StReBio'09: Statistical Relational Learning and Mining in Bioinformatics

Jan Ramon
Dept. of Computer Science
K.U.Leuven
Celestijnenlaan 200A
3001 Leuven, Belgium
jan.ramon@cs.kuleuven.be

Fabrizio Costa
Dept. of Computer Science
K.U.Leuven
Celestijnenlaan 200A
3001 Leuven, Belgium
fabrizio.costa@cs.kuleuven.be

Christophe Costa Florencio
Dept. of Computer Science
K.U.Leuven
Celestijnenlaan 200A
3001 Leuven, Belgium
chris@cs.kuleuven.be

ABSTRACT

Bioinformatics is an application domain where information is naturally represented in terms of relations between heterogenous objects. Modern experimentation and data acquisition techniques allow the study of complex interactions in biological systems. This raises interesting challenges for machine learning and data mining researchers, as the amount of data is huge, some information can not be observed, and measurements may be noisy.

This report presents a review on the ACM SIGKDD 2009 Workshop on Statistical Relational Learning and Mining in Bioinformatics (StReBio'09) which was held in Paris on June 28th, 2009. The aim of this workshop was to provide a forum to share challenges, results and ideas at the frontier between the field of statistical relational learning and the field of bioinformatics.

1. INTRODUCTION

Bioinformatics is an application domain where information is naturally represented in terms of relations between heterogenous objects such as DNA, RNA, proteins, peptides, chemicals, genes and organisms. Modern experimentation and data acquisition techniques allow the study of complex interactions in biological systems. This raises interesting challenges for machine learning and data mining researchers. First, the amount of data is huge and the structure of it is complex. At the same time, there is often a large fraction of missing data (as experiments are expensive, and not everything can be measured), while the available data is noisy due to a wide range of different effects that can interfere with the experiments.

This report presents a review on the Second Workshop on Statistical Relational Learning and Mining in Bioinformatics (StReBio'09) which was held with ACM-SIGKDD 2009 in Paris, France on June 28th, 2009. The first StReBio workshop was held at ECML/PKDD in Antwerpen, Belgium in September 2008. The aim of this workshop was to bring together researchers at the frontier between the field of statistical relational learning [4; 3] and the field of bioinformatics [1; 5]. To stimulate interaction between the different fields, next to the regular contributions we invited problem statements and invited a speaker from the field of statistical genetics, David Balding, who attracted quite some attention.

We will first present a review of the contributions to the workshop. After that, we will conclude with some discussion and perspectives.

2. CONTRIBUTIONS

The invited speaker of this workshop was David Balding, professor of statistical genetics [2] at Imperial College, London, UK. He presented joined work with Clive Hoggart, John Whittaker qnd Maria De Iorio on 'Simultaneous Analysis of all SNPs in Genome-wide and Resequencing Association Studies'. Testing one marker at a time does not fully realise the potential of genome-wide association studies to identify multiple causal variants, which is a plausible scenario for many complex diseases. Therefore, more complex models are needed taking many SNPs into account simultaneously. In this research the authors employ a recent search method to fit a model where every SNP can be considered for additive, dominant and recessive contributions to disease risk.

The paper by Lodhi, Muggleton and Sternberg "Multi-Class Protein Fold Recognition using Large Margin Logic based Divide and Conquer Learning" deals with the task of protein fold multiclassification. The multiclassification is transformed into a recursive binary classification problem using dynamic decision lists. The authors propose to use an ILP technique to extract features which are then ranked and selected using an information theoretic measure. Support Vector Machines based on Radial Basis Function kernels are then used as model for the predictive task.

The paper by Arevalillo and Navarro "Using Random Forests to uncover bivariate interactions in high dimensional small data sets" tackles the task of detecting pairwise feature interactions when data live in high dimensional spaces but are available in small quantities. The authors propose to use the out-of-bag error of an ensemble of decision trees trained in a supervised fashion on a binary classification problem to assess pairwise feature interactions. In order to reduce the computational burden they consider sets of variables and choose the set with the smallest generalization error. That set is then believed to contain the interacting features.

The paper by Ke?elj, Liu, Zeh, Blouin and Whidden "Finding Optimal Parameters for Edit-Distance Based Sequence Classification is NP-Hard" deals with the complexity analyses of finding optimal parameters for edit-distance measures.

Optimality is defined in respect of the accuracy achievable by a 1-nearest-neighbour classifier in the induced metric space. The authors motivate the relevance of the study with a bioinformatic application for the assessment of similarity between strings representing sentences in natural text that talk about interactions of proteins.

The paper by Hämalainen "Lift-based search for significant dependencies in dense data sets" tackles the task of detecting dependencies between binary attributes that can be expressed as association rules. The author proposes to use the notion of "lift" to build an anti-monotonic goodness measure called "potentially significant", which is then used to prune the search space for interesting association rules. The main concern of the author is to find a small set of less "redundant" rules (redundancy is defined once again using the lift notion). The experimental verification is done over a set of 5 biological/medical datasets and results are compared against the Apriori approach.

The paper by Cheng, Lu and Li "Identification of structurally important amino acids in proteins by graph-theoretic measures" deals with the task of identifying single residues in a protein that are responsible for the ovearll stability of the protein. The authors propose to use the graph theoretical notion of minimum vertex cover to rank single residues. As the proposed measure is NP-complete, the authors suggest a greedy approximation which boils down to iteratively remove vertices according to their degree. The authors find that this measure has a good linear correlation with the stability of the residue as numerically computed via software simulation. The rank is shown to be a better indicator as compared to betweenness and spectral methods, although from the experimental section is not clear if the comparison happens on an equal basis.

The paper 'Comparing Graph-based Representations of Protein for Mining Purposes' by Saidi, Maddouri and Nguifo deals with the preprocessing of protein data so that a representation (often graph-based) suitable for mining is obtained. They criticize existing methods from the perspective of a graph-theoretic analysis of important properties of proteins, and propose and evaluate a way to enhance existing methods.

The paper by Obradovic, Midic and Dunker, 'Protein Sequence Alignment and Intrinsic Disorder: A Substitution Matrix for an Extended Alphabet' proposes a way to extend substitution matrices, commonly used in protein alignment algorithms, to deal with intrinsically disordered proteins. This is achieved by extending the alphabet and using an iterative algorithm to estimate the associated matrix. This method performs well on realistic datasets.

Next to these regular contribution, the workshop featured also a problem statement. In 'Can we improve on the identification of Transcription Factor Binding Sites using string kernels?', Hugh Shanahan presents the problem of finding small sequences (TFBS's) with very specific properties in large non-coding regions of DNA. Relational information (i.e., across individuals/species) about the position and nature of TFBS's is becoming available, but mining this requires the deployment of new methods. Finding non-local correlations between TFBS's is another challenge, and string kernels seem a good candidate for dealing with this aspect. An SVM with mismatch string kernel is reported to outperform existing methods.

3. CONCLUSIONS

In this paper we presented a brief review of the StReBio'09 workshop. The workshop brought together researchers from the field of data mining and bioinformatics to exchange challenges, ideas and findings.

We hope this workshop contributed to the communication between experts and a better understanding of the numerous aspects of the biological problems and the computational and statistical methods. It is clear that the interaction between the involved fields will remain essential for progress in the future.

More information about the StReBio'09 workshop can be found on its website `http://www.cs.kuleuven.be/ dtai//events/St`

A special issue on the topic of the workshop is being prepared to publish the key results of the workshops in a more archival way.

4. ACKNOWLEDGEMENTS

This workshop is partially supported by the Fund for Scientific Research of Flanders. Jan Ramon is a post-doctoral fellow of the K.U.Leuven.

5. ADDITIONAL AUTHORS

Additional authors: Joost Kok (Leiden Institute of Advanced Computer Science, P.O. Box 9512, 2300 RA Leiden, The Netherland, `joost@liacs.nl`)

6. REFERENCES

[1] P. Baldi and S. Brunak. *Bioinformatics: the machine learning approach.* MIT Press, Cambridge, MA, USA, 2001.

[2] D. J. Balding, M. Bishop, and C. Cannings. *Handbook of Statistical Genetics.* Wiley, 2007.

[3] L. De Raedt, P. Frasconi, K. Kersting, and S. Muggleton. *Probabilistic Inductive Logic Programming.* Springer, 2008.

[4] L. Getoor and B. Taskar. *Introduction to statistical relational learning.* MIT press, 2007.

[5] M. S. Waterman. *Introduction to Computational Biology.* Chapman & Hall, 1995.