Data Analysis in the Life Sciences — Sparking Ideas —

Michael R. Berthold

ALTANA-Chair for Bioinformatics and Information Mining,
Dept. of Computer and Information Science, Konstanz University, Germany
Michael.Berthold@uni-konstanz.de

Data from various areas of Life Sciences have increasingly caught the attention of data mining and machine learning researchers. Not only is the amount of data available mind-boggling but the diverse and heterogenous nature of the information is far beyond any other data analysis problem so far. In sharp contrast to classical data analysis scenarios, the life science area poses challenges of a rather different nature for mainly two reasons. Firstly, the available data stems from heterogenous information sources of varying degrees of reliability and quality and is, without the interactive, constant interpretation of a domain expert, not useful. Furthermore, predictive models are of only marginal interest to those users – instead they hope for new insights into a complex, biological system that is only partially represented within that data anyway. In this scenario, the data serves mainly to create new insights and generate new ideas that can be tested. Secondly, the notion of feature space and the accompanying measures of similarity cannot be taken for granted. Similarity measures become context dependent and it is often the case that within one analysis task several different ways of describing the objects of interest or measuring similarity between them matter.

Some more recently published work in the data analysis area has started to address some of these issues. For example, data analysis in parallel universes [1], that is, the detection of patterns of interest in various different descriptor spaces at the same time, and mining of frequent, discriminative fragments in large, molecular data bases [2]. In both cases, sheer numerical performance is not the focus; it is rather the discovery of interpretable pieces of evidence that lights up new ideas in the users mind. Future work in data analysis in the life sciences needs to keep this in mind: the goal is to trigger new ideas and stimulate interesting associations.

References

1. Berthold, M.R., Wiswedel, B., Patterson, D.E.: Interactive exploration of fuzzy clusters using neighborgrams. Fuzzy Sets and Systems 149 (2005) 21-37
2. Hofer, H., Borgelt, C., Berthold, M.R.: Large scale mining of molecular fragments with wildcards. Intelligent Data Analysis 8 (2004) 376-385

J. Gama et al. (Eds.): ECML 2005, LNAI 3720, p. 1, 2005.