods include Support Vector Machines (SVM) with different kernels, Fisher Discriminant Analysis (FDA), Logistic Regression (LogReg), k-Nearest-Neighbors classifiers, Decision Trees and Random Forests [5]. In most cases, these can be complemented with a regularization procedure [2], encouraging model sparsity and providing helpful preprocessing to high dimensional and correlated data.

Given the wide range of high dimensional problems in the computational biology and high-throughput data analysis framework, there is no unique state-of-the-art pipeline for feature selection and data classification clearly outperforming others [5]. Therefore, the approach of this project will be to test a wide range of combinations of standard approaches and select the best ones for a more thorough analysis.

3. Dataset description

The database this project uses was jointly created by the Maechler Lab at UNIGE and the Zamboni Lab at ETHZ, and was transmitted to the authors under a confidentiality agreement. A colony of 56 mice was divided into two subgroups of 24 and 32 elements. The first group mice were genetically modified (knockout, KO) to induce diabetes, and the second group mice were kept as control (CTRL). Out of each group, samples of 6 to 8 mice were taken (without replacement) at weeks 4, 5, 6 and 10 of life (Table 1). For each selected mice, 2 samples of liver cells are processed through mass spectrometry [6] and the concentrations of metabolites are estimated (not uniquely but up to mass equivalence), for a total of 756 features per sample.

Type	Week 4	Week 5	Week 6	Week 10 (post-symptoms)	Total
CTRL	16	14	18	16	64
KO	14	10	12	12	48

Table 1: Dataset distribution, week of sampling and sick (KO) or healthy (CTRL) labels.

Our training set is the array of spectrometric measurements $(x^{(i)})_{1\leq i\leq 112} \in \mathbb{R}^{756}$, with labels giving the week of sampling $w^{(i)} \in \{4,5,6,10\}$ and the knockout state of the mice $y^{(i)} \in \{0,1\}$ for each of the $p = 112$ sampled liver samples. We note X the design matrix and Y the binary label vector.

Since the number of features is larger than the number of samples, data preprocessing is an important step to avoid critical overfits. Dimensionality reduction algorithms were applied prior to data classification:

PCA We compute the most significant n eigenvectors of the empirical covariance matrix $C = X^T X$, where n is selected by cross validation (Figure 1, left), and project the data in the span of these principal component vectors $e_1, \dots, e_n$: $X_{\text{PC}} = X \times [e_1 \ \dots \ e_n]$.

Correlation Thresholding We estimate the Pearson Correlation Coefficient $\rho(Y, X_{\cdot,j}) = \dfrac{\widehat{\text{Cov}}(Y, X_{\cdot,j})}{\hat{\sigma}_Y \hat{\sigma}_{X_{\cdot,j}}}$. The n features with highest absolute value of the correlation coefficient are selected, with n chosen by cross-validation. This method has been preferred over Mutual Information Thresholding, for the predictors being real-valued.

L_1 regularized logistic regression This classifier is applied, and the induced model parameter sparsity [7] is used as feature selector for a further classification by another algorithm.

4. Methods

Methodology The dataset was used to train different supervised classification algorithms, whose quality was quantified through their test scores in either leave-20-out or .632-bootstrap cross-validations. The best performing classification algorithms were used thereafter for quantitative and qualitative biomarker significance assessment.

Fisher Discriminant Analysis A Fisher Linear Discriminant (FDA) [8] was implemented hands-on as a starting point to assess the potential of the dataset, and then compared to the built-in matlab LDA and SVM with default configurations. The data is preprocessed using PCA (section 3), after which the first Fisher vector, normal the maximally separating hyperplane, is computed. The offset of the decision boundary between the two classes may be selected a posteriori, in order to adjust the false positive-false negative ratio to meet requirements.

Support Vector Machines classifiers were attempted with a variety of kernels over PCA-preprocessed data. The linear kernel yields preliminary results of the order of the FDA, although less stable over extensive test runs. Polynomial, Gaussian and Sigmoid kernels did not seem adapted to the dataset, leading to less satisfactory results.

REGULARIZED LOGISTIC REGRESSION Logistic regression -the generalized linear model with conditional probability $p(y|x)$ Bernoulli-distributed- is a natural choice for a 2-class linear classifier. The high dimensionality of the data at hand suggests the application of regularization procedures. This is equivalent to setting a Bayesian prior $p_\alpha(\theta)$ on the parameters of the model, and results in a modified log likelihood function:

$$l(\theta,\lambda) = \sum_i \log p(y^{(i)}|x^{(i)},\theta) + \lambda \log p(\theta) \ ,$$

where the regularization parameter λ is chosen by cross-validation. Most popular choices of prior distribution are $\mathrm{Exp}(\alpha)$ and $\mathcal{N}(0,\alpha^2 I)$, resulting respectively in so-called L_1 and L_2 regularizations. L_1 regularization is attractive in this framework as it induces sparsity in the model parameters [7]. A weighted choice of L_1 and L_2, called *Elastic Net* [9], has also been implemented. This learning algorithm was implemented using the python library `scikitlearn` [10].

RANDOM FORESTS Random forest (RF) classifier [11] is an ensemble learning approach based on decision trees. The latter sequentially divide the training data into subsets on single-decision rules that maximize class separation. As for Tree Bagging, RF classifiers bootstrap the training set and construct a decision tree for every bootstrapped sample, reducing the high variance of the decision tree classifier approach. RF differentiates form Tree Bagging by performing each split over a randomly chosen subset of features, of typical size $\sqrt{\#\text{features}}$. This randomized bootstrapping prevents correlated trees in the forest and improves accuracy. Interestingly, RF are invariant to rescaling of data, but tend to extract only a small subset of discriminative variables when features are highly correlated. This learning algorithm was implemented using the python library `scikitlearn` [10].

5. RESULTS AND DISCUSSION

PARAMETER SELECTION For feature extraction and regularized classifications, some hyperparameters must be picked to optimize the expected score of the algorithm. This is done by maximizing the CV-score over the full range of possible parameter combinations on the training set with a repeated leave-20-out procedure, which has been chosen because of the relatively small dataset size. This includes the regularization parameter for L_1-LogReg and the dimension of the PCA subspace for FDA. The resulting validation curves are displayed in Figure 1.

GENERAL RESULTS The scores for some of the preprocessor-classifier algorithms that have been applied in this analysis are reported in table 2. Out of the considered methods, those with best performance were studied in more detail, with a preference for methods allowing a direct ranking of discriminative features in order to facilitate biomarker extraction.

Method	Preproc.	CV score	Sparsity
LDA	PCA	97.4%	0%
Linear SVM	PCA	96.1%	0%
Quadratic SVM	PCA	92.7%	0%
L_1-penalized Logistic Reg.	CorrThresh	97.1%	90.6%
L_2-penalized Logistic Reg.	CorrThresh	96.1%	80.4%
ElNet-penalized Linear Reg.	CorrThresh	83.1%	54.4%
Random Forests	L_1-LogReg	93.0%	86.4 %

Table 2: Leave-20-out cross-validation scores for preprocessor-classifier pairs tried on the dataset.

MODEL EVALUATION The learning procedures listed in Section 4 were evaluated with different cross validation procedures: 0.632-bootstrap, 10-fold, leave-one-out and leave-20-out cross-validations. Detailed results are shown

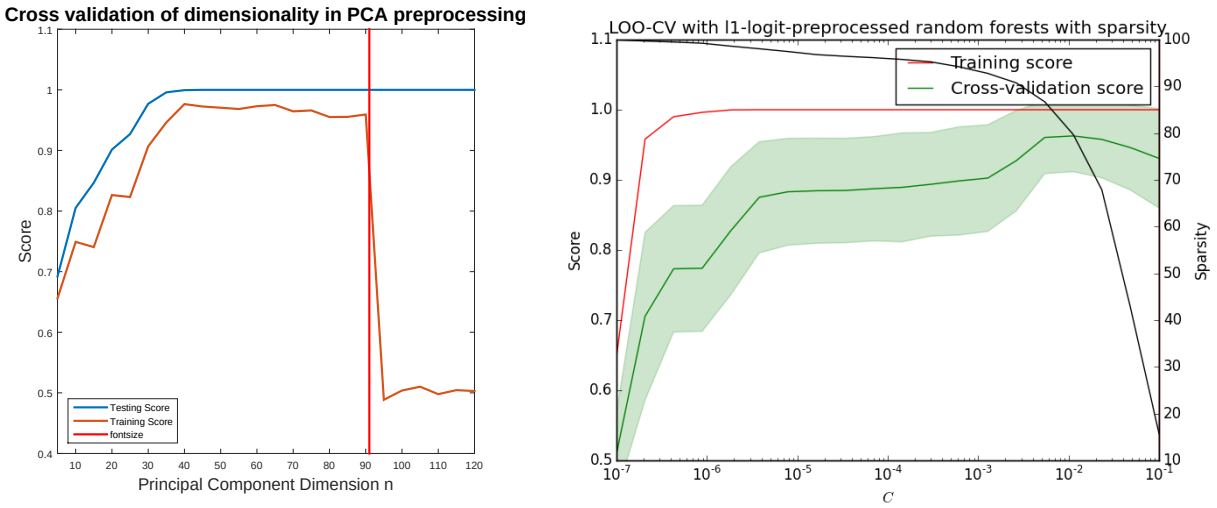


Figure 1: Left: Cross-validation of the number of PC dimensions in the FDA algorithm. The training score ceils at 1, whereas the testing score drops to 0.5 when dimensionality is higher than (number_of_points - 1) for separability reasons (red vertical line). Right: Cross-validation of the regularization parameter C in the L_1-LogReg preprocesser/RandomForest classifier algorithm. Colored areas correspond to the standard deviation on the errors. The black solid line represents the sparsity of model parameters as a function of the regularization parameter. Both cross-validations are realized in a repeated 300 times leave-20-out fashion.

for PCA preprocessing with FDA classifier in table 3, and for the two other algorithms in table 4. Cross-validations have been implemented manually in python and Matlab.

<table>
<tr><td>Total tests
6000</td><td>Actual
CTRL</td><td>Actual
KO</td><td colspan="2">Prevalence
57.04%</td></tr>
<tr><td>Predicted
CTRL</td><td rowspan="2">3342</td><td rowspan="2">88</td><td>Positive predictive
value: 97.41%</td><td>False discovery
rate: 2.59%</td></tr>
<tr><td>Predicted
KO</td><td>False omission
rate: 3.10%</td><td>Negative Predictive
value: 96.90%</td></tr>
<tr><td rowspan="2">Accuracy
97.19%</td><td>True positive
rate: 97.67%</td><td>False positive
rate: 3.44%</td><td>Positive likelihood
ratio: 28.39</td><td rowspan="2">Diagnostic odds
ratio: 1,176</td></tr>
<tr><td>False negative
rate: 2.33%</td><td>True negative
rate: 96.56%</td><td>Negative likelihood
ratio: 0.024</td></tr>
</table>

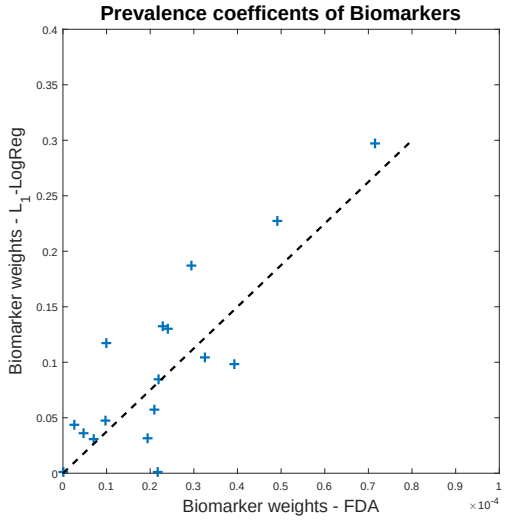
Table 3: Detailed confusion matrix for the Linear Discriminant classifier

Algorithm	True Positive	False Positive	True Negative	False Negative	Accuracy
L_1-reg Logistic Regression + Random Forest	3346	359	60	2235	**93.02%**
Correlation thresholding + L_1-reg Logistic Regression	3363	97	75	2467	**97.14%**

Table 4: Reduced confusion matrix for the L_1-LogReg and Random Forest Classifiers

FEATURE EXTRACTION AND BIOMARKER IDENTIFICATION The selected algorithms FDA, L_1-LogReg and Random Forests provided us with the ability to extract the importance of each feature in the discrimination between the two classes. We validate the extraction of these biomarkers by measuring the stability of their prevalence coefficients across algorithms. Figure 2 and the associated table show the stability of biomarker prevalence. The comparison between L_1-LogReg and FDA is most natural because the weight of the features are quantified by the corresponding component of the vector normal to the maximally separating hyperplane.

It is also a key point to verify our best found biomarkers against the medical knowledge as reflected by the current scientific literature. Preliminary research on the matter shows a consistent correlation between biomarker rankings and the amount of available scientific literature about diabetes mentioning the biomarker.



Compound \ Rank	L_1-LogReg	RF	LDA
Ajugalactone	1	39	1
UDP-glucosamine	2	1	2
Cholic Acid	3	21	48
UDP-glucose	4	12	36
Uridine	5	18	3

Figure 2: Left: Prevalence coefficients of the 17 most significant biomarkers in the L_1-LogReg algorithm versus their FDA prevalence. The dashed black line is provided as a guide to the eye. Right: rankings by prevalence of a selected subset of biomarkers across all three classifiers.

6. CONCLUSION AND FUTURE WORK

This work has addressed the problem of supervised classification and biomarker identification in metabolic profiling data of mice liver cells depending on the presence of Type II diabetes. The high dimensionality of data requires testing different preprocessors and learning algorithms for supervised classification. The preprocessor/classifier combinations that lead to the highest cross-validation scores are PCA/FDA, L_1-regularized Logistic Regression/Random Forest Classifier and Correlation thresholding/L_1-regularized Logistic Regression. For each of these three most successful methods, the features with the highest influence on the classifier have been ranked and compare favorably.

Possible directions for future work include the application of methods integrating prior biological knowledge -such as the known topology of the metabolic graph- for example through overlapping sparse group lasso or preconditioned PCA techniques.

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.

REFERENCES

[1] Y. Shi and F. B. Hu, "The global implications of diabetes and cancer," *The Lancet*, vol. 383, no. 9933, pp. 1947 – 1948, 2014.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.

[3] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[4] J. Quackenbush, "Computational analysis of microarray data," *Nature reviews genetics*, vol. 2, no. 6, pp. 418–427, 2001.

[5] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," vol. 20, no. 15, pp. 2429–2437, 2004.

[6] T. Fuhrer, D. Heer, B. Begemann, and N. Zamboni, "High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection–time-of-flight mass spectrometry," *Analytical Chemistry*, vol. 83, no. 18, pp. 7074–7080, 2011. PMID: 21830798.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[8] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[9] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] T. K. Ho, "The random subspace method for constructing decision forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 8, pp. 832–844, 1998.