Towards machine learning of predictive models from ecological data

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
[2] UMR 1347 Agroécologie, Pôle Ecoldur, BP 86510 21065 Dijon CEDEX, France
[3] Syngenta Ltd, Bracknell, Berkshire, RG42 6EY, UK

Abstract. In a previous paper we described a machine learning approach which was used to automatically generate food-webs from national-scale agricultural data. The learned food-webs in the previous study consist of hundreds of ground facts representing trophic links between individual species. These species food-webs can be used to explain the structure and dynamics of particular eco-systems, however, they cannot be directly used as general predictive models. In this paper we describe the first steps towards this generalisation and present initial results on (i) learning general functional food-webs (i.e. trophic links between functional groups of species) and (ii) meta-interpretive learning (MIL) of general predictive rules (e.g. about the effect of agricultural management). Experimental results suggest that functional food-webs have at least the same levels of predictive accuracies as species food-webs despite being much more compact. The results also suggest that when the number of training examples are limited, functional food-webs have a higher predictive accuracy. In this paper we also present initial experiments where predicate invention and recursive rule learning in MIL are used to learn food-webs as well as predictive rules directly from data.

1 Introduction

Machine Learning has previously been used in ecology (e.g. [5]), however, ecological data-mining is relatively a new emerging subject. For example large-scale ecological data from agricultural systems are nowadays being produced to evaluate the impacts of new technology, such as genetically modified crops. These large-scale data can be also used to develop models for predicting the effects of perturbation on agro-ecosystems. We have recently demonstrated [10] that a logic-based machine learning method can be used to automatically generate plausible and testable food webs from ecological census data. Through a review of the literature, it was found that many of the learned trophic links are corroborated by the literature. In particular, links ascribed with high probability by machine learning are shown to correspond well with those having multiple references in the literature. In some cases novel, high probability links were suggested, and some of these have recently been tested and confirmed by subsequent

empirical studies [11]. The learned species food-webs described in [10] and [11] consist of hundreds of ground facts (ground abductive hypotheses) representing trophic links between individual species. These food-webs can be used to explain the structure and dynamics of particular eco-systems. However, species-based food-webs cannot be directly used as general predictive models unless they are generalised or used together with general (i.e. non-ground) predictive models. In this paper we describe the first steps towards this generalisation and present initial results on (i) learning general functional food-webs (i.e. trophic links between functional groups of species) and (ii) meta-interpretive learning of general predictive rules (e.g. about the effect of agricultural management).

2 Background and related work

To make good decisions about ecosystem management, e.g. the management of agricultural land for the optimal delivery of ecosystem services, it is necessary to have theories that predict the effects of perturbation on ecosystems. Network ecology, and in particular food-web approach, holds great promise as an approach to modeling and predicting the effects of perturbation on ecosystems. Networks of trophic links, also known as food-webs, which describe the flow of energy/biomass between species, are important for explaining ecosystem structure and dynamics. However, relatively few ecosystems have been studied through detailed food-webs because establishing predation relationships between the many hundreds of species in an ecosystem is expensive and in many cases impractical.

We have recently developed [10] a logic-based machine learning method which can be used to automatically generate plausible and testable food-webs from ecological census data. The initial food-web was learned from an extensive Vortis suction sampling of invertebrates from 257 arable fields across the UK as part of the Farm Scale Evaluations (FSE) of genetically modified, herbicide-tolerant (GMHT) crops. Using a technique based on calculating a treatment effect ratio [6], this abundance count data was converted into up/down information and was regarded as the primary observational data for the learning. The set of observable (or training) data are 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 was compiled from FSE data as detailed in [10]. The knowledge gap that we initially aimed to fill was a predation relationship between species. Thus, we declared abducible predicate $eats(X, Y)$ capturing the hypothesis that species X eats species Y. In order to use abduction, we also provided the rules which describe the observable predicate (*abundance*) in terms of the abducible predicate (*eats*):

```
abundance(X,S,Dir):-
    predator(X),
    bigger_than(X,Y),
    eats(X,Y),
    abundance(Y,S,Dir).
```

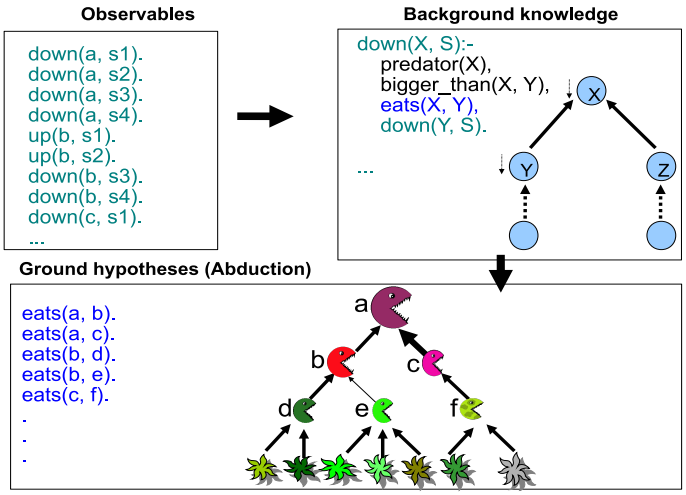
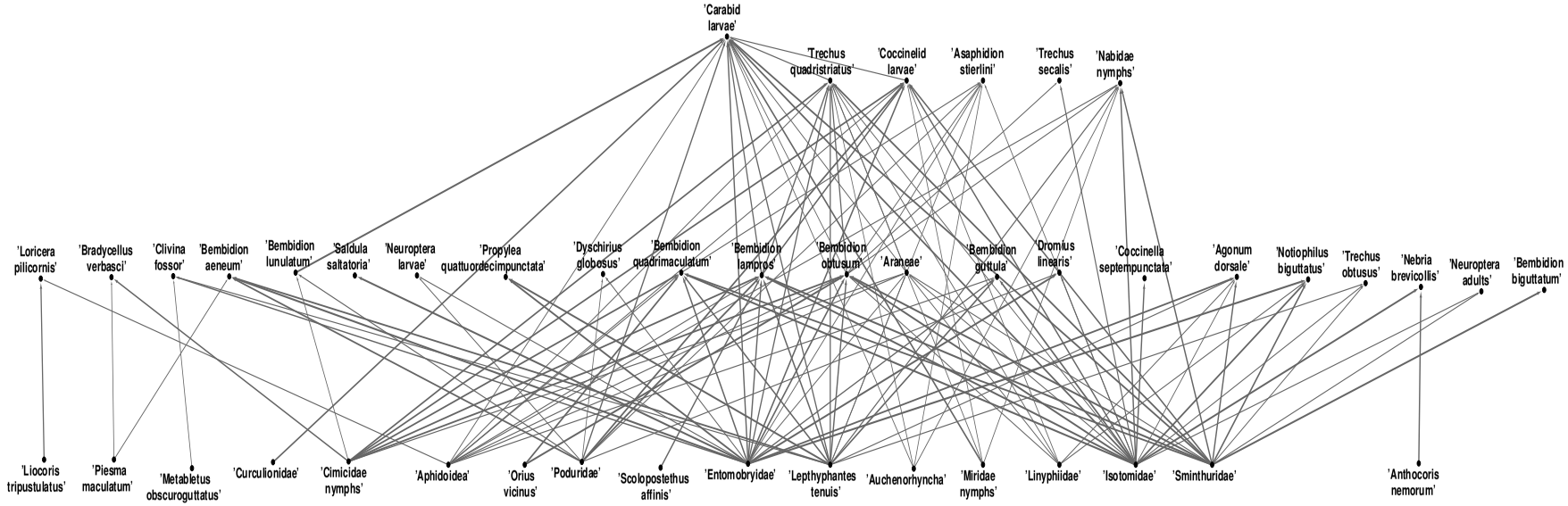


Fig. 1. Machine learning of species food-webs from ecological data using Abductive ILP.

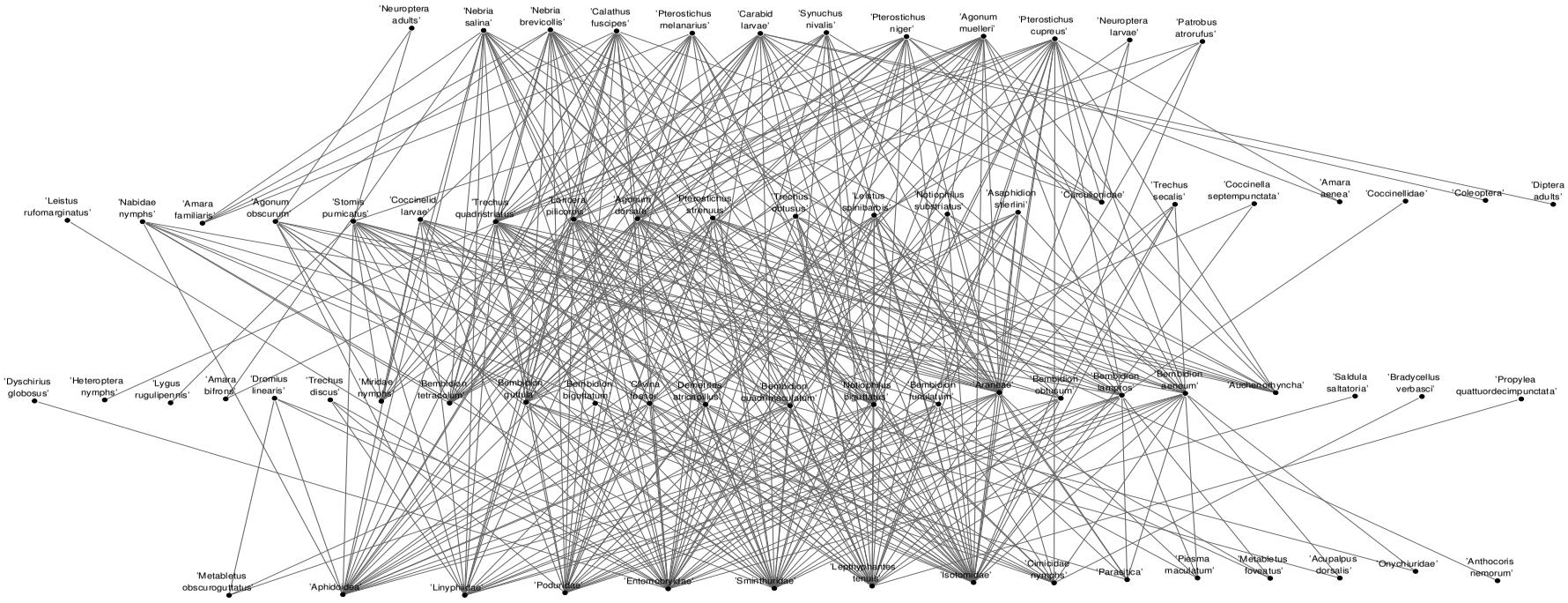
where *Dir* can be either *up* or *down*. This Prolog rule expresses the inference that following a perturbation in the ecosystem (caused by the management), the increased (or decreased) abundance of species X at site S can be explained by X eating species Y, which is lower in the food chain and the abundance of species Y is increased (or decreased). Given this model and the observable data, the Abductive ILP system Progol 5 [4] generates a set of ground abductive hypotheses in the form of 'eats' relations between species as shown in Figure 1. The set of ground hypotheses can be visualised as a network of trophic links (food-webs) as shown in Fig. 2a. In this network a ground fact $eats(a, b)$ is represented by a trophic link from species b to species a.

This food-web was examined [2] by domain experts from Rothamsted Research UK and it was found that many of the learned trophic links are corroborated by the literature. In particular, links ascribed with high probability by machine learning are shown to correspond well with those having multiple references in the literature. In some cases novel, high probability links were suggested, and some of these have recently been tested and confirmed by subsequent empirical studies. For example, in the hypothesised food-webs, some species of spiders always appeared as prey for other predators; a result that was unexpected because spiders are obligate predators. This hypothesis was tested using molecular analysis of predator guts and it was found that in this system spiders do appear to play an important role as prey [4]. Thus, even though some of the hypothesised links were unexpected, these were in fact confirmed later and this provided an extremely stringent test for the machine learning approach. The initial study was extended [11] by learning more complex food-webs from the national-scale

[4] Available from: http://www.doc.ic.ac.uk/~shm/Software/progol5.0/



(a)



(b)

Fig. 2. Species food-webs learned from farm-scale evaluations (FSE) of GMHT crops data collected using Vortis and pitfall sampling methods from 257 fields across the UK. **(a)** Food-web learned from the Vortis data and **(b)** food-web learned from merged Vortis and pitfall data.

pitfall sampling data (that was considerably larger than the initial Vortis data). Fig. 2b shows a species food-web learned from merged Vortis and pitfall data.

The learned species food-webs described in [10] and [11] consist of hundreds of ground facts (ground abductive hypotheses) representing trophic links between individual species. Figure 2 shows examples of species food-webs learned from the Vortis data and merged Vortis and pitfall data. These food-webs can be used to explain the structure and dynamics of particular eco-systems. However, species-based food-webs cannot be directly used as general predictive models unless they are generalised or used together with general (i.e. non-ground) predictive models. In this paper we describe the first steps towards this generalisation and present initial results on two different but related directions: (i) Learning general functional food-webs and (ii) Meta-interpretive learning (MIL) of general predictive rules.

Functional food-webs are more general than species food-web, a functional food-web represents interactions between functional groups of species while a species food-web represents trophic links between individual species as described in [2]. The machine learning of species food-web described above can be generalised by extending the approach to learn trophic links between functional groups of species, given the functional group memberships of species. Figure 4 shows examples of functional food-webs learned from the Vortis data and merged Vortis and pitfall data. More details about these functional food-webs and how they have been learned are given in the next section.

The machine learning of species food-webs (and functional food-webs) described above assume that the logical rules describing the problem, e.g. a rule which describes the observable predicate in terms of 'eats' relations (or the functional group memberships of species) are given as background knowledge. However, these may not be always available or they could be incomplete. In this paper we describe a new machine learning approach which allows automated discovery of trophic links as well as general predictive rules (and functional group memberships) directly from ecological data. This new setting requires both predicate invention and learning recursive rules which are not supported by most machine learning tools, including Progol which has been used for learning species and functional food-webs. On the other hand, Meta-interpretive Learning (MIL) [9, 8] is a new approach for predicate invention and recursive rule learning and can be used for learning ground hypotheses (e.g. trophic links) as well as non-ground hypotheses such as the general recursive rules. The MIL setting was initially used [9] for learning grammars from example sequences but was extended [8] to dyadic definite clauses. Unlike some ILP systems which either support predicate invention or recursion learning, MIL was shown to be a very efficient approach for predicate invention as well as learning recursive programs. For example, the ILP system ATRE has been used [1] for the discovery of mutual recursive patterns from text. However, ATRE does not support invention of first-order predicates. MIL is related to other studies where abduction has been used for predicate invention (e.g. [7]). One important feature of MIL, which distinguishes it from

other existing approaches, is that it introduces new predicate symbols which represent relations rather than new objects or propositions.

3 Machine learning of predictive models as an ILP problem

The machine learning tasks described above, i.e. learning of species food-webs, learning of functional food-webs and learning of predictive rules, can all be formally described by adopting the general ILP setting. ILP systems use given set of positive and negative examples $E = \{E^+ \cup E^-\}$ and background knowledge B to construct a hypothesis H that explains E^+ relative to B such that the extended theory is self-consistent:

- $B \cup H \models E^+$, and
- $B \cup H \cup E^-$ is consistent.

The components E, B and H are each represented as logic programs. In the case of machine learning of species and functional food-webs, abductive learning is used to learn ground hypotheses H (abducible) in the form of *eats* relations between species or functional group of species. In this case, background knowledge includes general rules $R \subseteq B$ which describe the observable examples in terms of the abducible predicate (e.g. see definition of *abundance/3* above).

In the case of machine learning of predictive rules, Meta-interpretive Learning (MIL) is used to learn a set of ground and non-ground hypotheses H. These include general predictive rules $R \subseteq H$ which describe the observable examples in terms of the invented predicates (e.g. see Fig. 6). In this case, background knowledge includes higher-order meta-rules $M \subseteq B$ which are activated during the proving of examples in order to generate hypotheses H.

Hence, machine learning of species and functional food-webs only require abductive learning where predicates can be separated into two disjoint sets: the observable predicates and the abducible predicates. In practice, observable predicates describe the empirical observations of the domain, i.e. *abundance* of species. The abducible predicates describe underlying relations in our model, i.e. *eats* relations between species or functional group of species, that are not observable directly but can, through the theory B, bring about observable information. By contrast, machine learning of predictive rules requires a combination of abduction and induction where the induction is needed to generate a set of non-ground hypotheses that contain universally quantified variables and can be used as general predictive rules. Meta-interpretive Learning (MIL) provides a tight integration of abduction and induction as described in [9].

In the following sections, machine learning of functional food-webs and meta-interepretive learning of predictive models are described with more details.

4 Machine learning of functional food-webs

In this section we explain how the approach for learning species food-webs has been extended for learning functional food-webs which are more general than

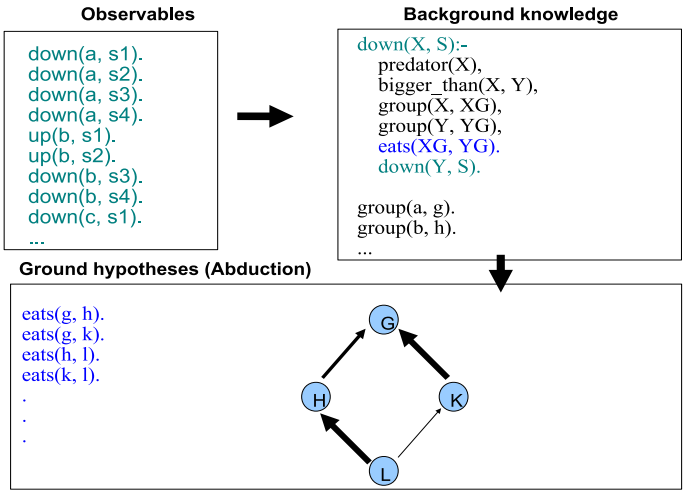


Fig. 3. Machine learning of functional food-webs from ecological data using Abductive ILP.

species food-webs. We also show that functional food-webs can lead to higher predictive accuracy than species food-webs.

As discussed in Section 2, species food-webs can be used to explain the structure and dynamics of a particular eco-system. However, functional food-webs (i.e. which represent trophic interactions between functional groups of species) are more important for predicting changes in agroecosystem diversity and productivity [3]. Given the background information on functional type of each species, trophic networks for functional groups can be also learned from ecological data using the machine learning approach described above.

As for the species food-webs, we need a rule which describes the observable predicate in terms of *eats* relation between functional groups:

abundance(X,S,Dir):-
predator(X),
bigger_than(X,Y),
ft(X,XFunc_ID),
ft(Y,YFunc_ID),
eats(XFunc_ID,YFunc_ID),
abundance(Y,S,Dir).

Given this new model and background information (functional types [3] of species in the form of $ft(X, XFunc\_ID)$) trophic networks can be constructed for functional groups in a learning setting similar to the one described above for individual species.

Figure 4a and 4b show functional food-webs learned from the Vortis data and from merged Vortis and pitfall data respectively. These food-webs are constructed by learning trophic interactions between functional groups rather than

individual species. Each functional group is represented by a species which can be viewed as an archetype for the functional group.

4.1 Empirical evaluation

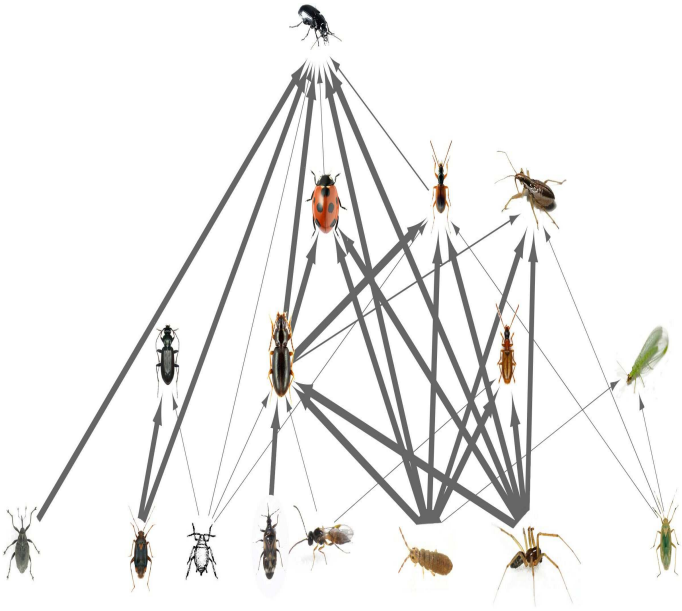
In this section we test the following null hypothesis:

Null hypothesis 1: A trophic network constructed by learning trophic links between functional groups has a lower predictive accuracy compared to the trophic network for individual species.

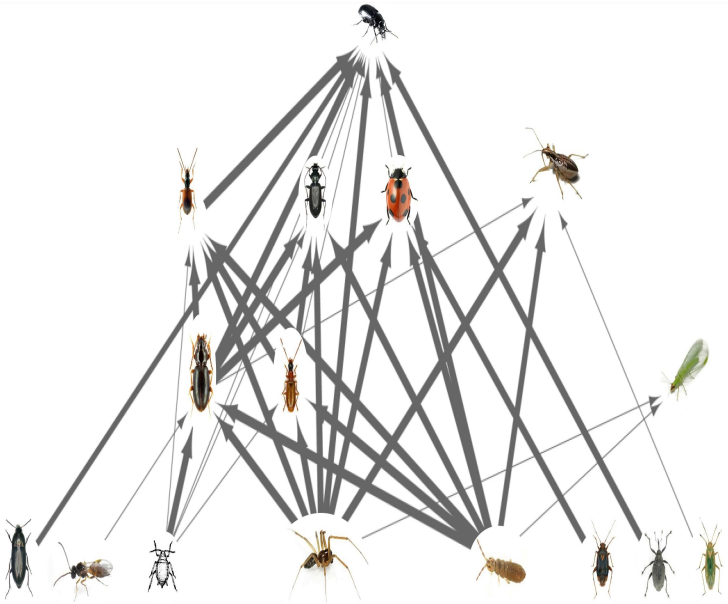
Materials and methods In this experiment Progol 5.0 [5] is used to abduce 'eats' relations between species and functional groups of species from observable data (i.e. up/down abundance of species at different sites). The observable data has been compiled from FSE data as described in [10]. The up/down abundance of species at different sites are then encoded as predicates $abundance(X, S, up)$ and $abundance(X, S, down)$. The background knowledge includes information about sites and species and Prolog rules for *abundance* as described in Sections 2 and 4. A probabilistic approach, called Hypothesis Frequency Estimation (HFE) [10], was used for estimating probabilities of hypothetical trophic links based on their frequency of occurrence when randomly sampling the hypothesis space. Using this technique, the thickness of trophic links in Figures 2 and 4 represent probabilities which are estimated based on the frequency of occurrence from 10 random permutations (a user selected parameter) of the training data (and hence different seeds for defining the hypothesis space). Relative frequencies are used in the same way probabilities are used in probabilistic ILP and the probabilistic inference is used to estimate probabilities of unseen data. For example, the probability $p(abundance(a, s, up))$ can be estimated by relative frequency of hypotheses which imply a at site s is up. Similarly, $p(abundance(a, s, down))$ can be estimated and by comparing these probabilities we can decide to predict whether the abundance is up or down. This method has been used in the leave-one-out experiments in [10] to compare the predictive accuracies of probabilistic trophic networks vs non-probabilistic trophic networks. We use similar leave-one-out experiments to compare the predictive accuracies of functional food-web vs species food-web from Vortis data (shown in Fig. 2a and 4a). Other materials and methods are similar to the experiments in [10], but we also include the rule and background knowledge for learning functional food-webs as described above.

Results and discussion The predictive accuracies of the functional and species food-webs are shown in Figure 5. According to this figure, the difference between the predictive accuracies of the probabilistic network for the species food-web and the functional food-web are not significant when more than 50% of training examples are provided. However, the predictive accuracy of functional food-web

[5] Available from: http://www.doc.ic.ac.uk/~shm/Software/progol5.0/



(a)



(b)

Fig. 4. Functional food-webs learned from the Vortis data **(a)** and from merged Vortis and pitfall data **(b)**. Each group in the functional food-web is represented by a species which can be viewed as an archetype for that functional group.

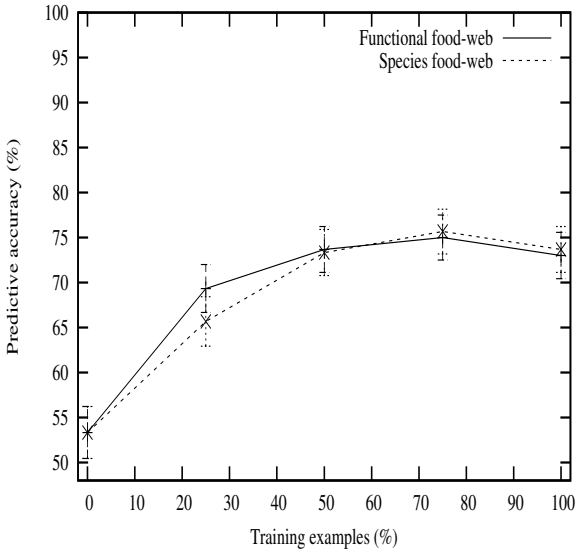


Fig. 5. Predictive accuracies of functional food-web vs. species food-web from leave-one-out cross-validation tests.

is significantly higher (p-value of 0.004 from t-test) than the predictive accuracy of species food-web when 25% of training examples is available. This result suggest that when the number of training examples are limited, the functional food-web (which is more general) has a higher predictive accuracy compared to the species food-web. The null hypothesis 1 can therefore be refuted. In general these experiments confirm that a network which is constructed by learning trophic links between functional groups is at least as accurate as the trophic network for individual species despite being less complex (i.e. having less nodes and edges). Note that in this experiment we used a leave-one-out test strategy and evaluated both food-webs on the data from the same agricultural system. The higher predictive accuracy of the functional food-web would be more evident if the food-webs are evaluated on a different agricultural system (e.g. different crops, climate etc) and we intend to demonstrate this as a future work.

5 Meta-interepretive learning of predictive models

In the machine learning setting described in the previous section, the recursive rules describing the observable predicate ($abundance/3$) and the functional groups were provided as part of background knowledge. However, these information may not be always available or they could be incomplete. Here we describe a new machine learning setting where these information could be learned directly from data. This new learning setting requires predicate invention and recursive rule learning and we use Meta-interpretive Learning (MIL) [9, 8] for this purpose. In the MIL framework described in this paper, predicate invention is conducted

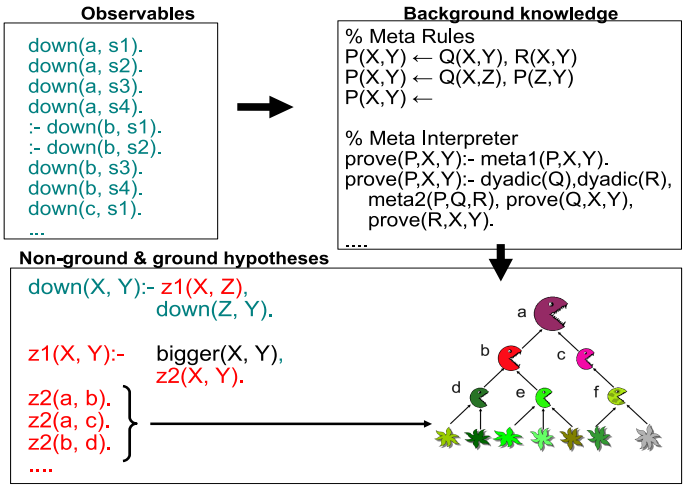


Fig. 6. Meta-interpretive learning of ground hypotheses (i.e. food-web) and non-ground hypotheses (i.e. prediction rule for down regulation) learned from a simplified food-web data (abundance of 6 species from 4 different sites). Predicates $z1$ and $z2$ are invented predicates, where $z2$ represents 'eats' relation.

via construction of substitutions for meta-rules employed by a meta-interpreter. The use of the meta-rules clarifies the declarative bias being employed. New predicate names are introduced as higher-order skolem constants, a finite number of which are introduced during every iterative deepening of the search as described in [8].

MIL is a technique which supports efficient predicate invention and learning of recursive logic programs built as a set of metalogical substitutions by a modified Prolog meta-interpreter which acts as the central part of the ILP learning engine. The meta-interpreter is provided by the user with meta-rules which are higher-order expressions describing the forms of clauses permitted in hypothesised programs. The meta-interpreter attempts to prove the examples and, for any successful proof, saves the substitutions for existentially quantified variables found in the associated meta-rules. When these substitutions are applied to the meta-rules they result in a first-order definite program which is an inductive generalisation of the examples.

Fig. 6 shows meta-interpretive learning of ground hypotheses (i.e. food-web) and non-ground hypotheses (i.e. prediction rule for down regulation) learned from a simplified ecological data on down regulation of species following an agricultural management. MIL works by proving examples via meta-interpreter. This figure shows three higher-order meta-rules which are activated during the proof in order to generate the hypotheses shown in this figure. These hypotheses include non-ground rules and ground facts. Predicates $z1$ and $z2$ are invented predicates, where $z2$ represents 'eats' relation. Hence, the ground facts $z2(a, c)$, $z2(c, f)$, etc represent the food-web which together with the non-ground rules for

'down' can be used for predicting down-regulation. The rule for 'down' shown in Fig. 6 is similar to the rule provided as background knowledge in the previous sections.

5.1 Empirical evaluation

In this section we test the following null hypothesis:

Null hypothesis 2: The MIL system $Metagol$ cannot outperform the ILP system $Prolog$ in learning prediction rules as well as trophic links from a simplified ecological data.

Materials and methods In this section we use $Metagol_D$ [6] to learn ground hypotheses (i.e. food-web) and non-ground hypotheses (i.e. prediction rule for down regulation) from a simplified food-web data consisting of abundance of 6 species from 4 different sites, as shown in Figure 6. We use leave-one-out experiments to compare the predictive accuracies of $Metagol$ vs $Prolog$. $Prolog$ has been also tested in an enhanced mode where the food-web is provided as background knowledge ($Prolog + foodweb$).

Results and discussion Fig. 7 compares predictive accuracies of $Metagol$ vs $Progol$ vs $Prolog + foodweb$ in learning ground hypotheses (i.e. food-web) and non-ground hypotheses (i.e. prediction rule for down regulation) from the simplified food-web data described above. According to this, the predictive accuracies of $Metagol$ are significantly higher than $Progol$. The accuracy of an enhanced Progol setting ,where the food-web is provided as background knowledge ($Prolog + foodweb$), reaches around 75%. However, $Metagol$, which can learn both food-web and prediction rules, reaches an accuracy of 100%. These results suggest that $Metagol$ can learn the recursive rules and the food-web at the same time but it is difficult for Progol to learn these recursive rules directly from data even if the food-web structure is provided as background knowledge. Fig. 7 also compares timings of $Metagol$ vs $Progol$. According to this figure Progol is significantly faster. But it should be noted that unlike Progol which fails to learn any recursive rule, $Metagol$ is learning and evaluating recursive rules.

6 Conclusions

We presented initial results on machine learning of general predictive models from ecological data. We have considered two different but related directions to extend our previous approach for machine learning of food-webs: (i) learning functional food-webs and (ii) meta-interpretive learning (MIL) of general predictive rules. Experimental results suggested that functional food-webs have at

[6] Available from: http://ilp.doc.ic.ac.uk/metagolD

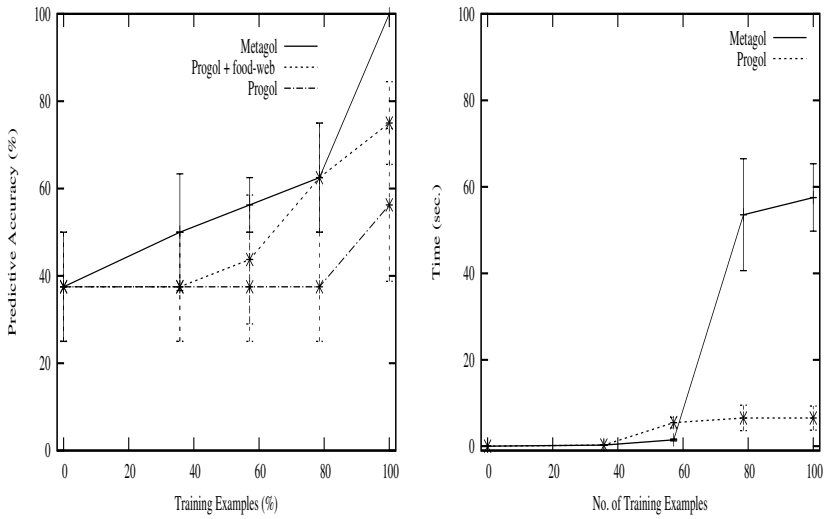


Fig. 7. Predictive accuracies and timing of Metagol vs Progol in learning ground hypotheses (i.e. food-web) and non-ground hypotheses (i.e. prediction rule for down regulation) from a simplified ecological data as shown in Fig. 6.

least the same levels of predictive accuracies as species food-webs and could also lead to higher predictive accuracy when the number of training examples are limited. We also presented initial results on using MIL for machine learning of predictive models. These results confirm that MIL can re-construct a simplified food-web and learn recursive predictive rules directly from data. In this paper we only demonstrated MIL on a simplified species food-web. However, initial experiments suggest that it is also possible to learn functional food-webs as well as functional groups membership directly from data using predicate invention.

References

1. Margherita Berardi and Donato Malerba. Learning recursive patterns for biomedical information extraction. In *Proceedings of the Int. Conf. on Inductive Logic Programming (ILP 2006)*, pages 79–93. Springer, 2007.
2. 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*, 6(12), 2011.
3. 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.
4. J Davey, I Vaughan, R Andrew King, J Bell, D Bohan, M Bruford, J Holland, and W Symondson. Intraguild predation in winter wheat: prey choice by a common epigeal carabid consuming spiders. *Journal of Applied Ecology*, 50(1):271–279, 2013.

5. 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.
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. K. Inoue, K. Furukawa, and I. Kobayashiand H. Nabeshima. Discovering rules by meta-level abduction. In L. De Raedt, editor, *Proceedings of the Nineteenth International Conference on Inductive Logic Programming (ILP09)*, pages 49–64, Berlin, 2010. Springer-Verlag. LNAI 5989.
8. S. H. Muggleton and D. Lin. Meta-interpretive learning of higher-order dyadic datalog: Predicate invention revisited. In *Proceedings of the 23rd International Joint Conference Artificial Intelligence (IJCAI 2013)*, pages 1551–1557, 2013.
9. S. H. Muggleton, D. Lin, N. Pahlavi, and A. Tamaddoni-Nezhad. Meta-interpretive learning: application to grammatical inference. *Machine Learning*, 94:25–49, 2014.
10. A. Tamaddoni-Nezhad, D. Bohan, A. Raybould, and S.H. Muggleton. Machine learning a probabilistic network of ecological interactions. In *Proceedings of the 21st International Conference on Inductive Logic Programming*, LNAI 7207, pages 332–346, 2012.
11. A. Tamaddoni-Nezhad, G. Milani, A. Raybould, S. Muggleton, and D. Bohan. Construction and validation of food-webs using logic-based machine learning and text-mining. *Advances in Ecological Research*, 49:225–289, 2013.