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
{a.tamaddoni-nezhad,s.muggleton}@imperial.ac.uk
[2] UMR 1347 Agroécologie, Pôle Ecoldur, BP 86510, 21065 Dijon Cedex, France
david.bohan@dijon.inra.fr
[3] Syngenta Crop Protection AG, Schwarzwaldallee 215, CH-4058 Basel, Switzerland
alan.raybould@syngenta.com

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. 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 [12] 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 [13].

The learned species food-webs described in [12,13] 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. This is mainly because establishing predation relationships between the many hundreds of species in an ecosystem is resource intensive, requiring considerable investment in field observation and laboratory experimentation.

We have recently developed [12] 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 [12]. 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 ($eats$):
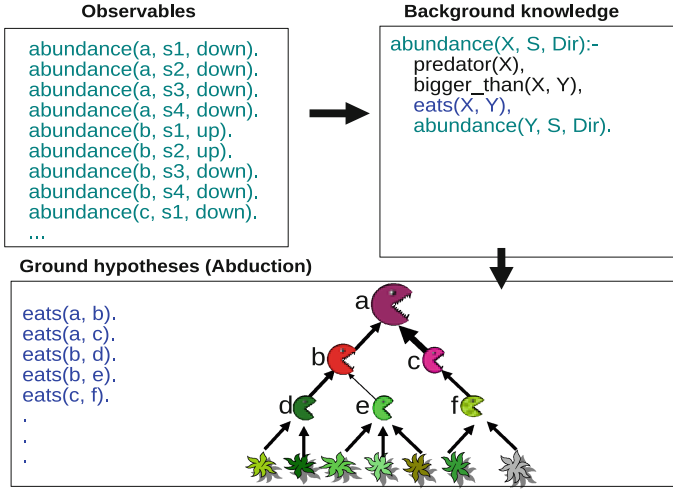
Fig. 1. Machine learning of species food-webs from ecological data using Abductive ILP.

$abundance(X,S,Dir):-$
   $predator(X),$
   $bigger\_than(X,Y),$
   $eats(X,Y),$
   $abundance(Y,S,Dir).$

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). It also includes additional conditions to constraint the search for abducible predicate $eats(X,Y)$, i.e. X should be a predator and bigger than Y. Predicates $predator(X)$ and $bigger\_than(X,Y)$ are provided as part of the background knowledge. Given this model and the observable data, the Abductive ILP system Progol 5 [10] was used to generate a set of ground abductive hypotheses in the form of 'eats' relations between species as shown in Fig. 1. These abductive hypotheses are automatically generated by matching the given information to the rule in order to abduce a fact which explains the observations. In this example, given the inputs, abduction will generate the hypotheses that a particular species a eats a particular species b. 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.

A probabilistic approach, called Hypothesis Frequency Estimation (HFE) [12], 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 Fig. 2 represent
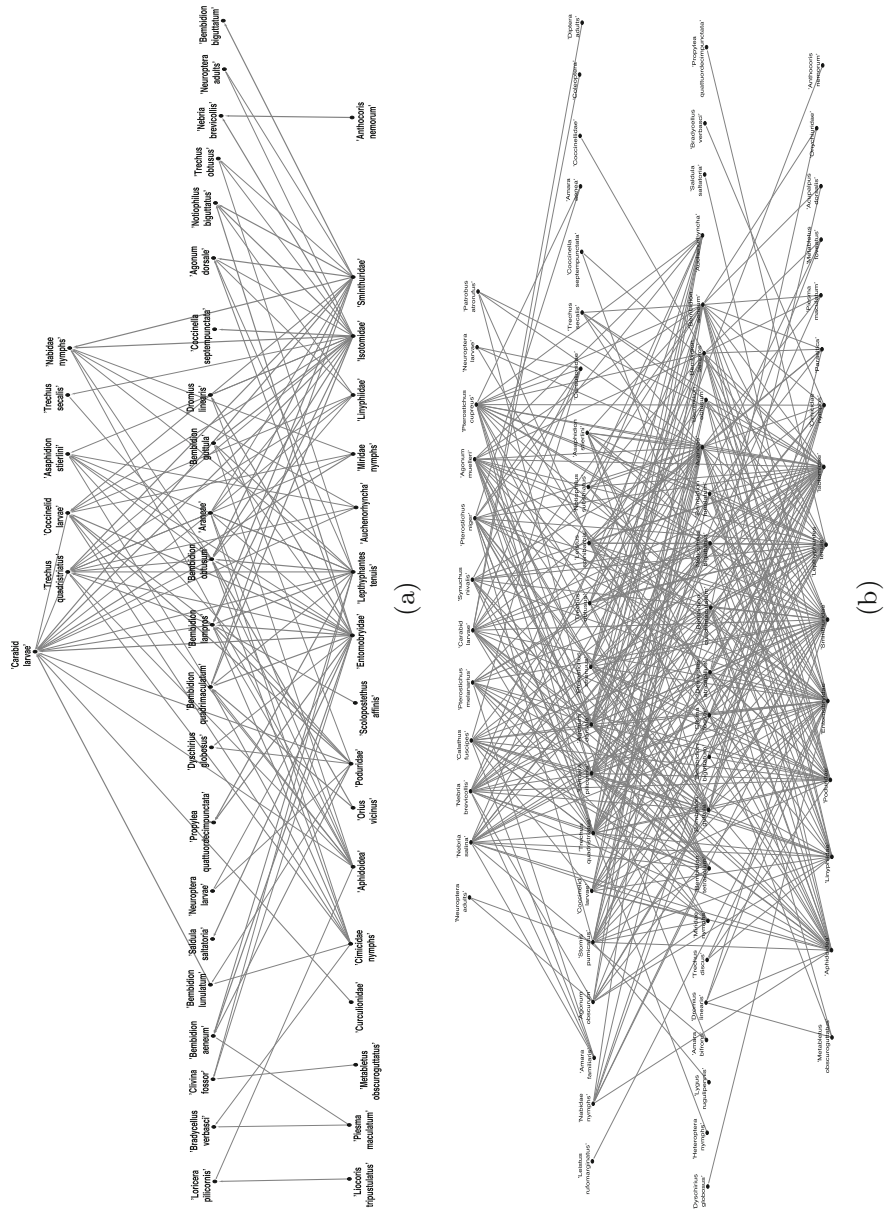
Fig. 2. Species food-webs learned from farm-scale evaluations (FSE) of GMHT crops data collected using Vortis and pitfall sampling methods from 257 fields across the UK. **(a)** Food-web learned from the Vortis data and **(b)** food-web learned from merged Vortis and pitfall data. Thickness of trophic links represents probabilities which are estimated using HFE [12].

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).

The food-web in Fig. 2a 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 [13] by learning more complex food-webs from the national-scale pitfall sampling data (that was considerably larger than the initial Vortis data). Vortis protocol is a kind of suction sampling which gathers a proportion of species present at the time of sampling, whereas pitfall protocol uses trap sampling over a period of time. Figure 2b shows a species food-web learned from merged Vortis and pitfall data.

The learned species food-webs described in [12,13] consist of hundreds of ground facts (ground abductive hypotheses) representing trophic links between individual species (see Fig. 2). 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, this information may not be always available or it 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) [9,11] 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 also 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. 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.
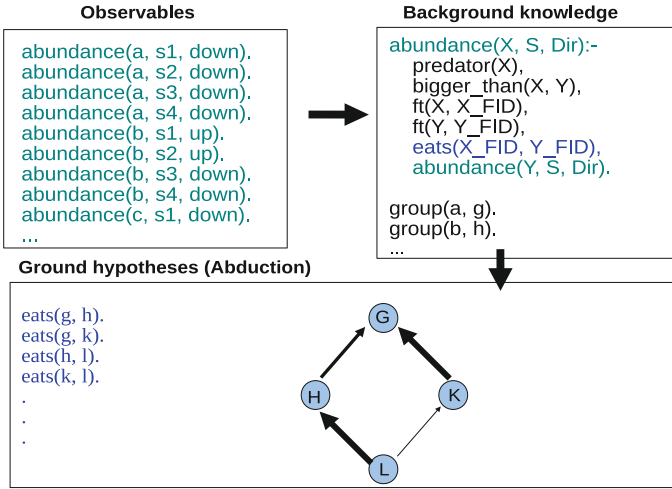
Fig. 3. Machine learning of functional food-webs from ecological data using Abductive ILP.

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. For this purpose, we use MIL which 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 species food-webs. We also show that functional food-webs can lead to higher predictive accuracy than species food-webs.

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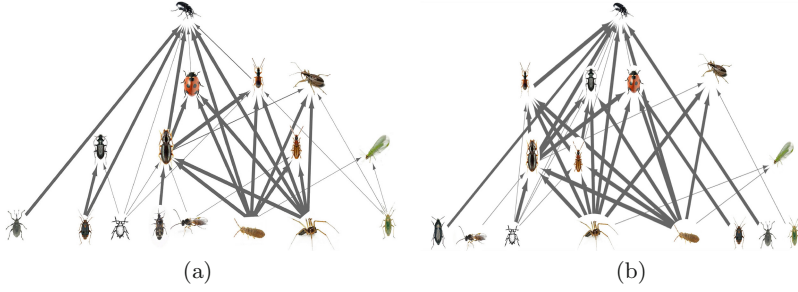
Fig. 4. Functional food-webs (a) learned from the Vortis data and (b) learned from merged Vortis and pitfall data. Each group in the functional food-web is represented by a species which can be viewed as an archetype for that functional group.

trophic networks for functional groups can be also learned from ecological data using the machine learning approach described above (See Fig. 3).

Here 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,X_FID),
  ft(Y,Y_FID),
  eats(X_FID,Y_FID),
  abundance(Y,S,Dir).

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

Figure 4a and b 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 food-web constructed by learning trophic links between functional groups has a lower predictive accuracy compared to the food-web for individual species.
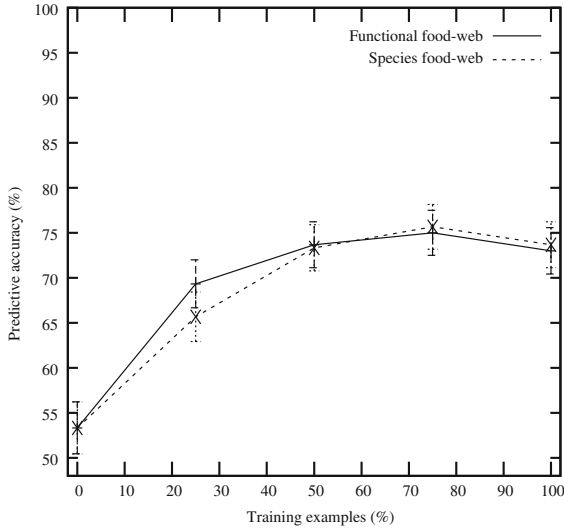
Fig. 5. Predictive accuracies of functional food-web vs. species food-web from leave-one-out cross-validation tests.

Materials and Methods. In this experiment Progol 5.0[1] 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 [12]. 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 Sects. 2 and 4. Hypothesis Frequency Estimation (HFE) [12], was used for estimating probabilities of hypothetical trophic links based on the frequency of occurrence from 10 random permutations of the training data. Relative frequencies and probabilistic inference were also used to estimate probabilities of unseen data as described in [12]. 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 predict whether the abundance of a at site s is *up* or *down*. This method used in the leave-one-out experiments in [12] to compare the predictive accuracies of probabilistic trophic networks vs non-probabilistic trophic networks. We use the same leave-one-out test strategy (described in Fig. 5 in [12]) to compare the predictive accuracies of species food-web vs functional food-web from Vortis data (food-webs shown in Figs. 2a and 4a). Other materials and methods are similar to those used in the experiments in [12], but we also include the rule and background knowledge for learning functional food-webs as described above.

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

Results and Discussion. The predictive accuracies of the functional and species food-webs are shown in Fig. 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 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 suggests 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 this experiment confirms that a food-web which is constructed by learning trophic links between functional groups is at least as accurate as the food-web 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. We expect the higher predictive accuracy of the functional food-web to be more evident if the food-webs are evaluated on a different agricultural system (e.g. different crops, climate etc.) where different species (not present in the training data) may exist. In this case we expect the species food-web to have a lower predictive accuracy and we intend to demonstrate this in 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, this information may not be always available or it 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) [9,11] for this purpose. In the MIL framework described in this paper, predicate invention is conducted 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 [11]. Both predicate invention and the learning of recursion are implemented in MIL via metalogical substitutions with respect to a modified Prolog meta-interpreter which acts as the 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.
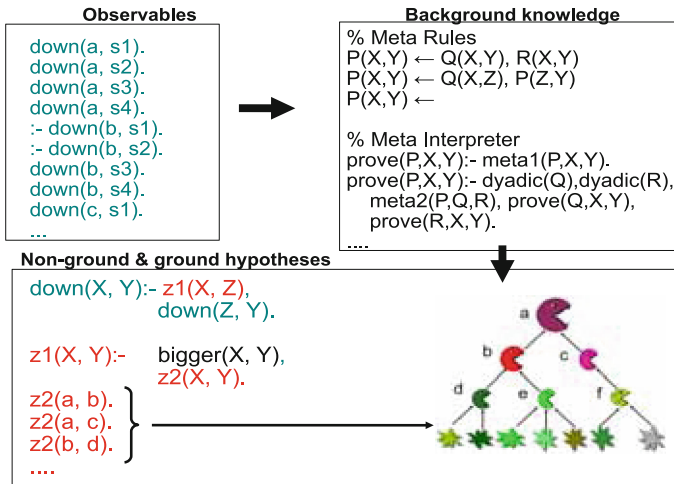
Fig. 6. Meta-interpretive learning of ground hypotheses (i.e. food-web) and non-ground hypotheses (i.e. prediction rule for down regulation) learned from a simplified food-web data (abundance of 6 species from 4 different sites). Predicates $z1$ and $z2$ are invented predicates, where $z2$ represents 'eats' relation.

Figure 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 comparable 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 *Progol* in learning prediction rules as well as trophic links from a simplified ecological data.

Materials and Methods. In this section we use $Metagol_D$[2] to learn ground hypotheses (i.e. food-web) and non-ground hypotheses (i.e. prediction rule for
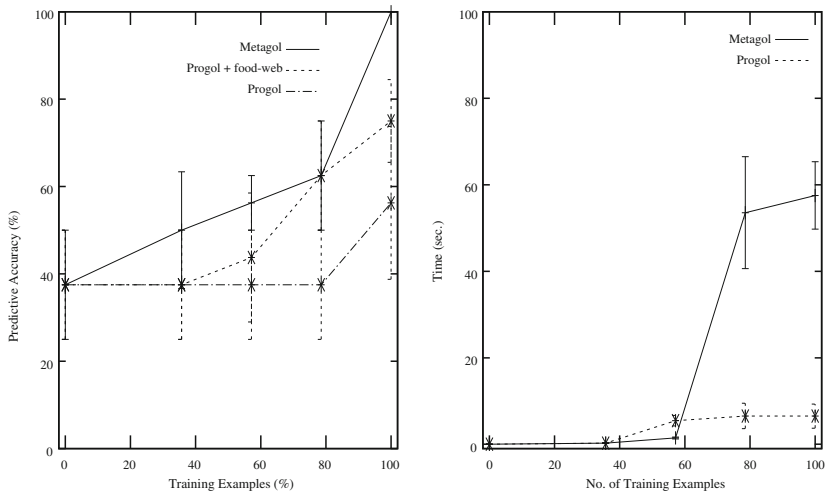
Fig. 7. Predictive accuracies and timing of Metagol vs Progol in learning ground hypotheses (i.e. food-web) and non-ground hypotheses (i.e. prediction rule for down regulation) from a simplified ecological data as shown in Fig. 6.

down regulation) from a simplified food-web data consisting of abundance of 6 species from 4 different sites, as shown in Fig. 6. We use a leave-one-out test strategy, similar to the one used in Sect. 4.1, to compare the predictive accuracies of $Metagol$ vs $Progol$. $Progol$ has been also tested in an enhanced mode where the food-web is provided as background knowledge ($Progol + foodweb$).

Results and Discussion. Figure 7 compares predictive accuracies of $Metagol$ vs $Progol$ vs $Progol + foodweb$ 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 ($Progol + foodweb$), 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. Figure 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 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. Berardi, M., Malerba, D.: Learning recursive patterns for biomedical information extraction. In: Muggleton, S.H., Otero, R., Tamaddoni-Nezhad, A. (eds.) ILP 2006. LNCS (LNAI), vol. 4455, pp. 79–93. Springer, Heidelberg (2007)
2. Bohan, D.A., Caron-Lormier, G., Muggleton, S.H., Raybould, A., Tamaddoni-Nezhad, A.: Automated discovery of food webs from ecological data using logic-based machine learning. PLoS ONE **6**(12), e29028 (2011)
3. Caron-Lormier, G., Bohan, D.A., Hawes, C., Raybould, A., Haughton, A.J., Humphry, R.W.: How might we model an ecosystem? Ecol. Model. **220**(17), 1935–1949 (2009)
4. Davey, J., Vaughan, I., King, R.A., Bell, J., Bohan, D., Bruford, M., Holland, J., Symondson, W.: Intraguild predation in winter wheat: prey choice by a common epigeal carabid consuming spiders. J. Appl. Ecol. **50**(1), 271–279 (2013)
5. Dietterich, T.G.: Machine learning in ecosystem informatics and sustainability. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI, Pasadena, Calif., pp. 8–13 (2009)
6. Haughton, A.J., Champion, G.T., Hawes, C., Heard, M.S., Brooks, D.R., Bohan, D.A., Clark, S.J., Dewar, A.M., Firbank, L.G., Osborne, J.L., et al.: Invertebrate responses to the management of genetically modified herbicide-tolerant and conventional spring crops. ii. within-field epigeal and aerial arthropods. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. **358**(1439), 1863 (2003)
7. Inoue, K., Furukawa, K., Kobayashi, I., Nabeshima, H.: Discovering rules by meta-level abduction. In: De Raedt, L. (ed.) ILP 2009. LNCS, vol. 5989, pp. 49–64. Springer, Heidelberg (2010)

8. Muggleton, S.H., Lin, D.: Meta-interpretive learning of higher-order dyadic datalog: predicate invention revisited. In: Proceedings of the 23rd International Joint Conference Artificial Intelligence (IJCAI 2013), pp. 1551–1557 (2013)

9. Muggleton, S.H., Lin, D., Pahlavi, N., Tamaddoni-Nezhad, A.: Meta-interpretive learning: application to grammatical inference. Mach. Learn. **94**, 25–49 (2014)

10. Muggleton, S.H., Bryant, C.H.: Theory completion using inverse entailment. In: Cussens, J., Frisch, A.M. (eds.) ILP 2000. LNCS (LNAI), vol. 1866, p. 130. Springer, Heidelberg (2000)

11. Muggleton, S.H., Lin, D., Tamaddoni-Nezhad, A.: Meta-interpretive learning of higher-order dyadic datalog: predicate invention revisited. Mach. Learn. 100, 49–73 (2015). doi:10.1007/s10994-014-5471-y

12. Tamaddoni-Nezhad, A., Bohan, D., Raybould, A., Muggleton, S.H.: Machine learning a probabilistic network of ecological interactions. In: Muggleton, S.H., Tamaddoni-Nezhad, A., Lisi, F.A. (eds.) ILP 2011. LNCS, vol. 7207, pp. 332–346. Springer, Heidelberg (2012)

13. Tamaddoni-Nezhad, A., Milani, G., Raybould, A., Muggleton, S., Bohan, D.: Construction and validation of food-webs using logic-based machine learning and text-mining. Adv. Ecol. Res. **49**, 225–289 (2013)