Knowledge discovery in biological and chemical domains

Stephen Muggleton

Department of Computer Science, University of York, Heslington, York, YO1 5DD, United Kingdom.

1 Extended abstract

The pharmaceutical industry is increasingly overwhelmed by large-volume-data. This is generated both internally as a side-effect of screening tests and combinatorial chemistry, as well as externally from sources such as the human genome project. The industry is predominantly knowledge-driven. For instance, knowledge is required within computational chemistry for pharmacophore identification, as well as for determining biological function using sequence analysis.

From a computer science point of view, the knowledge requirements within the industry give higher emphasis to "knowing that" (declarative or descriptive knowledge) rather than "knowing how" (procedural or prescriptive knowledge). Mathematical logic has always been the preferred representation for declarative knowledge and thus knowledge discovery techniques are required which generate logical formulae from data. Inductive Logic Programming (ILP) [6, 1] provides such an approach.

This talk will review the results of the last few years' academic pilot studies involving the application of ILP to the prediction of protein secondary structure [5, 8, 9], mutagenicity [4, 7], structure activity [3], pharmacophore discovery [2] and protein fold analysis [10]. While predictive accuracy is the central performance measure of data analytical techniques which generate procedural knowledge (neural nets, decision trees, etc.), the performance of an ILP system is determined both by accuracy and degree of stereo-chemical insight provided. ILP hypotheses can be easily stated in English and exemplified diagrammatically. This allows cross-checking with the relevant biological and chemical literature. Most importantly it allows for expert involvement in human background knowledge refinement and for final dissemination of discoveries to the wider scientific community. In several of the comparative trials presented ILP systems provided significant chemical and biological insights where other data analysis techniques did not.

In his statement of the importance of this line of research to the Royal Society [8] Sternberg emphasised the aspect of joint human-computer collaboration in scientific discoveries. Science is an activity of human societies. It is our belief that computer-based scientific discovery must support strong integration into existing the social environment of human scientific communities. The discovered knowledge must add to and build on existing science. The author believes that

the ability to incorporate background knowledge and re-use learned knowledge together with the comprehensibility of the hypotheses, have marked out ILP as a particularly effective approach for scientific knowledge discovery.

Acknowledgements

This work was supported partly by the Esprit Long Term Research Action ILP II (project 20237), EPSRC grant GR/K57985 on Experiments with Distribution-based Machine Learning and an EPSRC Advanced Research Fellowship held by the author. We would also like to thank both Pfizer UK and Smith-Kline Beecham for their generous support of some of this work.

References

1. I. Bratko and S. Muggleton. Applications of inductive logic programming. *Communications of the ACM*, 38(11):65–70, 1995.
2. P. Finn, S. Muggleton, D. Page, and A. Srinivasan. Pharmacophore discovery using the inductive logic programming system Progol. *Machine Learning*, 30:241–271, 1998.
3. R. King, S. Muggleton, R. Lewis, and M. Sternberg. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proceedings of the National Academy of Sciences*, 89(23):11322–11326, 1992.
4. R. King, S. Muggleton, A. Srinivasan, and M. Sternberg. Structure-activity relationships derived by machine learning: the use of atoms and their bond connectives to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences*, 93:438–442, 1996.
5. S. Muggleton, R. King, and M. Sternberg. Protein secondary structure prediction using logic-based machine learning. *Protein Engineering*, 5(7):647–657, 1992.
6. S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19,20:629–679, 1994.
7. A. Srinivasan, S. Muggleton, R. King, and M. Sternberg. Theories for mutagenicity: a study of first-order and feature based induction. *Artificial Intelligence*, 85(1,2):277–299, 1996.
8. M. Sternberg, R. King, R. Lewis, and S. Muggleton. Application of machine learning to structural molecular biology. *Philosophical Transactions of the Royal Society B*, 344:365–371, 1994.
9. M. Sternberg, R. Lewis, R. King, and S. Muggleton. Modelling the structure and function of enzymes by machine learning. *Proceedings of the Royal Society of Chemistry: Faraday Discussions*, 93:269–280, 1992.
10. M. Turcotte, S.H. Muggleton, and M.J.E. Sternberg. Protein fold recognition. In C.D. Page, editor, *Proc. of the 8th International Workshop on Inductive Logic Programming (ILP-98)*, LNAI 1446, pages 53–64, Berlin, 1998. Springer-Verlag.