Machine Learning a Probabilistic Network of Ecological Interactions

Alireza Tamaddoni-Nezhad[1], David Bohan[2],
Alan Raybould[3], and Stephen Muggleton[1]

[1] Department of Computing, Imperial College London, London, SW7 2AZ, UK
Email: {a.tamaddoni-nezhad,s.muggleton}@imperial.ac.uk
[2] Rothamsted Research, West Common, Harpenden, Herts, Al5 2JQ, UK
Email: david.bohan@rothamsted.ac.uk
[3] Syngenta Ltd, Bracknell, Berkshire, RG42 6EY, UK
Email: alan.raybould@syngenta.com

Abstract. In this paper we demonstrate that Abductive ILP can generate plausible and testable food webs from ecological data. In this approach, unlike previous applications, the abductive predicate 'eats' is entirely undefined before the start of the learning. We also explore a new approach for estimating probabilities for hypothetical 'eats' facts based on their frequency of occurrence when randomly sampling the hypothesis space. The results of cross-validation tests suggest that the trophic networks with probabilities have higher predictive accuracies compared to the networks without probabilities. The proposed trophic networks have been examined by domain experts and comparison with the literature shows that many of the links are corroborated by the literature. In particular, links ascribed with high probability are shown to correspond well with those having multiple references in the literature. In some cases novel high probability links are suggested, which could be tested.

1 Introduction

Machine Learning has the potential to address many challenging problems in ecological sciences [4]. Discovery of trophic links (food chains) which describe the flow of energy/biomass between species is one of these problems. Networks of trophic links (food webs) are important for explaining ecosystem structure and dynamics [2]. However, relatively few ecosystems have been studied through detailed food webs because finding out the predation relationships between the many hundreds of species in an ecosystem is difficult and expensive. Hence, any technique which can automate the discovery of trophic links from ecological data is highly desirable. Similar problems of network construction have been tackled in other complex systems, such as metabolic networks (e.g. [8]). In this paper we demonstrate that Abductive ILP can generate plausible and testable food webs from ecological data. In this approach the abductive predicate 'eats' is entirely undefined before the start of the learning process. This contrasts with previous applications of Abductive ILP where partial, non-empty, definitions exist and the gaps are filled by abduced hypotheses. In this paper we also explore a new approach for estimating probabilities for hypothetical 'eats' facts based on their frequency of occurrence when random permutations of the training data (and hence different seeds for defining the hypothesis space) are considered. We empirically evaluate the hypothetical trophic networks using leave-one-out cross-validation tests on the observable data. The results of cross-validation tests for the networks with and without probabilities are presented. The proposed trophic networks have been also examined by domain experts and the results of comparison with the literature are presented.

2 Ecological data

The data set was sampled from 257 fields across the UK in the Farm Scale Evaluations (FSE) of GM, herbicide tolerant (GMHT) crops. This national-scale experiment evaluated the change in weed plants and invertebrates between the current, conventional herbicide management of spring-sown Maize, Beet and Oilseed Rape and winter-sown Oilseed Rape, and the herbicide management of GMHT varieties of the same crops using a split-field design. We use data from the Vortis suction sampling protocol for epigeal invertebrates [6, 1] to calculate a treatment effect ratio. The counts from each conventional and GMHT half-field pair were converted to multiplicative treatment ratio, R, and as in [6, 1] treatment ratio values of $R < 0.67$ and $R > 1.5$ were regarded as important changes in count with direction of down (decreased) and up (increased), respectively. This information on up and down abundances is regarded as our primary observational data for the learning.

3 Machine learning of trophic links using Abductive ILP

The main role of abductive reasoning in machine learning and its use in the development of scientific theories [5] is to provide hypothetical explanations of the empirical observations. Then based on these explanations we try to inject back into the current scientific theory, new information that helps complete the theory. This process of generating abductive explanations and then updating in some way the theory with them can be repeated several times when new observational data is made available. In many implementation of abductive reasoning, such as that of Progol 5 [7], which is used in this paper, the approach taken is to choose an explanation that best generalises under some form of inductive reasoning (e.g. simplest explanation approximated by compressibility). We refer to this approach as Abductive ILP (A/ILP). We believe that ecological data in this study fulfil the conditions for the use of A/ILP: firstly, the given background knowledge is incomplete; and secondly, the problem requires learning in the circumstance in which the hypothesis language is disjoint from the observation language. In our problem, the set of observable data can be represented by predicate $abundance(X, S, up)$ (or $abundance(X, S, down)$) expressing the fact that the abundance of X at site S is *up* (or *down*). This information is compiled from FSE data as described in Section 2. The knowledge gap that we initially aim to fill is a predation relationship between species. Thus, we declare abducible predicate $eats(X, Y)$ capturing the hypothesis that species X eats species Y. In order to use abduction, we also need to provide the rules which describe the observable predicate in terms of the abducible predicate. An example of such a rule is shown below.

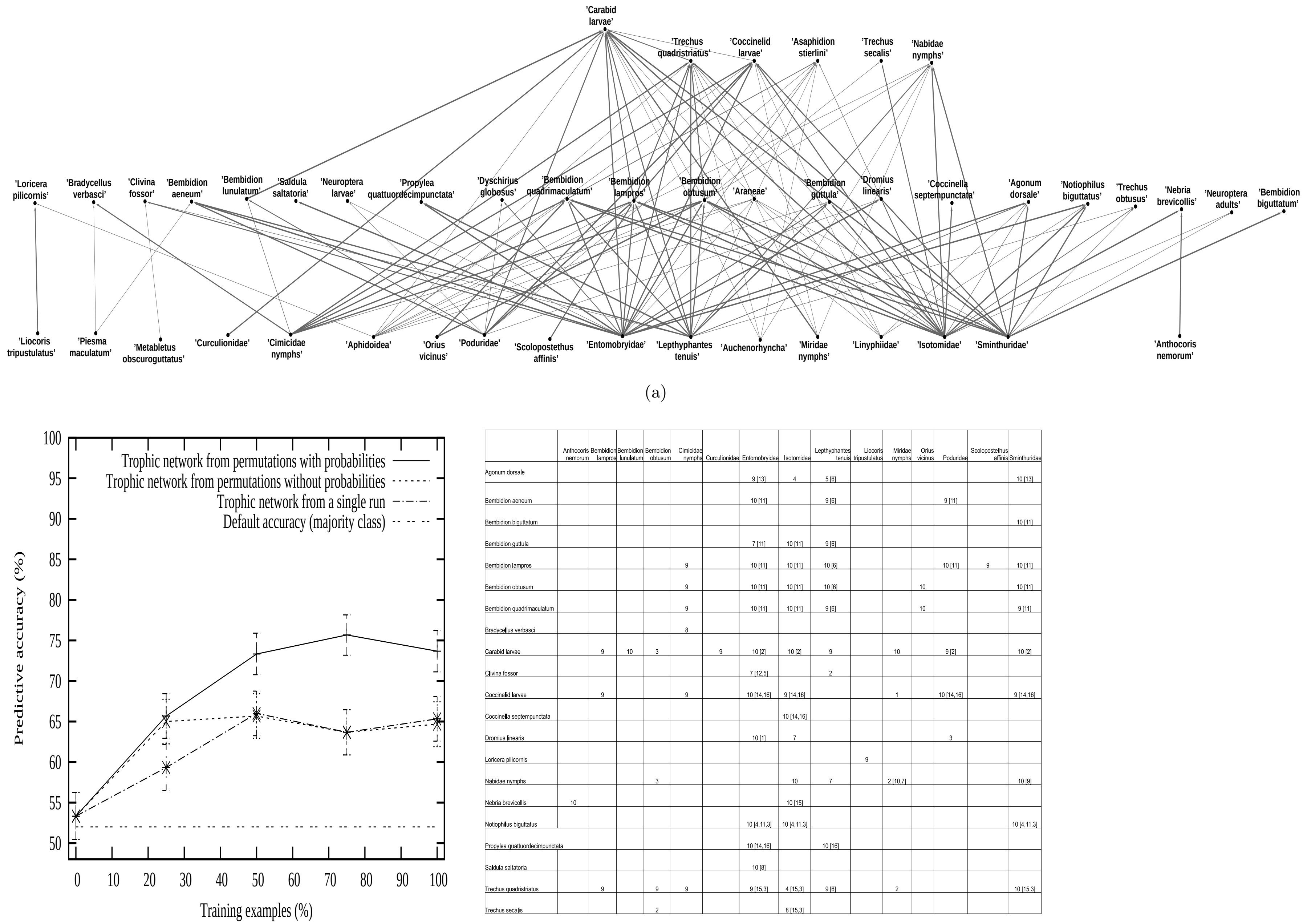
abundance(X, S, up):- predator(X), co_occurs(S, X, Y), bigger_than(X, Y), abundance(Y, S, up),eats(X, Y).

Similarly, a rule for $abundance(X, S, down)$ can be defined. This Prolog rule expresses the inference that following a perturbation in the eco-system (caused by the management), the increased (or decreased) abundance of species X at site S can be explained by the fact that X eats species Y which is further down in the food chain and the abundance of species Y is increased (or decreased). It also includes additional conditions to constraint the search for abducible predicate $eats(X, Y)$, i.e. X should be a predator, X and Y should co-occur and that X should be bigger than Y. Predicates $predator(X)$, $co\_occurs(S, X, Y)$ and $bigger\_than(X, Y)$ are provided as part of the background knowledge. Given this model and the observable data, Progol 5 generates a set of ground abductive hypotheses in the form of 'eats' relations between species. This set of ground hypotheses can be visualised as a network of trophic links (food webs) as shown in Figure 1a. In this network a ground fact $eats(a, b)$ is represented by a trophic link from b to a.

4 Probability estimation and evaluation of hypotheses

In order to get probability estimates for ground hypotheses, we use a technique which is based on direct sampling from the hypothesis space. In some ILP systems, including Progol, training examples also act as seeds to define the hypotheses space (e.g. a most specific clause is build from the next positive example). Hence, different permutations of the training examples define different parts of the hypothesis space. We use this property to sample from the hypothesis space by random permutations of the training data. Probability of ground hypotheses can be estimated based on the frequency of occurrence when random permutations of the training data (and hence different seeds for defining the hypothesis space) are considered. Using this technique, the thickness of trophic links in Figure 1a represent probabilities which are estimated based on the frequency of occurrence from 10 random permutations of the training data.

In order to empirically evaluate the hypothetical trophic links, we use leave-one-out cross-validation test on the observable data for species in the network, i.e. leaving out the abundance of each predator at each site and trying to predict whether the abundance is up or down, given the trophic network generated from the rest of the data. For the trophic network with probabilities, we first calculate the relative frequencies of hypotheses which imply that the abundance of the test example e is *up* or *down*. Let $p_{\uparrow}(e)$ be the relative frequency of hypotheses which imply e is *up* and $p_{\downarrow}(e)$ is defined analogously. If $p_{\uparrow}(e) > p_{\downarrow}(e)$ then we predict that the abundance of the test example e is *up* and otherwise it is *down*. $p_{\uparrow}(e)$ and $p_{\downarrow}(e)$ can be calculated using a Stochastic Logic Program (SLP) with the rules for $abundance(X, S, up)$ and $abundance(X, S, down)$ (as described in Section 3)



(a)

(b)

(c)

	Anthocoris nemorum	Bembidion lampros	Bembidion lunulatum	Bembidion obtusum	Cimicidae nymphs	Curculionidae	Entomobryidae	Isotomidae	Lepthyphantes tenuis	Liocoris tripustulatus	Miridae nymphs	Orius vicinus	Poduridae	Scolopostethus affinis	Sminthuridae
Agonum dorsale							9 [13]	4	5 [6]						10 [13]
Bembidion aeneum							10 [11]		9 [6]				9 [11]		
Bembidion biguttatum															10 [11]
Bembidion guttula							7 [11]	10 [11]	9 [6]						
Bembidion lampros					9		10 [11]	10 [11]	10 [6]				10 [11]	9	10 [11]
Bembidion obtusum					9		10 [11]	10 [11]	10 [6]			10			10 [11]
Bembidion quadrimaculatum					9		10 [11]	10 [11]	9 [6]			10			9 [11]
Bradycellus verbasci					8										
Carabid larvae		9	10	3		9	10 [2]	10 [2]	9		10		9 [2]		10 [2]
Clivina fossor							7 [12,5]		2						
Coccinelid larvae		9			9		10 [14,16]	9 [14,16]			1		10 [14,16]		9 [14,16]
Coccinella septempunctata								10 [14,16]							
Dromius linearis							10 [1]	7					3		
Loricera pilicornis										9					
Nabidae nymphs				3				10	7		2 [10,7]				10 [9]
Nebria brevicollis	10							10 [15]							
Notiophilus biguttatus							10 [4,11,3]	10 [4,11,3]							10 [4,11,3]
Propylea quattuordecimpunctata							10 [14,16]		10 [16]						
Saldula saltatoria							10 [8]								
Trechus quadristriatus		9		9	9		9 [15,3]	4 [15,3]	9 [6]		2				10 [15,3]
Trechus secalis				2				8 [15,3]							

Fig. 1. a) Hypothetical trophic network (food web) constructed by A/ILP. Thickness of trophic links represent probabilities which are estimated based on the frequency of occurrence from 10 random permutations of the training data. **b)** Predictive accuracies of probabilistic trophic network vs. non-probabilistic networks from leave-one-out cross-validation tests. **c)** Tabulated trophic links for some prey (columns) and predator (rows) species combination in Figure 1a. Each pairwise hypothesised link has a strength (i.e. frequency between 1 to 10) followed by references (in square brackets) in the literature (listed in Appendix) supporting the link.

and the abduced predicates *eats* with probabilities which correspond to their frequencies. Figure 1b compares the predictive accuracy of probabilistic and non-probabilistic networks, i.e. networks generated from 10 random permutations or from a single run. The results suggest that the predictive accuracies for the non-probabilistic networks are increased from around 65% to around 75% for the probabilistic network when more than 50% of the training data are provided. In all cases the predictive accuracies are significantly higher than the default accuracy of the majority class (i.e. *down*).

The trophic network in Figure 1a has been examined by the domain experts and corroboration of many of the links in the literature have been found. Table 1c is a tabular representation for some prey (columns) and predator (rows) species combination in Figure 1a. Each pairwise hypothesised link has a strength (i.e. frequency between 1 to 10) followed by references (in square brackets) in the literature (listed in Appendix) supporting the link. In this table, only prey/predators are shown which have at least one link with strength more than or equal to 7. This table shows that many of the links, suggested by the model, are corroborated by the literature. In particular, links in the model ascribed with high probability are shown to correspond well with those having multiple references in the literature. In some cases novel high probability links are suggested, which could be tested.

5 Conclusions

We find that machine learning, using A/ILP, produced a convincing food web from sample ecological data. Many of the important abduced trophic links are supported either by the literature or the expert knowledge of agricultural ecologists. This food web representing probabilistic interactions between species can readily be interpreted by Ecologists and the logical framework for learning trophic links can be openly discussed, a priori, and the hypothesised links are not an abstract, statistical product of the data. Two aspects of the use of A/ILP in this paper are particularly novel. Firstly, unlike previous applications of A/ILP, the abductive predicate 'eats' is entirely undefined before the start of the learning process. This setting is close to the classic hard problem of predicate invention within ILP. The second novel aspect of the approach relates to the assignment of probabilities to hypothetical 'eats' facts based on their frequency of occurrence when randomly sampling the hypothesis space. The resulting probabilistic network is a compact summary of the hypothesis space with a posterior distribution which could be viewed as a Bayes predictor and is expected to have lower error. The results of cross-validation tests suggest that the trophic networks with probabilities have higher predictive accuracies compared to the networks without probabilities. In this paper we have reported the predictive accuracies for binary classification. However, we have also used expected utilities implemented as Decision-Theoretic Logic Programs (DTLPs) [3] for estimating R values (treatment effect ratio as described in Section 2). Initial results suggest that using probabilities leads to reduced mean square errors when estimating R values in cross-validation tests. The probabilistic trophic network together with the expected utility approach can be viewed as a Decision-Theoretic representation which we call an Acyclic Expectation Network (AEN). We intend to study different aspects of this representation in a follow up paper.

Acknowledgements We would like to thank G. Caron-Lormier, J. Chen, S. Dunbar and V. Lesk for useful discussions and advise. This work is funded by Syngenta Ltd.

References

1. D.A. Bohan, C.W.H. Boffey, D.R. Brooks, S.J. Clark, A.M. Dewar, L.G. Firbank, A.J. Haughton, C. Hawes, M.S. Heard, M.J. May, et al. Effects on weed and invertebrate abundance and diversity of herbicide management in genetically modified herbicide-tolerant winter-sown oilseed rape. *Proceedings of the Royal Society B: Biological Sciences*, 272(1562):463, 2005.
2. G. Caron-Lormier, D.A. Bohan, C. Hawes, A. Raybould, A.J. Haughton, and R.W. Humphry. How might we model an ecosystem? *Ecological Modelling*, 220(17):1935–1949, 2009.
3. J. Chen and S. Muggleton. Decision-theoretic logic programs. In *Proceedings of ILP*, 2009.
4. T.G. Dietterich. Machine learning in ecosystem informatics and sustainability. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, Calif.: IJCAI*, pages 8–13, 2009.
5. P.A. Flach and A.C. Kakas. *Abduction and Induction: Essays on their relation and integration*. Springer Netherlands, 2000.
6. AJ Haughton, GT Champion, C. Hawes, MS Heard, DR Brooks, DA Bohan, SJ Clark, AM Dewar, LG Firbank, JL Osborne, et al. Invertebrate responses to the management of genetically modified herbicide–tolerant and conventional spring crops. ii. within-field epigeal and aerial arthropods. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1439):1863, 2003.
7. S. Muggleton and C. Bryant. Theory completion using inverse entailment. *Inductive Logic Programming*, pages 130–146, 2000.
8. A. Tamaddoni-Nezhad, R. Chaleil, A. Kakas, and S. Muggleton. Application of abductive ILP to learning metabolic network inhibition from temporal data. *Machine Learning*, 64(1):209–230, 2006.

Appendix: References used for corroboration in Table 1c

1. K.N.A. Alexander. The invertebrates of living and decaying timber in britain and ireland–a provisional annotated checklist. *English Nature Research Reports*, 467:1–142, 2002.
2. T. Bauer. Prey-capture in a ground-beetle larva. *Animal Behaviour*, 30(1):203–208, 1982.
3. J.R. Bell, R. Andrew King, D.A. Bohan, and W.O.C. Symondson. Spatial co-occurrence networks predict the feeding histories of polyphagous arthropod predators at field scales. *Ecography*, 33(1):64–72, 2010.
4. K. Berg. *The role of detrital subsidies for biological control by generalist predators evaluated by molecular gut content analysis.* PhD thesis, Universitäts-und Landesbibliothek Darmstadt, 2007.
5. K. Desender and M. Pollet. Ecological data on clivina fossor(coleoptera, carabidae) from a pasture ecosystem. ii. reproduction, biometry, biomass, wing polymorphism and feeding ecology. *REV. ECOL. BIOL. SOL.*, 22(2):233–246, 1985.
6. A. Dinter. Intraguild predation between erigonid spiders, lacewing larvae and carabids. *Journal of Applied Entomology*, 122(1-5):163–167, 1998.
7. J.D. Lattin. Bionomics of the nabidae. *Annual review of entomology*, 34(1):383–400, 1989.
8. J. Lindsey. Ecology of Commanster, http://www.commanster.eu/commanster/insects/bugs/spbugs/saldula.saltatoria.html.
9. X. Pons, B. Lumbierres, and R. Albajes. Heteropterans as aphid predators in inter-mountain alfalfa. *European Journal of Entomology*, 106(3):369–378, 2009.
10. C.W. Schaefer and A.R. Panizzi. *Heteroptera of economic importance.* CRC, 2000.
11. KD Sunderland. The diet of some predatory arthropods in cereal crops. *Journal of Applied Ecology*, 12(2):507–515, 1975.
12. K.D. Sunderland, G.L. Lovei, and J. Fenlon. Diets and reproductive phenologies of the introduced ground beetles harpalus-affinis and clivina-australasiae (coleoptera, carabidae) in new-zealand. *Australian Journal of Zoology*, 43(1):39–50, 1995.
13. S. Toft. The quality of aphids as food for generalist predators: implications for natural control of aphids. *European Journal of Entomology*, 102(3):371, 2005.
14. B.D. Turner. Predation pressure on the arboreal epiphytic herbivores of larch trees in southern england. *Ecological Entomology*, 9(1):91–100, 1984.
15. D.J. Warner, LJ Allen-Williams, S. Warrington, AW Ferguson, and IH Williams. Mapping, characterisation, and comparison of the spatio-temporal distributions of cabbage stem flea beetle (psylliodes chrysocephala), carabids, and collembola in a crop of winter oilseed rape (brassica napus). *Entomologia experimentalis et applicata*, 109(3):225–234, 2003.
16. D.C. Weber and J.G. Lundgren. Assessing the trophic ecology of the coccinellidae: their roles as predators and as prey. *Biological Control*, 51(2):199–214, 2009.