Methods for class prediction with high-dimensional gene expression data

Jana Šilhavá*
Department of Computer Graphics and Multimedia
Faculty of Information Technology
Brno University of Technology
Božetěchova 2, 612 66 Brno, Czech Republic
silhava@fit.vutbr.cz

Abstract

An increasing amount of genomic data has become available. The work deals with class prediction with high-dimensional gene expression data. Combining gene expression data with other data can improve the prediction of disease prognosis. The main part of the work is aimed at combining gene expression data with clinical data. We use logistic regression models that can be built through various regularized techniques. Generalized linear models enable us to combine models with different structure of data. It is shown that such a combination may yield more accurate predictions than those obtained based on the use of gene expression or clinical data alone. Suggested approaches are not computationally intensive.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Knowledge acquisition; I.2 [**Artificial Intelligence**]: Applications and Expert Systems; I.2.1 [**Applications and Expert Systems**]: Medicine and science

Keywords

predictive classification, generalized linear models, model evaluation, high-dimensional data, combining of heterogenous data, gene expression, clinical data

1. Introduction

Microarray class prediction [1] is an important application of gene expression data in biomedical research. Microarray experiments monitor gene expressions associated with different phenotypes. Prediction of prognosis based on different phenotypes is challenging due to relatively small number of samples, high-dimensionality of gene expression data and a low signal to noise ratio, which influence a quality of prediction model construction and predictions. Combining gene expression data with other relevant data may add valuable information and can generate more accurate predictions.

In this paper, we combine gene expressions with clinical data. Clinical data is heterogeneous and measures various entities (e.g. lymph nodes, tumor size), while gene expression data is homogeneous and measures gene expressions. We assume that the combination of gene expressions with clinical data can involve complementary information, which may yield more accurate (disease outcome) predictions than those obtained based on the use of gene expression or clinical data alone. In literature, there are studies aimed at integrative prediction with gene expression and clinical data, e.g. see [10] and [8]. On the other side, redundant and correlated data can have contradictory impact on prediction accuracy.

Methods combining biomedical data can be divided into categories depending on the stage of integration [2]. We propose an approach that combines data at the stage of late integration. We use logistic regression models that can be built through various regularized techniques and can be applied to high-dimensional data as well. A key to combining gene expression and clinical data is a framework of generalized linear models (GLMs), which is offered for many statistical models.

Simple logistic regression has been widely used with clinical data in clinical trials to determinate the relationship between variables and outcome and to assess variable significance. Clinical data is usually low-dimensional because gene expression data sets include just a few clinical variables. That is why we use simple logistic regression models with clinical data and regularized logistic regression models with high-dimensional gene expression data.

According to [10], the penalized estimation methods for integrative prediction and gene selection are promising but computationally intensive. We experimented with R packages 'mboost', 'glmnet', 'grplasso', 'glmpath' that regularize high-dimensional data with penalties and at the same time these statistical models were developed for fitting in GLM framework. R packages 'mboost' and 'glmnet' performed very well and models fitting were not time-consuming. We built the algorithms from these R packages in our classifiers that combine gene expression and clinical data. In case of R package 'mboost', we use a version of boosting that utilizes componentwise linear least squares (CWLLS) as a base procedure, that closely

*Recommended by thesis supervisor: Dr. Pavel Smrž

corresponds to fitting a logistic regression model [4]. R package 'glmnet' is an application of elastic net [14], which is a regularization and variable selection method that can include both L_1 and L_2 penalties. The algorithms for fitting GLMs with elastic net penalties were developed by [5], which also described logistic regression model with elastic net penalties. Our approaches that combine gene expression and clinical data improve prediction performances and are not computationally intensive.

The rest of this paper is organized as follows: The relevant models and the proposed approaches that combine data are described in Section 2. Section 3 presents evaluation methodology. Simulations are performed with several generated data sets in redundant and nonredundant setting together with some tests applied publicly abailable breast cancer data sets in Section 4. It also includes a comparison of execution times of applied approaches. This paper is concluded in Section 5.

2. Methods

Notation: Let $\mathbf{X}$ be $p \times n$ gene expression data matrix with an element x_{ij}, p genes and n samples. Let $\mathbf{Z}$ be $q \times n$ clinical data matrix with an element z_{ij}, q clinical variables and n samples. $\mathbf{y}$ is $n \times 1$ response vector with an element y_i and with ground truth class labels $\mathbf{y} \in \{A, B\}$, where A and B can denote poor and good prognosis. In the following text, the upper indexes X, Z, distinguish from variables with gene expression data, clinical data.

2.1 Generalized Linear Models

GLMs [**?**] are a group of statistical models that model the response as a nonlinear function of a linear combination of the predictors. These models are linear in the parameters. The nonlinear function (link) is the relation between the response and the nonlinearly transformed linear combination of the predictors. We employ GLMs in data combining due to nice shared properties such as linearity. GLMs are generalization of normal linear regression models and are characterized by the following features:

1. Linear regression model:

$$\eta_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \epsilon_i, \tag{1}$$

where $i = 1, \ldots, n$. β are regression coefficients and ϵ is a random mean-zero error term.

2. The link function:

$$g(y_i) = \eta_i, \tag{2}$$

where g is a link function, $i = 1, \ldots, n$. η_i is a linear predictor. Respectively $y_i = g^{-1}(\eta_i)$, where g^{-1} is an inverse link function.

2.2 Logistic Regression Model

We use linear logistic regression model with clinical data. The linear logistic regression model is an example of GLM, where the response variable y_i is considered as a binomial random variable p_i and the link function is logistic:

$$\eta = \log\left(\frac{p}{1-p}\right). \tag{3}$$

Logistic regression model with clinical data can be described with the following equation:

$$g(y_i) = \eta_i = \beta_0^Z + \sum_{l=1}^{q} \beta_l^Z z_{il}, \tag{4}$$

where $i = 1, \ldots, n$. g is the link function (3). y_i or p_i are outcome probabilities $\mathbb{P}(y_i = A | z_{i1}, \ldots, z_{iq})$.

2.3 Boosting Model

A boosting with componentwise linear least squares as a base procedure is applied to gene expression data. A linear regression model (1) is considered again. A boosting algorithm is an iterative algorithm that constructs a function $\hat{F}(x)$ by considering the empirical risk:

$$n^{-1} \sum_{i=1}^{n} L(y_i, F(x_i)). \tag{5}$$

$L(y_i, F(x_i))$ is a loss function that measures how close a fitted value $\hat{F}(x_i)$ comes to the observation y_i. In each iteration, the negative gradient of the loss function is fitted by the base learner. The gradient descent is an optimization algorithm that finds a local minimum of the loss function. The base learner is a simple fitting method which yields as estimated function:

$$\hat{f}(\cdot) = \hat{f}(\mathbf{X}, \mathbf{r})(\cdot), \tag{6}$$

where $\hat{f}(\cdot)$ is an estimate from a base procedure. The response $\mathbf{r}$ is fitted against $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

The functional gradient descent (FGD) boosting algorithm, which has been given by [6] is as follows [4]:

1. Initialize $\hat{F}^{(0)} \equiv \sum_{i=1}^{n} L(y_i, a) \equiv \bar{y}$. Set $m = 0$.
2. Increase m: $m = m + 1$. Compute the negative gradient (also called pseudo response), which is the current resudial vector:
   $r_i = -\frac{\partial}{\partial F} L(y, F)|_{F=\hat{F}^{(m-1)}(x_i)}$
   $r_i = y_i - \hat{F}^{(m-1)}(x_i),\ i = 1, \ldots, n.$
3. Fit the residual vector $(r_1, \ldots, r_n)$ to $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ by base procedure (e.g. regression)
   $(\mathbf{x}_i, r_i)_{i=1}^n \xrightarrow{\text{base procedure}} \hat{f}^{(m)}(\cdot),$
   where $\hat{f}^{(m)}(\cdot)$ can be viewed as an approximation of the negative gradient vector.
4. Update $\hat{F}^{(m)}(\cdot) = \hat{F}^{(m-1)}(\cdot) + \nu \cdot \hat{f}^{(m)}(\cdot)$, where $0 < \nu < 1$ is a step-length (shrinkage) factor.

5. Iterate steps 2 to 4 until $m = m_{stop}$ for some stopping iteration m_{stop}.

The CWLLS base procedure estimates are defined as:

$$\hat{f}(\mathbf{X}, \mathbf{r})(x) = \hat{\beta}_{\hat{s}} \hat{x}_{\hat{s}}, \quad \hat{s} = \arg\min_{1 \le j \le p} \sum_{i=1}^{n} (r_i - \beta_j x_{ij})^2,$$
$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} r_i x_{ij}}{\sum_{i=1}^{n} (x_{ij})^2}, \quad j = 1, \ldots, p.$$

$\hat{\beta}$ are coefficient estimates. $\hat{s}$ denotes the index of the selected (the best) predictor variable in iteration m. For every iteration m, a linear model fit is obtained.

BinomialBoosting [4], which is the version of boosting that we utilize, use the negative log-likelihood loss function: $L(y, F) = \log_2(1+e^{-2yF})$. It can be shown that this population minimizer has the form [4]: $F(x_i) = \frac{1}{2} \log\left(\frac{p}{1-p}\right)$, where p is $\mathbb{P}(y_i = A | x_{i1}, \ldots, x_{ip})$ and relates to logit function, which is analogous to logistic regression.

2.4 Elastic Net Model

We also use elastic net [14] with gene expression data. The linear regression model (1) is considered again. The elastic net optimizes the following equation with respect to β [7]:

$$\hat{\beta}(\lambda) = \arg\min \frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \mathbb{P}_\alpha(\beta),$$

where: $\mathbb{P}_\alpha(\beta) = (1-\alpha)\frac{1}{2}\|\beta\|^2_{L_2} + \alpha\|\beta\|_{L_1}$ or $\mathbb{P}_\alpha(\beta) = (1-\alpha)\frac{1}{2}\beta_j^2 + \alpha|\beta_j|$ is the elastic net penalty.

$$\mathbb{P}_\alpha(\beta) = \begin{cases} L_1 \text{ penalty} & \text{if } \alpha = 1, \\ L_2 \text{ penalty} & \text{if } \alpha = 0, \\ \text{elastic net penalty} & \text{if } 0 < \alpha < 1. \end{cases} \quad (7)$$

In our case, elastic net builds logistic regression model with elastic net penalties. The regularized equation is fitted by maximum (binomial) log-likelihood and solved by coordinate descent, see [7]. The coordinate update has the form:

$$\hat{\beta}_j \leftarrow \frac{S(\frac{1}{n}\sum_{i=1}^{n} x_i r_{ij}, \lambda\alpha)}{1 + \lambda(1-\lambda)} = \frac{S(\beta_j^*, \lambda\alpha)}{1 + \lambda(1-\lambda)}, \quad (8)$$

where r_{ij} is the partial residual $y_i - \hat{y}_{ij}$ for fitting $\hat{\beta}_j$ and $S(\kappa, \gamma)$ is the soft-thresholding operator, which takes care of the lasso contribution to the penalty. More detailed description is given in [5].

A simple description of CCD algorithm for elastic net is as follows [7]:

The authors assume that the x_{ij} are standardized:

$$\sum_{i=1}^{n} x_{ij} = 0, \ \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2.$$

- Initialize all the $\hat{\beta}_j = 0$.
- Cycle around till convergence and coefficitents stabilize:
  1. Compute the partial residuals: $r_{ij} = y_i - \hat{y}_{ij} = y_i - \sum_{k \ne j} x_{ik}\beta_k$.
  2. Compute the simple least squares coefficient of these residuals on jth predictor: $\beta_j^* = \frac{1}{n}\sum_{i=1}^{n} x_{ij} r_{ij}$.
  3. Update $\hat{\beta}_j$ by soft-thresholding: $\hat{\beta}_j \leftarrow S(\beta_j^*, \lambda)$, which equals (8).

2.5 Combining Gene Expression and Clinical Data

In GLMs, the linear models are related to the response variable via a link function (2). For binary data, we expect that the responses y_i come from binomial distribution. Therefore, logit link function is used in all models with clinical and gene expression data. η_i is a linear model, which is a linear part of logistic regression and a linear regression model in boosting with CWLLS described in Subsection 2.3. We combine the data by summing the linear predictions of clinical and gene expression data:

$$\eta_i = \eta_i^Z + \eta_i^X. \quad (9)$$

According to the additivity rule that is valid for linear models, it is possible to sum the linear models:

$$\eta_i = \beta_0^Z + \sum_{l=1}^{q} \beta_l^Z z_{il} + \sum_{j=1}^{p} \beta_j^X x_{ij}. \quad (10)$$

Then the inverse link function g^{-1}, which is the inverse logit function, is applied to the sum of linear predictions η_i :

$$g^{-1}(\eta_i) = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \quad (11)$$

For better readibility in results, we denote this approach LOG+B.

Similarly we combine logistic regression and regularized logistic regression models from elastic net. For better readability in results, we denote this approach LOG+EN.

3. Evaluation

Monte-Carlo cross-validation (MCCV) was used as a validation strategy. MCCV generates learning set in that way that the learning data sets are drawn out of $\{1, \cdots, n\}$ samples randomly and without replacement. The test data set consists of the remaining samples. The random splitting in non-overlapping learning and test data set was repeated 100-times. The splitting ratio of training and test data set was set to 4 : 1. Responses consisted of predicted class probabilities were measured with the area under the ROC curve (AUC).

	1.	2.	3.	4.
μ_Z, μ_X	0, 0	1, 0.25	0.5, 0.5	1, 0.5
Method	(no power)	$\mu_Z > \mu_X$	$\mu_Z < \mu_X$	(strong p.)
LOG	0.55 ± 0.05	0.65 ± 0.06	0.56 ± 0.05	0.65 ± 0.06
B	0.51 ± 0.05	0.60 ± 0.05	0.72 ± 0.05	0.72 ± 0.05
EN	0.50 ± 0.03	0.60 ± 0.05	0.74 ± 0.05	0.74 ± 0.05
LOG+B	0.53 ± 0.05	0.69 ± 0.04	0.73 ± 0.04	0.79 ± 0.04
LOG+EN	0.55 ± 0.04	0.68 ± 0.05	0.75 ± 0.04	0.81 ± 0.04

Table 1: Simulated data: Non-redundant data sets. AUCs evaluated over 100 MCCV iterations.

	1.	2.	3.	4.
μ_Z, μ_X	0, 0	1, 0.25	0.5, 0.5	1, 0.5
Method	(no power)	$\mu_Z > \mu_X$	$\mu_Z < \mu_X$	(strong p.)
LOG	0.49 ± 0.05	0.94 ± 0.02	0.78 ± 0.04	0.94 ± 0.02
B	0.51 ± 0.05	0.71 ± 0.04	0.98 ± 0.01	0.98 ± 0.01
EN	0.48 ± 0.04	0.72 ± 0.04	0.97 ± 0.01	0.97 ± 0.01
LOG+B	0.51 ± 0.05	0.94 ± 0.02	0.96 ± 0.02	0.98 ± 0.01
LOG+EN	0.48 ± 0.05	0.95 ± 0.02	0.98 ± 0.01	0.99 ± 0.01

Table 2: Simulated data: Redundant data sets. AUCs evaluated over 100 MCCV iterations.

4. Results

We tested the described approaches with simulated and publicly available breast cancer data sets. The performances of individual models were evaluated as well because of a comparison of the models. We performed experiments in R environment[1]. We used glm function from 'base' package to fit the logistic regression models with clinical data; 'mboost' and 'glmnet' packages to fit the logistic regression models with gene expression data.

4.1 Simulated data sets

We tested approaches with non-redundant and redundant data sets. We generated simulated data sets through the use of R script available in [3]. In case of redundant sets, microarray and clinical variables are generated using exactly the same model. Such variables discriminate classes in the same way and giving redundant information. In case of non-redundant sets, the observations are assumed to form two distinct subgroups [3]. Then we considered different predictive powers for the clinical variables μ_Z and different predictive powers for the microarray variables μ_X. In present simulations, $\mu_Z = 0$ denotes no power, $\mu_Z = 0.5$ moderate power and $\mu_Z = 1$ strong power for Z. Similarly $\mu_X = 0$, 0.25, 0.5 for X. Difference in μ_Z and μ_X ranges compensates for ranges of predictor values for microarray and clinical variables.

The following Tables 1 and 2 display selected results of LOG+B and LOG+EN for different predictive powers of Z and X. AUCs from test data sets (including mean AUCs and standard deviations) were evaluated over 100 MCCV iterations. In case of non-redundant data sets, LOG+B and LOG+EN increase AUCs. LOG+B and LOG+EN have a good performance on redundant data sets as well.

4.2 Publicly avalilable data sets

We also evaluted the described approaches with publicly available breast cancer data sets. We used two data sets in this paper. The van't Veer data set [13] gives the expression levels of 22483 genes for 78 breast cancer patients. Based on existence of distant metastases, 34 of these samples are classified into the poor prognosis group, the rest 44 samples belong to the the good prognosis group. The used data set is prepared as described in [13] and is included in R package 'DENMARKLAB'. This data set includes 4348 resulting genes. Clinical variables are age, tumor grade, estrogen receptor status, progesterone receptor status, tumor size and angioinvasion. The second data set, which is the Pittman data set [11], gives the expression levels of 12625 genes for 158 breast cancer patients. According to recurrence of disease, 63 of these patients are classified into the poor prognosis group, the remaining 95 patients belong to the good prognosis group. Gene expression data was prepared with Affymetrix Human U95Av2 GeneChips. The data was pre-processed using packages 'affy' and 'genefilter' to normalize and filter the data. The genes that showed a low variability across all samples were cleared out. The resulting data set includes 8961 genes. Clinical variables are age, lymph node status, estrogen receptor status, family history, tumor grade and tumor size.

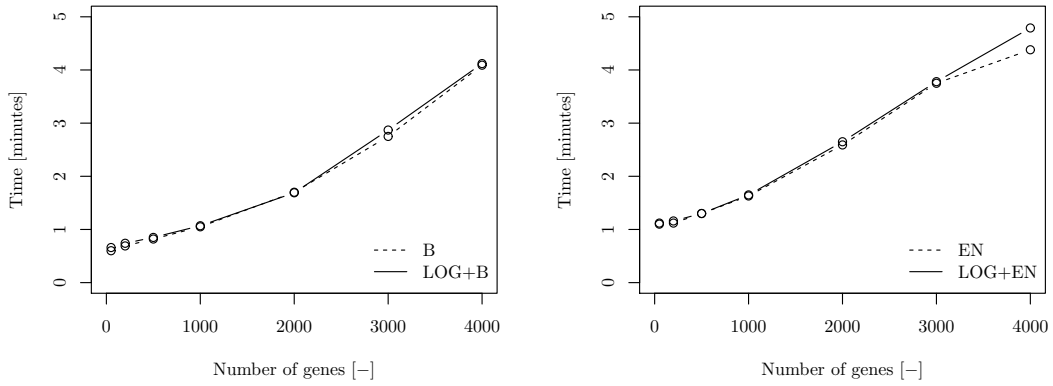
Table 3 shows AUCs of breast cancer data sets. AUCs from test data sets were evaluated over 100 MCCV iterations. Considering the results with the Pittman data set, the combined models have a positive effect on prediction performances and increase AUCs. The combined models, built with the data of van't Veer, do not improve AUC performances and it is better to use for prediction of prognosis clinical data alone. The conclusion with the van't Veer data set also corresponds with findings, e.g. [9]. The performances of the combined models are similar.

The execution time of the combined models is dominated by the execution times of the models built with high-dimensional data; therefore, we compared the execution times of the FGD boosting algorithm from the package 'mboost' (B and LOG+B) with the CCD algorithm from the package 'glmnet' (EN and LOG+EN). Figure 1 depicts this comparison. Increasing numbers of variables are on the horizontal axes, while total execution times for 100 MCCV iterations (in minutes) are on the vertical axes. The plots indicate that both methods need similar time to be computed. The execution times grow almost

[1] www.r-project.org

Data set	*Methods*				
van't Veer	LOG	B	EN	LOG+B	LOG+EN
	0.82 ± 0.10	0.65 ± 0.11	0.64 ± 0.12	0.79 ± 0.11	0.79 ± 0.11
Pittman	LOG	B	EN	LOG+B	LOG+EN
	0.67 ± 0.09	0.78 ± 0.08	0.77 ± 0.07	0.82 ± 0.08	0.80 ± 0.07

Table 3: Breast cancer data sets. AUCs evaluated over 100 **MCCV iterations.**



(a) B and LOG+B.

(b) EN and LOG+EN.

Figure 1: The comparison of computation times and their dependence on increasing number of variables. The graphs are drawn for 100 **MCCV iterations.**

linearly. Besides, FGD boosting grows with the number of boosting iterations (in our simulations $m_{max=700}$). A grid of 100 λ values is computed in each iteration of EN. The simulations were run on a standard PC (Intel T72500 Core 2 Duo 2.00 GHz, 2 GB RAM) and a 32-bit operating system.

5. CONCLUSION

In this work, we combine gene expression and clinical data to predict disease prognosis. We used logistic regression models built by different ways. GLMs enabled combining of these models. Two suggested approaches were evaluated with simulated and publicly available breast cancer data sets. Both approaches performed well and showed similar performances.

This apporach is extended in the thesis with pre-validation of models built with microarray and clinical data followed by weights calculation. Pre-validation of build models increases outcome prediction performances in cases of redundant data sets. The thesis includes more topics. We evaluated classifiers combining gene expression, clinical and SNP data. Our approach can be used with more than two various data sources. The predictive performance is influenced by the quality of data. The presented approaches can be used for determing additional predictive value of microarray data. Researchers do not pay too much attention to the clinical data which are usually included in microarray data set. However, clinical data can have higher predictive power than microarray data. We dealt with this topic.

An aim of class prediction studies is to identify a set of genes (signature) that can accurately classify prognosis of new data. It is interesting to compare selected features of described classifiers on different data sets. However, we did not expect too much because different gene signatures usually have very few genes in common [12]. We compared selected features of three feature selection methods evaluated with five breast cancer data sets. Boosting with CWLLS and elastic net classifiers in the setting with L1 penalty (the lasso) select the most similar gene sets.

Incorporating gene-to-gene relations and interactions into the class prediction process can increase prediction performance and can help to identify more interpretable genes. A way of gene-to-gene relations presentation is the similarity matrix, which is often used in connection with ontologies. We presented preliminary results with four variants of the feature selection method combining GO information with gene expression data. Microarray data gene ontology feature selection based on maximum gene cliques improved prediction performances.

Prediction accuracy of combination of microarray and other data depends on complementarity of these data sources. If data sources or data models are complementary, i.e. they contain some nonredundant information, combination of models leads to increased prediction accuracy. It could also be interesting to combine more than two or three different data together. The described approaches combining clinical and microarray data can combine more than two or three data. Integrating distinct and multiple information resources is and will be an important task in the future.

Acknowledgements. My research has been supported by the Technology Agency of the Czech Republic, project

TA01010931 – GenEx – System for support of the FISH method evaluation, and the operational programme 'Research and Development for Innovations' in the framework of the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070.

References

[1] D. Amaratunga and J. Cabrera. *Exploration and Analysis of DNA Microarray and Protein Array Data*. John Wiley & Sons, Hoboken, 2004.

[2] F. Azuaje. *Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine*. John Wiley & Sons, Singapore, 2010.

[3] A. L. Boulesteix, C. Porzelius, and M. Daumer. *Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value*. Bioinformatics 24, 1698-1706, 2008.

[4] P. Bühlmann and T. Hothorn. *Boosting Algorithms: Regularization, Prediction and Model Fitting*. Statist. Sci., 22, 477-505, 2007.

[5] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. *Pathways Coordinate Optimization*. Ann. Appl. Stat., 1, 302-332, 2007.

[6] J. H. Friedman. *Greedy Function Approximation: A Gradient Boosting Machine*. Ann. Statist., 29, 1189-1232, 2001.

[7] J. H. Friedman, T. Hastie, and R. Tibshirani. *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Journal of Statistical Software, 33 (1), 1-24, 2010.

[8] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor. *Predicting the Prognosis of Breast Cancer by Integrating Clinical and Microarray Data with Bayesian Networks*. Bioinformatics, 22 (14), 147-157, 2007.

[9] S. K. Gruvberger, M. Ringner, and P. Eden. *Expression Profiling to Predict Outcome in Breast Cancer: the Influence of Sample Selection*. Breast Cancer Res., 5(1), 23-26, 2003.

[10] L. Li. *Survival Prediction of Diffuse Large-B-Cell Lymphoma Based on both Clinical and Gene Expression Information*. Bioinformatics, 22(04), 466-471, 2006.

[11] J. Pittman, E. Huang, and H. Dressman. *Integrated Modeling of Clinical and Gene Expression Information for Personalized Prediction of Disease Outcomes*. Proc.Natl.Acad.Sci., 101(22), 8431-8436, 2004.

[12] C. Sotiriou and L. Pusztai. Gene-expression signatures in breast cancer. *N. Engl. J. Med.*, (360):790–800, 2009.

[13] L. J. van't Veer, H. Dai, and M. J. van de Vijver. *Gene Epression Profiling Predicts Clinical Outcome of Breast Cancer*. Nature, 530-536, 2002.

[14] H. Zou and T. Hastie. *Regularization and Variable Selection via the Elastic Net*. Journal of the Royal Statistical Society, Series B, 67, 301-320, 2005.

Selected Papers by the Author

Jana Šilhavá, Pavel Smrž. Additional Predictive Value of Microarray Data Compared to Clinical Variables. In *4th IAPR International Conference on Pattern Recognition in Bioinformatics*, pages 1–6, 2009.

Jana Šilhavá, Pavel Smrž. Gene Ontology Driven Feature Filtering from Microarray Data. In *Znalosti'2010*, pages 263–266, 2010.

Jana Šilhavá, Pavel Smrž. Improved Disease Outcome Prediction Based on Microarray and Clinical Data Combination and Pre-validation. In *Biomedical Engineering Systems and Technologies*, pages 36–41, 2010.

Jana Šilhavá, Pavel Smrž. Combining Gene Expression and Clinical Data to Increase Performance of Prognostic Breast Cancer Models. In *ICAART'2012 - 4th International Conference on Agents and Artificial Intelligence*, pages 589–594, 2012.