Institution: Imperial College London

Unit of Assessment: Computing

Title of case study: Case Study 1: Machine Learning for Agrisciences (Syngenta)

1. Summary of the impact (indicative maximum 100 words)

The world is facing challenges in feeding its growing population. Climate change and increasing urbanisation have led to estimates of a 50% increase required in food production by 2050 according to report of the FAO on food security in 2012. Agriscience research has been developing high yielding crop varieties, which in turn, requires integration of incomplete complex data sets. Research in machine learning and predictive modelling at Imperial has addressed these challenges by filling the gaps in the descriptions of biological networks. This is having a significant economic impact in agriscience areas such as tomato ripening (market value > $100m), herbicide toxicity (market value ca. $20bn) and environmental modelling of herbicide-based crop management (> $100m).

2. Underpinning research (indicative maximum 500 words)

Within this case study predictive modelling involves the use of a form of Machine Learning known as Inductive Logic Programming (ILP) [2, 4, i, ii, iii]. Professor Stephen Muggleton is the founder of ILP, the only subarea of Machine Learning which supports the use of explicitly encoded prior knowledge within a first-order logic representation. Within ILP, the key components for a learning problem consist of examples (E), background knowledge (B) and hypotheses (H) in the form of logic programs. Hypotheses are formulated based on a search through a refinement graph under the constraint that B,H |= E, i.e. the examples are explained by the hypotheses conjoined with the background knowledge. Additionally, hypotheses are selected to maximise the posterior probability of the observations. In biological applications, such as those studied in the Syngenta University Innovation Centre (UIC – http://www3.imperial.ac.uk/syngenta-uic), the examples E represent observations of up/down regulation of both cellular metabolites and expressed genes, or up/down regulation of species count. Background knowledge takes the form of encoded biological networks or known trophic relationships and hypotheses add suggest new arcs and labels in the networks.

The primary ILP system used in this case study is CProgol5 [2], which supports the integration of abductive and inductive inference within an inverse entailment framework. The novelty of this approach with respect to learning in scientific domains, such as the modelling of gene product flux within cells, was its ability to revise multiple theoretical predicate definitions on the basis of examples given for a single observational predicate. This form of integration was first demonstrated to be effective for Systems Biology applications in the Robot Scientist project [3] where observations were related to metabolic concentrations at the cell boundary, while hypotheses were described in terms of enzyme modulation of biochemical reactions. Subsequently, the approach was extended to the context of toxicological modelling in which examples were associated with metabolic concentration variations and hypotheses were related to reactions in the cell that were inhibited by the toxin. Moreover, variants of this approach which support the learning of Stochastic Logic Programs [4] have proved critical within the work at the UIC on Ecological modelling. In this case probability estimates associated with individual hypotheses are based on frequency of occurrence of abductively derived hypotheses within samples of consistent hypotheses.

The large hypothesis spaces in Systems Biology applications at the UIC motivated the development of the TopLog sytem [5, iv] which uses a Top theory as a form of explicit declarative bias to significantly increase the efficiency of the search. This approach has been further refined within the MC-TopLog system [6] in which a multi-clause search was shown complete for integrating abduction and induction. While the single clause hypotheses derived by CProgol5 and TopLog are effective for suggesting individual cause effect pairs, the multi-clause solutions of MC-TopLog have allowed hypotheses involving sub-networks of multiple inter-related causes and

effects. In scientific terms this characterises the shift within Systems Biology from Reductionist to Systems-level hypotheses. Evidence of the significance of our research for Syngenta has been their commitment to sponsor Professor Muggleton's Royal Academy of Engineering Chair [v] for five years from 2013.

3. References to the research (indicative maximum of six references)

Publications that directly describe the underpinning research

* References that best indicate quality of underpinning research.

[1] S.H. Muggleton et al. ILP turns 20: biography and future challenges. Machine Learning, 86:3-23, 2012. http://dx.doi.org/10.1007/s10994-011-5259-2

* [2] S.H. Muggleton and C.H. Bryant. Theory completion using inverse entailment. In Proc. of the 10th International Workshop on Inductive Logic Programming (ILP-00), pp. 130-146, 2000. http://dx.doi.org/10.1007/3-540-44960-4_8

*[3] R.D. King, K.E. Whelan, F.M. Jones, P.K.G. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell, and S.G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature, 427:247-252, 2004. http://dx.doi.org/10.1038/nature02236

[4] S.H. Muggleton. Learning structure and parameters of stochastic logic programs. In Proceedings of the 12th International Conference on Inductive Logic Programming, pp. 198-206. 2002. http://dx.doi.org/10.1007/3-540-36468-4_13

*[5] S.H. Muggleton, J. Santos, and A. Tamaddoni-Nezhad. TopLog: ILP using a logic program declarative bias. In Proceedings of the International Conference on Logic Programming 2008, pp. 687-692, 2008. http://dx.doi.org/10.1007/978-3-540-89982-2_58

Grants that directly funded the underpinning research

[i] Closed Loop Machine Learning. EPSRC GR/M56067/02, S. Muggleton (PI), £52,213, September 2001 – March 2002.

[ii] Metalog - Integrated Machine Learning of Metabolic Networks Applied to Predictive Toxicology. DTI Beacon project, S. Muggleton (PI), £1.2M, October 2002 – March 2006.

[iii] Application of Probabilistic Inductive Logic Programming II, EU FP6-508861, S. Muggleton (PI), £301K. January 2004 – December 2006.

[iv] Four year PhD programme in Bioinformatics at Imperial College London, Wellcome Trust No. 069962, S (£1.2M), S. Muggleton (CI), October 2005 – September 2012.

[v] Automated Microfluidic Experimentation using Probabilistic Inductive Logic Programming. RAEng Research Chair for S. Muggleton, RAE/10143/55, £608K, January 2007 – December 2011.

4. Details of the impact (indicative maximum 750 words)

Syngenta is the world's third largest company specialising in seeds and pesticides. The company is driven by its biotechnology and genomic research. The challenges facing the world's food supply are also key issues facing agriscience companies such as Syngenta. A key issue is that the science is complex and the data are incomplete. In order to solve these challenges, following a world-wide search among leading Universities, Syngenta identified the modelling work in Professor Muggleton's group at Imperial as being the best match for their Systems Biology modelling requirements. In 2008 Syngenta founded a Systems Biology Innovation Centre at Imperial under Prof Muggleton's supervision to enable collaboration within a $3m programme using Inductive Logic Programming techniques in an industrial context [A].

Since the start of the collaboration in 2008, ILP techniques developed in Prof Muggleton's group have had profound impact on Syngenta's business in areas such as improving consumer traits in tomatoes; environmental benefits in better prediction of herbicide impact, and reducing risk by prediction of unwanted toxicological effects. In particular, Muggleton's ILP approach represents background knowledge, allowing description of biochemical networks. This has been a key advantage with respect to other Systems Biology modelling approaches. Since ILP is a Machine

Learning technique, this allows the network description to be automatically revised on the basis of high-throughput biological data. ILP has the advantage of suggesting readily comprehensible hypotheses. Biologists can then examine the hypotheses using their existing knowledge. Those plausible hypotheses that are impossible to disprove can be considered for further experimental validation, while a biologically non-meaningful hypothesis may indicate that insufficient background knowledge has been provided.

The specific commercial impact is detailed below.

Tomato ripening (2008-2013)[A]:

In the context of tomato-ripening, the tools developed by Professor Muggleton have highlighted new pathways that have been integrated into tomato-breeding programs. The outputs of the tomato-ripening project have changed Syngenta's research direction by focusing on manipulating pathways using tools developed at the UIC rather than individual proteins. ILP is used for analysis of gene expression and metabolite changes across fruit development to identify new genetic targets that play a role in controlling the ripening process. This allows Syngenta to focus on these genetic control points in breeding new tomato varieties, thus producing the most favourable combination of fruit quality characters in the ripe fruit. A Senior Syngenta Fellow confirms that this research "has informed a breeding program where we are using this information to breed new varieties using marker assisted technologies"[A]. The cost of breeding a new tomato variety is $5m [B]. Syngenta assess that the market size of improved flavour earlier in the ripening process is >$100m; Syngenta is the world's largest supplier of tomato seed and it is estimated that Syngenta can capture between $40 -50m of this market.

This is a high impact problem for Syngenta, and the UIC has delivered novel solutions addressing these challenges.

Herbicide toxicity (2008-2013)[A]:

When new agrochemicals fail due to adverse toxicology, it is often because of liver cancer. The research has been used to identify molecular markers that are indicative of liver cancer. These markers have been used to develop a simple biochemical screen to reduce the chances of failure at the late stage. In addition to saving costs, this reduces reliance on whole animal mammalian tests, which has a positive ethical outcome. These toxicology experiments are expensive (multi-million $), and use whole animal in-vivo tests prescribed by regulators. When compounds fail at this stage, Syngenta has had to spend around $100-250m. The Imperial research has enabled successful identification of new pathways containing molecular markers that have been linked to liver cancer. Early removal of potentially carcinogenic herbicides highly impacts Syngenta's reputation and benefits society. In addition, key Syngenta employees have been trained in the use of computational and visualisation tools developed in this collaboration. These tools have been employed in Syngenta's research into product safety, changing their working practices. "The downstream value of this will be very large indeed regarding reduction in our toxicological failure rate once the pathways have been fully investigated" quote from [A].

Environmental modelling of herbicide-based crop management (2009-2013)[A]:

Within Ecology the network of predatory inter-relationships between species is referred to as a Food Web – what eats what. Food Webs are key to understanding the effects of agriculture on the environment. However, detailed Food Webs are rare since each link requires intensive field studies. The application of the underpinning ILP research in environmental modelling has led to production of a machine-suggested food web involving 45 species, generated automatically from a large-scale field dataset provided by Syngenta, as detailed in [C]. The proposed web correlates strongly with links suggested within the literature. The industrial significance of this is that food-webs are indicators of biodiversity [C]. The automatic generation of food-webs will enable Syngenta to perform experiments at a landscape scale that have been previously impossible to do. Evaluating environmental impact of a change in farming practice, such as a new herbicide, at a species level requires sampling across farms over a few years. Prediction using machine learning helps provide guidance to farmers on the impact of farming practices. The economic and environmental impact can be quantified. Syngenta wishes to ensure their chemistry has minimal impact on the environment. Syngenta estimates that use of machine learning based Food Web

generation will facilitate opening markets worth >$100m per year for a new herbicide [B].

Equinox Pharma Ltd

The same underpinning ILP technology is being used within the Imperial College spinout Equinox Pharma, founded by Prof Muggleton and two other Imperial College professors in 2008. The company applies Support Vector ILP to initial molecule candidate discovery within both drug discovery and herbicide discovery problems. Initial contracts obtained by the Equinox Pharma have been from Japanese pharmaceutical company Astellas and Syngenta. Due to the commercially sensitive information we cannot go into details but in both cases the machine learning led to promising candidates that were taken forward for internal validation within the companies. [E]

5. Sources to corroborate the impact (indicative maximum of 10 references.)

[A] Senior Syngenta Fellow, Global Head of Biochemistry, Syngenta to corroroate impact on tomatoe ripening

[B] S.H. Muggleton et al. Variation of background knowledge in an industrial application of ILP. Proc. 20th Int. Conf. on Inductive Logic Programming, pp. 158-170, 2011. http://dx.doi.org/10.1007/978-3-642-21295-6_19

[C] D.A. Bohen, G. Caron-Lormier, S.H. Muggleton, A. Raybould, and A. Tamaddoni-Nezhad. Automated discovery of food webs from ecological data using logic-based machine learning. PLoS ONE, 2011. http://dx.doi.org/10.1371/journal.pone.0029028

[D] D. Lin et.al. Does multi-clause learning help in real-world applications? Proc. 21st Int. Conf. on Inductive Logic Programming, pp 221-237, 2012. http://dx.doi.org/10.1007/978-3-642-31951-8_21

[E] http://www.equinoxpharma.com/index.html . Archived on 22/1-/2013 https://www.imperial.ac.uk/ref/webarchive/lyf