České vysoké učení technické v Praze
Fakulta elektrotechnická

Czech Technical University in Prague
Faculty of Electrical Engineering

Ing. Jiří Kléma, Ph.D.

Klasifikace dat z genových čipů založená na množinách genů

Set-level Microarray Classification

Summary

Molecular classification of biological samples based on their annotated gene expression profiles represents a natural task. Although there are several success stories, the problem is conceptually difficult for the sake of high cost of microarrays resulting in a low number of analyzed biological samples, a large number of genes being screened and measurement inaccuracy. These characteristics often cause overfitting. Classifiers do not sufficiently generalize and instead of the underlying relationships rather capture perturbations in training data. This problem can be minimized by regularization, i.e., introduction of additional knowledge. This talk focuses on utilization of prior molecular knowledge in order to extract features representing functionally related genes. The main issues are obvious: 1) how to merge genes into groups, 2) how to calculate their group expression and 3) how to select best extracted features for further learning. The overall goal is to maximize accuracy of resulting set-level molecular classifiers as well as their biological interpretability.

Souhrn

Molekulární klasifikace biologických vzorků na základě jejich anotovaných profilů genové exprese je přirozenou úlohou. Přes prokazatelné úspěchy jde však o úlohu obtížnou, zejména vzhledem k ceně genových čipů a nízkému počtu biologických vzorků, vysokému počtu sledovaných genů a nepřesnostem v měření. Úlohy s těmito charakteristikami často vedou k přeučení, kdy klasifikátory dostatečně nezobecňují a namísto základních vztahů popisují nahodilé vazby v trénovacích datech. Řešením je regularizace, tedy zavedení dodatečné znalosti. Tématem přednášky je využití apriorní molekulární znalosti k vytváření odvozených příznaků reprezentujících funkčně či jinak příbuzné množiny genů. Hlavní otázky jsou zřejmé: 1) jak skupiny genů tvořit, 2) jak počítat jejich skupinovou expresi a 3) jak vybrat vhodné odvozené příznaky před samotným učením. Cílem je maximalizovat přesnost molekulárních klasifikátorů založených na odvozených příznacích a také jejich srozumitelnost pro biology.

Klíčová slova

Umělá inteligence, strojové učení, učení s učitelem, klasifikace, přeučení, apriorní znalost, extrakce příznaků, bioinformatika, molekulární genomika, genová exprese, genová ontologie, metabolické a signální dráhy.

Keywords

Artificial intelligence, machine learning, supervised learning, classification, prior knowledge, feature extraction, bioinformatics, molecular genomics, gene expression, gene ontology, metabolic and signalling pathways.

Contents

1 Background

1.1 Supervised Learning

Supervised learning is one of the principal forms of machine learning with frequent practical application. In supervised learning a learner observes some example input-output pairs and learns a function that maps from input to output [32]. More precisely, the learner is provided with a training multiset of n training examples

$$T = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)\}$$

where $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{im})$ represents a random sample from the input space X of m independent variables called features and y_i is a value of random output variable y. Each y_i is supposed to be generated by an unknown function $y = f(\mathbf{x})$. The learner discovers an approximation h of the true function f. The approximation is called hypothesis, it is selected from a hypothesis space of possible functions $\mathcal{H}$.

When the output y is one of a finite set of values, i.e., the output variable is categorical, the learning problem is referred to as *classification*. When y is continuous, the learning problem is called *regression*. In this text we will further deal with classification only.

Learning is a search through the space of possible hypotheses for one that will perform well [32]. First, the well performing hypothesis shall be consistent and agree with all the training data, i.e. $\forall i = 1, \ldots, n\ h(\mathbf{x_i}) = y_i$. Second, it shall also generalize well and correctly assign the value of dependent variable for unseen examples. Although it may seem that $\mathcal{H}$ shall be as large as possible to guarantee that $f \in \mathcal{H}$ and no prior preference among hypotheses is needed, this approach contradicts a fundamental property of inductive inference: a learner that makes no a priori assumptions regarding the identity of the target function has no rational basis for classifying any unseen examples [27]. Consequently, one needs to resolve a fundamental trade-off between complex hypotheses that are consistent with training data but may overfit them and more regular hypotheses that do not agree with all the training outputs but may generalize better.

Learning from examples requires certain prior assumptions, called *inductive bias*. $\mathcal{H}$ is said to define a *hard bias* of the learner, hypotheses not belonging to $\mathcal{H}$ cannot be acquired. The actual form of the hard bias depends on the domains of the m attributes. To exemplify, $\mathcal{H}$ can be constrained to the functions representable as conjunctive logical formulas, decision trees, if-then rules or a linear classifier such as per-

ceptron. At the same time, there may exist a *soft bias* that gives a soft preference for one hypothesis over another. The soft bias can be implemented, e.g., in the form of a probability distribution over the hypothesis space. A less probable hypothesis can be acquired, but it requires stronger evidence, which means more training examples in favor of the hypothesis, to be learned. General soft bias strategies applicable across a wide range of domains are the smoothness assumption most explicitly used in nearest neighbor classifier, the related low-density assumption where classes are split by low density areas, or Occam's razor that prefers simple hypotheses over complex ones. When choosing a function from $\mathcal{H}$, the learner may minimize the total hypothesis cost defined as follows [32]:

$$Cost(h) = \frac{1}{n} \sum_{\{x_i, y_i\} \in T} L(y_i, h(x_i)) + \lambda \rho(h)$$

where L is a loss function defined as the amount of utility lost by predicting $h(x_i)$ instead of true y_i, $\rho(h)$ is a regularization term defining the soft bias and $\lambda \in R$ is a conversion rate between loss and regularization. Another approach to overfitting control without the immediate regularization penalty represents cross-validation.

In general setting, the function f does not need to be deterministic but stochastic and the learner approximates a conditional probability distribution $P(Y|X)$. Bayesian decision theory [5] provides an alternative view of the problem of finding the hypothesis function h that fits the stochastic target functions. The problem is posed as *risk minimization*. Empirical risk minimization seeks the function that best fits the training data while structural risk prevents overfitting through adding a regularization penalty that typically quantifies model complexity and can be seen as a variant of Occam's razor.

1.2 The Role of Prior Knowledge in Learning

The term prior knowledge refers to all information about the problem available in addition to the training data [34]. Only prior knowledge makes it possible to generalize from the training examples to novel test examples. The previous section mentions several examples of general prior knowledge applicable across a wide range of domains (smoothness assumption, low-density assumption, Occam's razor). The last and probably most widely used criterion prefers simple models. One way to keep the models simple is to reduce the feature space dimension by

means of feature selection or extraction [23]. In a typical dimensionality reduction scenario, there is no prior domain knowledge involved. In feature selection, the features empirically exhibiting little mutual information with the dependent variable are removed. In feature extraction, one often aims to reach the intrinsic dimensionality of a training set.

However, there are approaches that go beyond statistical significance and analysis of training data. Their key to successful learning rests in the selection of proper hard and soft biases stemming from the actual domain. Feature selection may benefit from the knowledge of features that are a priori denoted to be less relevant or irrelevant [2] or metadata on feature value domains [41]. There could also be prior feature relations, feature links could take values such as plausible, unknown or unlikely [9] and serve to constrain features with specific relationship to the currently used feature set (i.e., the current seed of a decision rule). Similar prior knowledge could also be incorporated using informative prior distributions such as in [15]. The cost of acquiring feature values can make an additional criterion for model construction [25]. The authors of [35] show that straightforward incorporation of domain knowledge in a form of binary recommendation significantly improves classification performance. [33] represents a more-sophisticated method that incorporates a human-built approximate predictive model into boosting. The learner is supposed to fit the training data as well as the prior model. [24] proposes an ensemble method which builds multiple weak classifiers based on spatially defined subsets of features, there are prior feature groups to deal with. Human knowledge can also be incorporated into a Bayesian network learnt from a set of observations, a recent application in image interpretation is in [38].

In general, there is a great variety of ad hoc approaches to incorporate prior knowledge into learning process. The previous paragraph roughly maps their actual range. The studies strongly attest to the positive influence of prior knowledge in the form of partial domain theory, higher-level attributes, monotonicity constraints, or structural properties [35]. Inductive logic programming [21] solves incorporation of prior knowledge into learning process in a systematic way. Statistical relational learning [10] also aims at tasks with complex relational structure, it additionally deals with uncertainty. On the other hand, generality of both the paradigms may cause intractability and integration of non-trivial prior knowledge in learning task still remains a challenging task [41].

1.3 Molecular Biology

Hereditary information encoding the development and functioning of an organism is stored in a macromolecule called *deoxyribonucleic acid* (DNA). The information is stored as a sequence of nucleotides also called *bases*, namely adenine, cytosine, guanine and thymine. The information carried by DNA is held in the sequence of distinguishable regions of DNA called *genes*. *Gene expression* (GE) is the cellular process by which information from a gene is used in the synthesis of a functional product, most often a protein. A gene is first transcribed into a *ribonucleic acid* (RNA) that serves for passing the genetic instructions from the cell nucleus to the cytoplasm where the RNA is *translated* into a *protein*. Each protein has its own unique amino acid sequence given by the nucleotide sequence of its encoding gene. A three-nucleotide combination called codon translates into an amino acid. Proteins already perform a large scale of biological functions. To sum up, the genes expressed into proteins specify the structure and function of the biological system. The above-described flow of genetic information is referred to as *the central dogma of molecular biology*.

Knowing the essential terms, the field of molecular biology can formally be defined. Molecular biology studies biological activity at the molecular level. Its main goal is to understand the aspects of DNA, RNA and protein biosynthesis including their interactions and regulation in a cellular context. It goes far beyond the simplified central dogma view. The pathway from the genotype, the inherited instructions, to the phenotype, observable characteristics such as morphology or physiological properties, is actually much more complex. Let us briefly mention the most important resources of this complexity.

Firstly, cell functioning is based on which genes get expressed under what circumstances. The expression is influenced by *transcription factors*. Transcription factor is a protein that physically *binds* to particular DNA sequences adjacent to a gene. This binding is specific but not exclusive and technically corresponds to the many-to-many relationship. One transcription factor binds to enhancers or promoters of certain *target genes* only. However, one transcription factor can contain more DNA-binding domains and thus bind to several different sequences and it does not have to aim at single gene. At the same time, multiple factors can bind to a single enhancer of promoter. The binding affinity results from the spatial structure and folding of the transcription factor and the order of nucleotides in the sequence. A transcription factor may either *promote* or *block* RNA polymerase in transcription, i.e., it may control gene expression in both directions. Transcription factors

may be activated and deactivated through their signal-sensing domain. By default they are typically inactive and cannot bind to DNA, the binding ability originates as a result of interaction with other transcription factors or external signals coming through the membrane from outside the cell. In this way, a cell responds to external stimuli and produces specific proteins under specific circumstances. While the simplified view suggests that cell can be modelled as a feedforward linear system that proceeds from DNA towards proteins and phenotype in the end, the existence of transcription factors and corresponding regulation mechanism gives rise to an extremely complex dynamic and non-linear regulatory network containing frequent feedback loops.

Secondly, there are non-protein coding DNA regions whose product is a functional RNA. They contribute to the transcriptional and translational regulation of protein-coding sequences. The recent ENCODE project suggested that over 80% of DNA in the human genome serves some purpose [30]. The term junk DNA for DNA with no known biological function tends to reduce its DNA coverage.

There are also reasons why a particular DNA sequence does not necessarily mean that a particular phenotype is produced. A single gene may code for multiple proteins. Particular exons of a gene may be included or excluded when constructing the final messenger RNA. The process is called alternative splicing, abnormal splicing variants are suspected to contribute to the development of diseases such as cancer. There are external influences, for example, changes at epigenetic level such as acetylation and methylation can cause different expression patterns and phenotypes without a change in underlying DNA sequence. In general, any environmental change can trigger a change in orchestration of the complex molecular regulatory network.

High-throughput technologies, like DNA microrarrays and RNA-Seq for transcriptome profiling that examine the expression level of mRNAs in a given cell population or ChIP-on-chip and ChIp-seq technologies used to analyze protein interactions with DNA, allow researchers to simultaneously conduct a huge number of genetic tests or measurements. However, GE data analysis, which is the main focus of this text, represents a difficult task as the data usually show an inconveniently low ratio of samples (biological situations) against variables (genes). Datasets are often noisy and they contain a great part of variables irrelevant in the context under consideration. Independent of the platform and the analysis methods used, the result of a GE experiment should be driven, annotated or at least verified against genomic prior knowledge.

There is also an issue of measurement dichotomy. When classifying

phenotypes, RNA amount is less biologically relevant than abundance of proteins and metabolites, however, it is much easier to quantify. Utilization of GE data stems from the assumption that gene expression levels correspond to protein levels. Although we know that the transcript abundance does not tell us everything, we believe it tells us a lot more than we knew before. This assumption cannot be taken for granted, agreement between mRNA levels and protein levels can be poor.

2 Set-level Microarray Classification

2.1 Micorarray Analysis and Classification

Section 1.3 briefly reviewed high-throughput technologies for parallel measuring of biological systems. Here we focus on a single technology, microarrays. We explain the principle of its operation and the main methods of the consequent data analysis and learning. Although microarrays seem to be overcome by the recent RNA-Seq method at least in several biological aspects [40], the problems of analysis and learning from the measurements remain very close for both the technologies.

Microarrays can measure the expression of thousands of genes (i.e., the amount of RNA corresponding to a given gene) under different conditions (e.g., in different tissues) in parallel. A variety of DNA microarray and DNA chip devices and systems have been developed and commercialized. DNA hybridization microarrays are generally fabricated on glass, silicon, or plastic substrates. DNA probes are selectively spotted or addressed to individual test sites, the probes can include synthetic oligonucleotides, amplicons, or larger DNA/RNA fragments attached to support material. Depending on the array format, probes can be the target DNA or RNA sequences to which other "reporter probes" would subsequently be hybridized. DNA arrays can be fabricated using physical delivery techniques such as inkjet or microjet deposition technology or in situ synthesis using a photolithographic process [14].

The output of a single microarray experiment is a colored image. The joint output of a series of n microarray experiments is a rectangular matrix $\mathbf{X} = (x_{ij})_{n \times p}$ with columns corresponding to one of p genes and rows corresponding to one of n biological samples relating to different conditions. Often, only few different conditions are taken into account (typically diseased versus normal) and multiple samples are analyzed under each of the conditions. The conditions are commonly available and taken as a categorical dependent variable. The expression x_{ij} is

most often from $\mathbb{R}$, its magnitude corresponds to the color intensity in the specific region of the colored image. Eventually, the collected GE data can be seen as the classic attribute-value data.

Probably the most frequent objective of microarray analysis is the identification of genes differentially expressed between samples obtained under different conditions. Biologically, the detected genes can help to address the origin of the phenotype under study. In the simplest setting, the detection can be carried out with the aid of common statistical tests such as t-test. The main difficulty lies in the fact that a large number of genes is tested simultaneously, i.e., in extreme multiple hypothesis testing. That is why, specialized hypothesis tests have been developed. [37] represents an example of popular dedicated technique for significance analysis of microarrays, the test has its own test statistic d whose significance is verified in terms of repeated data permutations, the false discovery rate is used instead of the family-wise error rate to control the number of false positive test results.

Microarray data can also serve for subgroup discovery (uncovers disease subtypes which allow specific future treatment), functional characterization of unknown genes (an unannotated gene can share an annotation with the better functionally explored genes having a similar expression profile), bi-clustering (groups genes showing a local expression similarity pattern, answers the problem of gene multi-functionality), and many other purposes.

Classification based on GE monitoring by DNA microarrays (often referred to as molecular classification) is a natural learning task with immediate practical uses. There have been several early success stories [1, 6, 12], followed by a large number of studies with the main goal of predicting cancer outcome (an overview is provided, e.g., in [22]).

2.2 Features For Functionally Related Genes

However, later surveys [7, 26] demonstrated serious technical flaws in a large proportion of these studies, which were published in high-impact biomedical journals, and found that most of the published results are overly optimistic. The routine application of GE classification is limited by frequent inaccuracy in the resulting classifiers and their inability to be understood by physicians. Molecular classifiers based solely on GE in most cases cannot be considered useful decision-making and decision-supporting tools.

Recent efforts in the field of molecular classification aim to employ additional information available for genes, proteins and tissues that are

being studied. They follow the major trend that is currently prevailing in the area of general GE data analysis. The analysis that was formerly aimed at identifying *individual genes* that are differentially expressed across sample classes [37] now focuses on identifying entire sets of genes with significantly differential expression [4, 16, 36]. The genes share a set of characteristics that are defined by prior biological knowledge. The *set-level techniques* applied to GE classification develop new features that correspond to gene sets that represent pathways, their sub-clusters or gene-ontology terms at various levels of generality [18, 19]. The authors of [31] propose a method that integrates a priori the knowledge of a gene network into a classification that results in classifiers with biological relevance, a good classification performance and an improved interpretability of the results. An overview of knowledge-based high-throughput data analysis can be found in [29].

In our BMC Bioinformatics paper (see Section 3 for the reference), we employed genuine curated gene sets to build robust features for subsequent classification. We proposed the whole learning workflow (see Figure 1) and suggested the optimal procedures for its crucial steps: 1) the selection of the initial pool of genuine gene sets including the comparison with their random counterparts, 2) the aggregation of member genes expression into a unique feature value, and 3) the selection of the best extracted features for further learning. The overall goal was to maximize accuracy of resulting set-level molecular classifiers as well as their biological interpretability.

We concluded that the genuine curated gene sets constitute better features for classification than the sets assembled without biological relevance. This statement is not obvious, since constructing randomized gene sets in fact corresponds to the machine learning technique of stochastic feature extraction [17] and as such may itself contribute to learning good classifiers. Nevertheless, relevant prior knowledge resting in the prior definition of biologically plausible gene sets contributes further to increasing the predictive accuracy. Smaller gene sets and sets pertaining to chemical and genetic perturbations were particularly successful.

For identifying the best gene sets for classification, we employed both dedicated gene-set ranking methods and generic feature selection methods known from machine learning such as information gain [27] and support vector machine with recursive feature extraction [13]. The first class of methods ranks the gene sets using the raw gene expression profiles while the second one deals with the aggregated feature values. The Global test [11] outperforms the gene-set methods GSEA [36] and

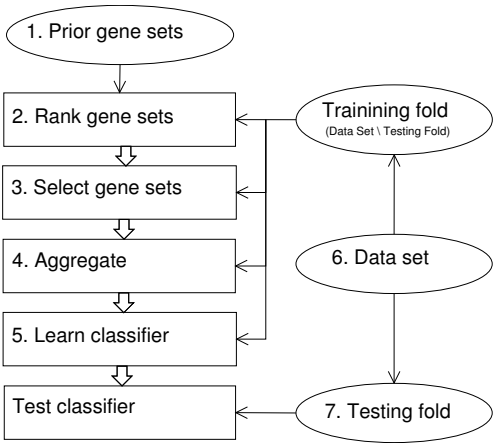


Fig. 1: The workflow of a set-level learning experiment conducted multiple times with varying alternatives in the numbered steps. For compatibility with the learned classifier, testing fold samples are also reformulated to the set level. The reformulation is done using gene sets selected in Step 3 and aggregation algorithm used in Step 4. The diagram abstracts from this operation.

SAM-GS [4] as well as two generic feature selection methods. To aggregate expressions of genes into a feature value, the singular value decomposition (SVD) method as well as the SetSig [28] technique improve on simple arithmetic averaging. These conclusions are probably the most significant for practitioners in set-level predictive modeling of gene expression as so far there has been no clear guidance to choose from the two triples of methods.

Set-level classifiers learned with 10 features constituted by the Global test slightly outperform baseline gene-level classifiers learned with all original data features although they are slightly less accurate than gene-level classifiers learned with a prior feature-selection step. In summary, set-level classifiers do not boost predictive accuracy, however, they do achieve competitive accuracy if learned with the right combination of ingredients.

Another way to introduce set-level features is studied in the paper of Krejnik and Klema (see Section 3 for the reference). Features of biological samples that originally corresponded to genes are replaced by features that correspond to the centroids of gene clusters and are then used for classifier learning. The paper focuses on functional clustering, which groups genes according to their functional similarities. The si-

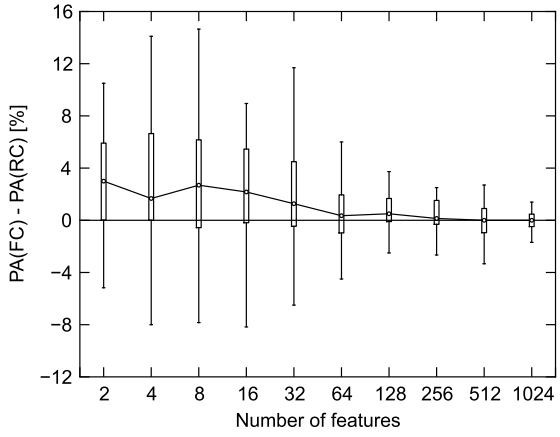
milarity measure is computed from binary vectors of the annotation terms assigned to the genes. The terms were collected from 14 annotation categories including Gene Ontology, KEGG Pathways, BioCarta Pathways and Swiss-Prot Keywords (the term can be present or absent for the given gene). The κ similarity measure adopted from [3, 20] was used.

Functional clustering was compared with random clustering without knowledge of biological relevance and gene expression clustering, which groups genes according to the similarity of their expression profiles. Using ten benchmark datasets, we demonstrated that functional clustering significantly outperforms random clustering without biological relevance. We also showed that functional clustering performs comparably to gene expression clustering, which groups genes according to the similarity of their expression profiles. The detailed results can be seen in Figure 2.

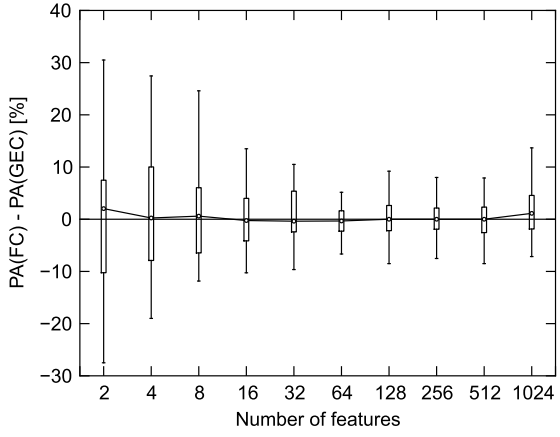
We also showed that functional clustering can provide a reasonable dimensionality reduction without sacrificing the predictive accuracy achieved with the full set of features. All the clustering approaches were also compared with the parallel approach to dimensionality reduction, feature selection. We ranked the genes by t-test, selected the most differentially expressed genes (the thresholds were gradually set to match the number of clusters) and ran the same set of classification algorithms as for clustering. It holds that functional clustering does not achieve a predictive accuracy that is comparable to that achieved by feature selection, and combining the two techniques would maximize performance.

2.3 Future Work – Further Data Integration Steps

Recent molecular research further emphasized the role of non-coding DNA, it is estimated that only about 20% of transcription across the human genome is associated with protein-coding genes. Consequently, accurate molecular models (such as disease models) need to concern more than the gene expression levels monitored by microarrays and other technologies that quantify the amount of messenger RNA. At the same time, availability of non-coding RNA data increases, new techniques such as RNA-Seq provide information on differential gene expression including differently spliced transcripts as well as non-coding RNAs. Both the above-mentioned factors ask for the concurrent analysis of the various types of transcriptional data. To exemplify, microRNAs, small non-coding RNA molecules commonly 22 nucleotides long,



(a)



(b)

Fig. 2: Box plots for the PA differences for a given number of features and pairs of feature extraction approaches: (a) FC versus RC; (b) FC versus GEC. Each box plot is computed from 50 (10 datasets $\times$ 5 classification algorithms) values for the predictive accuracy difference.

play a role in translational regulation of gene expression often resulting in gene silencing. Increasing amount of their regulatory targets can be obtained from public databases such as miRWalk [8] and TarBase [39], i.e., their interaction with genes, respectively their mRNA, is partly known. Their longer counterparts, lncRNAs, also prove to be

functional but their role is yet to be further explored.

Moreover, epigenetic data such as methylation measurements can help to explain unexpected transcriptional irregularities observed in microarrays. In order to reach deeper understanding of molecular nature of complexly orchestrated biological processes, all the available measurements and genomic knowledge need to be fused. Currently we develop algorithms allowing to immediately combine all the types of the above mentioned high-throughput data as well as the current structural genomic prior knowledge. The algorithms are principally based on prior-knowledge driven feature extraction, matrix factorization and ensemble classification.

References

[1] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, pages 54–64, 2000.

[2] M. Berens, H. Liu, L. Parsons, L. Yu, and Z. Zhao. Fostering biological relevance in feature selection for microarray data. *IEEE Intelligent Systems*, 20(6):29–32, 2005.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[4] I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, and Y. Yasui. Improving gene set analysis of microarray data by SAM-GS. *BMC bioinformatics*, 8(1):242, 2007.

[5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley, New York, 2. edition, 2001.

[6] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.

[7] A. Dupuy and R. M. Simon. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *JNCI Journal of the National Cancer Institute*, 99(2):147, 2007.

[8] H. Dweep, C. Sticht, P. Pandey, and N. Gretz. miRWalk – database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes. *Journal of Biomedical Informatics*, 44(5):839–847, 2011.

[9] L. Frey, M. E. Edgerton, D. H. Fisher, L. Tang, and Z. Chend. Using prior knowledge and rule induction methods to discover molecular markers of prognosis in lung cancer. *AMIA Annu Symp Proc.*, pages 256–260, 2005.

[10] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning.* Adaptive Computation and Machine Learning Series. MIT Press, 2007.

[11] J. J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, February 2007.

[12] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, March 2002.

[14] M. J. Heller. DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4:129–153, 2002.

[15] S. M. Hill, R. M. Neve, N. Bayani, W. Kuo, S. Ziyad, P. T. Spellman, J. W. Gray, and S. Mukherjee. Integrating biological knowledge into variable selection: an empirical bayes approach with an application in cancer biology. *BMC Bioinformatics*, 13:94, 2012.

[16] Y. Hippo, H. Taniguchi, S. Tsutsumi, N. Machida, J.M. Chong, M. Fukayama, T. Kodama, and H. Aburatani. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980, 2007.

[17] T. K. Ho. The random subspace method for constructing decision forests. *Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–44, 1997.

[18] M. Holec, F. Železný, J. Kléma, and J. Tolar. Integrating multiple-platform expression data through gene set features. In *Proceedings of the 5th International Symposium on Bioinformatics Research and Applications*, pages 5–17. Springer-Verlag Berlin, Heidelberg, 2009.

[19] M. Holec, F. Železný, J. Kléma, and J. Tolar. A comparative evaluation of gene set analysis techniques in predictive classsification of expression samples. In *International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics (BCBGC-10)*, 2010.

[20] D. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9)(R183), 2007.

[21] N. Lavrač and S. Džeroski. *Inductive Logic Programming: Tecniques and Applications*. Ellis Horwood Series in Artificial Intelligence. Ellis Horwood, 1994.

[22] J. W. Lee, J. B. Lee, M. Park, and S. H. Song. An extensive evaluation of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis*, 48:869–885, 2005.

[23] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer, 1998.

[24] M. Liu, D. Zhang, and D. Shen. Ensemble sparse classification of alzheimer's disease. *NeuroImage*, 60(2):1106–1116, 2012.

[25] S. Lomax and S. Vadera. A survey of cost-sensitive decision tree induction algorithms. *ACM Comput. Surv.*, 45(2):16:1–16:35, March 2013.

[26] S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365(9458):488–492, 2005.

[27] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[28] M. Mramor, M. Toplak, G. Leban, T. Curk, J. Demsar, and B. Zupan. On utility of gene set signatures in gene expression-based cancer class prediction. In *JMLR Workshop and Conference Proceedings Volume 8: Machine Learning in Systems Biology*, pages 55–64, 2010.

[29] M. F. Ochs. Knowledge-based data analysis comes of age. *Briefings in Bioinformatics*, 11(1):30–39, 2010.

[30] E. Pennisi. ENCODE project writes eulogy for junk DNA. *Science*, 337(6099):1159–1161, 2012.

[31] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J. P. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8(1):35+, 2007.

[32] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition, 2009.

[33] R. E. Schapire, M. Rochery, M. G. Rahim, and N. K. Gupta. Incorporating prior knowledge into boosting. In *ICML*, pages 538–545, 2002.

[34] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[35] A. P. Sinha and H. Zhao. Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decis. Support Syst.*, 46(1):287–299, December 2008.

[36] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(23):15545–1550, 2005.

[37] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, 2001.

[38] M. Velikova, P. J. F. Lucas, M. Samulski, and N. Karssemeijer. On the interplay of machine learning and background knowledge in image interpretation by bayesian networks. *Artif Intell Med*, 57(1):73–86, 2013.

[39] T. Vergoulis, I. S. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, and A. G. Hatzigeorgiou. Tarbase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Research*, 40:D222–D229, 2012.

[40] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, January 2009.

[41] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

3 Ing. Jiří Kléma, Ph.D.

Jiří Kléma received the PhD in artificial intelligence and biocybernetics from the Czech Technical University in Prague (CTU) in 2002. In 2005–2006 he carried out postdoctoral training at the University of Caen, France. Currently, he is an Assistant Professor at CTU. His main research interest is data mining and its applications in industry, medicine and bioinformatics. He focuses namely on knowledge discovery and learning in domains with complex background knowledge. He is a co-author of 12 international journal publications and 6 book chapters, a reviewer for several international journals and a member of the Presidium of The Czech Society for Cybernetics and Informatics.

Selected IF Journal Papers

- M. Krejnik, J. Klema. *Empirical Evidence of the Applicability of Functional Clustering through Gene Expression Classification.* **IEEE/ACM Trans. on Computational Biology and Bioinformatics**, 9:3, pp. 788-798, 2012.
- M. Holec, J. Klema, F. Zelezny, and J. Tolar: *Comparative Evaluation of Set-Level Techniques in Predictive Classification of Gene Expression Samples.* **BMC Bioinformatics**, 13, Suppl. 10, S15, 2012.
- E. Girmanova, I. Brabcova, J. Klema, P. Hribova, M. Wohlfartova, J. Skibova, and O. Viklicky. *Molecular Networks Involved in The Immune Control of BK Polyomavirus.* **Clinical and Developmental Immunology**, Vol. 2012, Article ID 972102, 9 p., 2012.
- H. Libalova, K. Uhlirova, J. Klema, M. Machala, R. Sram, M. Ciganek, and J. Topinka. *Global Gene Expression Changes in Human Embryonic Lung Fibroblasts Induced by Organic Extracts from Respirable Air Particles.* **Particle and Fibre Toxicology**, 9:1, 2012.
- J. Klema, L. Novakova, F. Karel, O. Stepankova, and F. Zelezny. *Sequential Data Mining: A Comparative Case Study in Development of Atherosclerosis Risk Factors.* **IEEE Trans. on Systems, Man, and Cybernetics: Part C: Applications and Reviews**, Vol. 38, no. 1, pp. 3-15, 2008.
- J. Klema, J. Kubalik, and L. Lhotska. *Optimized Model Tuning in Medical Systems.* **Computer Methods and Programs in Biomedicine**. Vol. 80, no. 3, pp. 17-28, 2005.

Web of Science Citations

- Citations: 38 (non-auto), h-index: 3.

Teaching

- Lecturing in 10 master and bachelor courses (in 2012: Advanced Methods for Knowledge Representation, Machine Learning and Data Mining, Foundations of Artificial Intelligence, Bioinformatics),
- 20 diploma and bachelor thesis supervised (4 of them awarded), introduction of 6 new courses, a member of The Open Informatics Board,
- a supervisor of 1 PhD graduate.