Proceedings of the Third International Workshop on Machine Learning in Systems Biology: Revised Selected Papers

Sašo Džeroski   Saso.Dzeroski@ijs.si
Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia

Pierre Geurts   p.geurts@ulg.ac.be
Department of EE and CS & GIGA-R, University of Liège, Belgium

Juho Rousu   juho.rousu@cs.helsinki.fi
Department of Computer Science, University of Helsinki, Finland

Editor: Neil Lawrence

Molecular biology and all the biomedical sciences are undergoing a true revolution as a result of the emergence and growing impact of a series of new disciplines and tools sharing the '-omics' suffix in their name. These include in particular genomics, transcriptomics, proteomics and metabolomics, devoted respectively to the examination of the entire systems of genes, transcripts, proteins and metabolites present in a given cell or tissue type. The availability of these new, highly effective tools for biological exploration is dramatically changing the way one performs research in at least two respects. First, the amount of available experimental data is not a limiting factor any more; on the contrary, there is a plethora of it. Given the research question, the challenge has shifted towards identifying the relevant pieces of information and making sense out of it (a 'data mining' issue). Second, rather than focus on components in isolation, we can now try to understand how biological systems behave as a result of the integration and interaction between the individual components that one can now monitor simultaneously, so called 'systems biology'.

Machine learning naturally appears as one of the main drivers of progress in this context, where most of the targets of interest deal with complex structured objects: sequences, 2D and 3D structures or interaction networks. At the same time bioinformatics and systems biology have already induced significant new developments of general interest in machine learning, for example in the context of learning with structured data, graph inference, semi- supervised learning, system identification, and novel combinations of optimization and learning algorithms.

This volume contains extended versions of selected papers presented at the Third International Workshop on Machine Learning in Systems Biology (MLSB-2009), held in Ljubljana, Slovenia from September 5 to 6, 2009. The aim of the workshop was to contribute to the cross-fertilization between the research in machine learning methods and their applications to systems biology (i.e., complex biological and medical questions) by bringing together method developers and experimentalists.

The technical program of the workshop consisted of 6 invited lectures, 12 oral presentations and 22 poster presentations. Invited lectures were given by Diego di Bernardo, Roman

Jerala, Nick Juty, Yannis Kalaidzidis, Ross D. King, and William Stafford Noble. The workshop attracted around 60 scholars from around the world. The workshop was organized as a core event of the PASCAL2 Network of Excellence, under the IST programme of European Union, and also sponsored by the Slovenian Research Agency, Jožef Stefan Institute, and University of Helsinki. All the lectures at the workshop were recorded and are available for viewing via the `videolectures.net` portal. More information on the workshop can be found at `mlsb09.ijs.si`.

The selected papers cover a broad spectrum of topics under the theme of the workshop.

Function prediction is the topic of three papers. Cesa-Bianchi and Valentini study cost-sensitive hierarchical classification method for gene function prediction, and are able to show significant benefits over "flat" classification approaches. Re and Valentini show that simple ensemble methods reach predictive performance on par with state-of-the art methods. Rahmani, Blockeel, and Bender focus on function prediction in the protein-protein interaction network context.

Biological networks are another frequent topic in the papers. Saeys, Van Landgehem, and Van De Peer consider the use of machine learning and text mining for the reconstruction of heterogeneous interaction networks. Isik, Atalay and Cetin-Atalay present an evaluation method for signaling cascades using microarray and ChIP-seq data. Umek and co-workers present a method combining unsupervised and supervised machine learning for discovering relationships between chemical structure and biochemical pathways from growth-fitness characterization data.

Biomarker discovery is discussed in two papers. Mramor et al. present a machine learning approach that classifies cancer types by using so-called gene set signatures. Slavkov, Ženko and Džeroski study methods for evaluating and aggregating feature rankings obtained by different algorithms or from different datasets.

Two papers present advances in methods for evaluating machine learning performance. Airola and co-workers point out problems in the use of the area under ROC curve as a performance measure in cross-validation schemes and argue for the leave-pair-out cross-validation over other commonly used cross-validation schemes. Nadeem, Zucker, and Hanzar propose Accuracy-Rejection curves for classifier comparison in settings where the classifier has the reject-option, i.e. the option to refrain from predicting a label.

Finally, Hullám and co-workers put forward Bayesian approaches for feature selection in Genetic Association Studies.

The editors of these proceedings of MLSB-09, containing revised selected papers presented at the workshop, would like to thank all authors of this volume, as well as all MLSB-2009 contributors, the scientific program committee, the local organizers and the sponsors for making the workshop the success it was.

The next workshop on the topic, the Fourth International Workshop on Machine Learning in Systems Biology, MLSB-2010, will be held October 15 to 16 in Edinburgh, Scotland (`mlsb10.ijs.si`).

Ljubljana, Liège, Helsinki, February 2010,

Sašo Džeroski, Pierre Geurts, Juho Rousu
MLSB-2009 Program Chairs