Using Machine Learning to Understand Top-Down Effects in an Ecosystem: Opportunities, Challenges, and Lessons Learned

Douglas A. Talbert
Computer Science
Tennessee Tech University
dtalbert@tntech.edu

Paul Tinker
Computer Science
Tennessee Tech University
pjtinker42@students.tntech.edu

Tom Crowther
Terrestrial Ecology
Netherlands Inst of Ecol.
T.Crowther@nioo.knaw.nl

Donald Walker
Biology
Tennessee Tech University
dmwalker@tntech.edu

Abstract

The soil decomposer community is a primary driver of carbon cycling in forest ecosystems. Understanding the processes that regulate this community is critical to our understanding of the global carbon cycle and fungal mediated impact on climate change. Inadequate statistical strength in traditional soil food web studies has limited our capacity to disentangle the cascading effect of top-level predators on the composition of complex fungal communities. We hypothesize that machine learning can help with this complex problem. This paper examines the opportunities for machine learning in this domain, presents initial results from such analysis, identifies challenges encountered with initial effort, and charts a path forward.

Introduction

The soil decomposer community is a primary driver of carbon cycling in forest ecosystems (Crowther et al. 2013). Understanding the processes that regulate this community is critical to our understanding of the global carbon cycle and fungal mediated impact on climate change (Wardle et al. 2004; Bardgett 2005; Van der Heijden et al. 2008).

To improve our understanding of the decomposer community, we performed a study in which an upland deciduous forest in West Virginia, USA, was simulated within a controlled laboratory experiment. Using this mesocosm environment, we measured how the removal of a top predator species (the red-backed salamander) affects the soil decomposer community. Specifically, the red-backed salamander has been described as a keystone species, due to its capacity to feed on soil invertebrate communities that regulate the activity of primary decomposers (fungi) in soil (Burton and Likens 1975).

Populations of red-backed salamanders are increasingly threatened by habitat fragmentation and land use change in temperate US forests. Generating a mechanistic understanding of how salamander removal affects decomposer communities is, therefore, critical for us to comprehend the impacts of future biodiversity loss on the functioning of temperate forest ecosystems. Yet inadequate statistical strength in traditional soil food web studies has limited our capacity to disentangle the cascading effect of salamanders on the composition of complex fungal communities.

Our hypothesis is that machine learning, specifically classification models, can help us (1) determine the strength of the relationship between the keystone species and other members of the decomposer community and (2) identify the nature of the relationship between community members (if a relationship does exist).

Background

Soil Decomposer Food Web

The structure of the soil decomposer community can be thought of as a pyramid with a small number of top-predators (salamanders) at the pyramid tip, invertebrates (insects) in the center, and a huge biomass of soil microbes (bacteria and fungi) at the base. The flow of energy within an ecosystem can be considered either top-down, where predators influence consumers, or bottom-up, where producers influence consumers (Pace et al. 1999; Shurin et al. 2002).

Fungi represent a critical food source for fungus-consuming invertebrates, while invertebrates mix soil and shred leaf litter, allowing nutrients to be cycled through the microbial food web (Crowther et al. 2013). The red-backed salamander is a terrestrial keystone predator of the decomposer food web that feeds on invertebrates within the forest ecosystem (Burton and Likens 1975).

Given a salamander's capacity to govern the composition of invertebrates via predation, it is likely that this top-down control will alter the structure of fungal communities, and therefore, the functioning (carbon and nutrient cycling) of temperate forest ecosystems.

To test this, we investigated the impact of the removal of the red-backed salamander on the decomposer food web in a mesocosm study simulating an upland deciduous forest in West Virginia, USA.

Data Collection Methodology

A total of 14 salamanders were captured and a block of soil including a cover object and all associated leaf litter was excavated from the exact point of salamander capture and placed into a plastic Tupperware box (mesocosm). All mesocosms were treated in a standardized fashion with respect to abiotic conditions (temperature, moisture, light) throughout the duration of the study. Each mesocosm contained one salamander at the beginning of the experiment.

On day 1 of the experiment, four mesocosms were randomly assigned to the 'absence' treatment, and salamanders were evicted from these mesocosms. On day 42 of the experiment, an additional five salamanders were randomly chosen and evicted from their respective mesocosms ('eviction' treatment). Salamanders were allowed to remain in the last five mesocosms for the duration of the 84-day (12-week) experiment ('presence' treatment). Table 1 summarizes the three experimental treatment groups.

Table 1. *Experimental design.*

Treatment	Label	Size	Description
1	absence	4	Salamanders removed on day 1
2	eviction	5	Salamanders removed on day 42
3	presence	5	Salamanders not removed

Soil cores were taken on a weekly basis. Once at the start of the experiment (week 0), and once at the end of each week for the duration of the 12-week experiment (weeks 1-12). Thus, for each of the 14 mesocosms, we took 13 soil cores, for a total of 182 samples.

DNA was extracted (n=182 total samples) using a fungal-specific polymerase chain reaction (PCR), and sequencing completed on the Illumina MiSeq platform (2 × 250 bp reads). Fungal sequences (14,247,564 total DNA sequences) were analyzed and quality controlled with the software package mothur (Schloss et al. 2009) and clustered into operational taxonomic units (OTUs – e.g. molecular species) using VSEARCH (Rognes et al. 2016). A total of 7,860 fungal OTUs were identified and used in subsequent machine learning analyses.

Machine Learning

Machine learning occurs when a system is able to use experience, *E*, to improve its performance at a task, *T*, as measured by performance metric, *P* (Mitchell 1997).

For the task of predicting discrete classes (the task in which we are interested) a learning system is given experience in the form of example instances, each labeled with the correct corresponding class. From this experience, the learning system constructs a model that is capable of predicting the class for previously unseen instances. This is called *classification*.

Specific classification techniques include decision tree induction (Quinlan 1986), artificial neural networks (Bishop 1995), k-nearest neighbor classifiers (Cover, Hart 1967), RIPPER (Cohen 1995), support vector machines (Cortes, Vapnik 1995), and random forests (Breiman 2001).

Related Work

Analysis of Microbial Communities and the Top-Down Effect

The use of metabarcoding and high-throughput sequencing of microbial communities is a cutting-edge approach to aid in the understanding of complex microbial systems that cannot be studied using traditional microbiological approaches. A typical discovery pipeline consists of collecting samples (*e.g.*, soil cores), extracting DNA, using high-throughput sequencing, and bioinformatics tools to describe and compare microbial community composition and diversity between samples.

The standard approach to data analysis utilizes multivariate statistics to infer patterns between samples. The limitation of multivariate statistics when compared to machine learning techniques is the relatively low predictive power and limited mechanistic understanding of observed trends. In contrast, the capacity of machine learning techniques to learn, predict, and model complex biological systems means that they have the chance to revolutionize our understanding of community and ecosystem ecology.

The top-down effect of the red-backed salamander on invertebrate communities has been extensively investigated and verified in field and lab studies (Burton and Likens 1975; Hairston 1987; Petranka 1998; Davic and Welsh 2004). Crowther et al. (2011, 2013) provide conclusive evidence that these invertebrate communities exert a strong top-down control on fungal communities. As such, we expect that the top-down control by salamanders should cascade to observable effects on the structure of the fungal community.

Similar to this study, Walker et al. (2014) investigated the direct top-down effect of red-backed salamanders on fungal communities in a field (*i.e.*, not mesocosm) experiment and

concluded that salamanders do not have an overall effect on fungal communities but do impact specific groups of fungi. However, it is likely that the precise mechanisms of this top-down control were obscured by the abiotic variability under natural field conditions. As such, a controlled lab-based study may be necessary to isolate and describe the impacts of this top-down salamander control on the composition of fungi in soil.

Machine Learning/Fungal Community and Ecosystem/Food Web Research

Machine learning is a developing bioinformatics tool for biologists and ecologists. Despite the clear advantages for comprehending complex datasets, machine learning has rarely been applied to study terrestrial microbial communities and food web dynamics. Of the relevant machine learning studies that have been conducted, Kampichler et al. (2000) applied neural network and tree-based models to relate the abundance and species composition of Collembola (invertebrates) to habitat characteristics (*e.g.*, carbon and nitrogen content, microbial biomass, respiration). These approaches were compared to the predictive power of simple statistical models (*e.g.*, regression). They found that neural networks and tree-based models outperformed traditional statistical models and that invertebrate communities can be predicted by total carbon content using these techniques (Kampichler et al. 2000).

Leckberg et al (2014) employed machine learning to understand the effects of using differing thresholds to partition fungal samples into OTUs. For example, a threshold of 97% would mean that two microbes must have at least 97% DNA sequence similarity to be considered the same 'molecular' species. There is debate over the appropriate threshold. Their experiments tested the impact that varying the threshold from 90% to 99% had on a classification accuracy of a machine learning algorithm. Results showed that the accuracy was similar regardless of the threshold used to differentiate the samples. The authors claim that these machine learning experiments suggest that microbial community pattern differences are deeply phylogenetically rooted.

Research Questions

To examine the role of machine learning in helping understand relationships in this ecosystem, we posed two research questions:

Research Question 1:
Does the presence or absence of a salamander in the mesocosm environment impact the community of fungal species?

Research Question 2:
If so, what are the patterns of change in the OTU read counts that are associated with the presence or absence of the salamander?

We address these questions using machine learning algorithms for classification. Before discussing the experiments, however, we describe the data and data preparation process.

Data Descriptions and Preparation

The data describes the fungal communities found in 182 soil samples. Recall that there were 14 mesocosms from three different treatments (presence, eviction, and absence) each sampled 13 times over the course of the 12-week study. The data description for each sample consisted of 7,860 integers, each one indicating the number of times the corresponding OTU (molecular species) was detected in the sample.

Each sample was then augmented with a label to support classification. Each sample was labeled as either PR (salamander present) or AB (salamander absent). Table 2 describes the labeling scheme for each treatment, and Table 3 shows the number of samples assigned to each class.

Table 2. *Labeling process by treatment*

	Presence 5 mesocosms 65 samples	***Eviction*** 5 mesocosms 65 samples	***Absence*** 4 mesocosms 52 samples
	Class	**Class**	**Class**
Wk 0	PR	PR	PR
Wk 1	PR	PR	AB
Wk 2	PR	PR	AB
Wk 3	PR	PR	AB
Wk 4	PR	PR	AB
Wk 5	PR	PR	AB
Wk 6	PR	PR	AB
Wk 7	PR	AB	AB
Wk 8	PR	AB	AB
Wk 9	PR	AB	AB
Wk 10	PR	AB	AB
Wk 11	PR	AB	AB
Wk 12	PR	AB	AB

Table 3. *Sample size for each label*

Label	Sample count
PR	104
AB	78

After labeling, a classification file was generated containing the 7,860 OTU features and the categorical label for each of the 182 samples.

Initial Experimental Methodology

Our research questions focus on patterns related to either the presence or absence of the salamander in the environment. 'Present' or 'absent' represents two distinct situations in our

data. Thus, from a machine learning perspective, it lends itself to analysis through classification algorithms.

To that end, we performed nine classification experiments (Table 4) to build and evaluate models that search for patterns among the fungi that can be leveraged to predict whether or not a salamander is present in the mesocosm at the time the fungal sample was collected.

Table 4. *Experiments for research questions*

Exp.	Induction Algorithm	Feature Set
1	Decision tree induction	Full set
2	Artificial neural network	
3	RIPPER	
4	Random Forests	
5	Artificial neural network	Selected by decision tree induction
6	Decision tree induction	Selected by wrapper method with decision tree induction
7	Artificial neural network	
8	RIPPER	
9	Random Forests	

Evaluation on all nine experiments was performed using 10-fold cross validation (Kohavi 1995).

For experiment 5, we used decision tree induction as a feature selector. In this experiment, decision tree induction was first run using the full feature set. The features used in the decision tree were selected as the subset of features to use in learning an artificial neural network.

For experiments 6-9, we used Kohavi and John's Wrapper Feature Selection technique (Kohavi and John 1997). This feature selection technique embeds classification within the feature selection process to explicitly evaluate the utility of a feature subset to support the learning task. The classification algorithm we embedded in the wrapper was decision tree induction.

Research Question 1

To assess our first research question, we examined whether or not a classification algorithm could learn a model that accurately predicts whether or not a salamander is present in the environment. The intuition behind this is that the accuracy of such a model should be directly proportional to the extent of the difference in the fungal community between environments with a salamander and those without.

Research Question 2

To address our second research question, we examined the details of the models themselves rather than just the performance of those models. The intuition behind this approach is that patterns identified and used by the models should provide the scientific researcher with valuable insight regarding which fungi are impacted by the presence or absence of the salamander.

Not all nine experiments in Table 4 are conducive to this analysis. The algorithm must produce a model that the researcher can understand. Of the algorithms explored, only decision tree induction and RIPPER produce human-understandable models. Thus, this research question is only addressed by experiments 1, 3, 6 and 8.

Initial Results

Research Question 1

The results for the nine experiments in Table 4 are listed in Table 5 below.

Table 5. *Classification results*

Exp.	Accuracy	AUROC	F-Measure
1	67.22%	0.68	0.71
2	69.20%	0.73	0.71
3	68.61%	0.69	0.71
4	69.20%	0.85	0.78
5	83.69%	0.92	0.86
6	79.68%	0.80	0.82
7	79.97%	0.88	0.83
8	78.21%	0.79	0.81
9	78.04%	0.84	0.82

Research Question 2

We extracted the rules generated by the decision tree experiments (1 and 6) and the rule induction experiments (3 and 8).

Table 6 illustrates the rules generated during experiment 8 and is indicative of all the extracted rules.

Table 6. *Rules from experiment 8*

Rule	Definition
1	(Otu00028 <= 1) and (Otu00224 <= 0) and (Otu00146 <= 0) and (Otu00149 <= 2) => State=ABSENT
2	(Otu00614 >= 1) => State=ABSENT
3	(Otu00416 >= 2) => State=ABSENT
4	(Otu00972 >= 1) => State=ABSENT
5	(Otu00675 >= 3) => State=ABSENT
6	=> State=PRESENT [DEFAULT RULE]

The rules listed in Table 6 are executed in order, proceeding from rule 1 down to rule 6. The first rule satisfied by the instance is the one used to classify the instance. If none of the first five rules are satisfied by the instance, the default rule (rule 6, in this case) is used.

Observations

While the accuracies in Table 5 appear to support the existence of a relationship between the presence (or absence) of a salamander and patterns in the fungal communities, the discovered rules raised questions about the nature of the patterns.

The rules appear to rely on essentially the presence or absence of particular OTUs (presence/absence of fungal species). This caused us to question whether the patterns are due to the response of the fungal community to the presence or absence of the salamander or are due to random initial differences (noise) in the mesocosm's fungal communities.

Two factors lent weight to this suspicion. First, our data set is small compared to the number of features. This increases the chance of overfitting (fitting noise in the training data). Second, we realized that our use of standard 10-fold cross validation was inappropriate given that multiple samples are drawn from each mesocosm.

The random distribution of samples across the 10-fold cross validation experiment would likely result in samples from the same mesocosm appearing in both the training set and test set. This raises the possibility that the results in Table 5 are from the algorithms leveraging differences in the initial random conditions of the mesocosms to recognize the mesocosms rather that finding differences relating to the presence (or absence) of a salamander.

We examined this new hypothesis through a series of experiences described in the next section.

Revised Methodology and Results

To ensure that idiosyncratic mesocosm patterns in the training sets do not give the learned model an unfair advantage, we must make sure no samples from the mesocosm(s) in the test data are present in the training data.

To accomplish this, we designed a 14-fold cross validation approach (14-fold cv) in which each fold consists of the 13 soil samples from one of the 14 mesocosms. This isolates mesocosm-specific patterns in the test set from those in the training set. This approach comprised a set of 14 tests. Each one using all 169 samples from 13 of the mesocosms to learn a classification model that is evaluated using the 13 samples from the remaining mesocosm. For this series of experiments, we used decision tree induction.

Table 7 includes the both the aggregate error rate on the training sets as well as the aggregate error rate on the test sets for our three 14-fold cv experiments. The first of these experiments was run using the full feature set and appears to confirm our suspicions. The low training error coupled with the high test error indicate that patterns were found that matched the training set but did not generalize to the test set – suggesting that the patterns fit the noise in the training set.

Table 7. *Training and test errors from first 14-fold cv experiment*

Exp. #	Description	Training error	Test error
1	Full set of instances; full set of features	4.14%	46.7%
2	Full set of instances; reduced set of features	9.7%	39%
3	"Balanced" set of instance; reduced set of features	8%	42.9%

In an effort to reduce the presence of OTUs that are unique to specific mesocosms, we ran experiments using only the OTUs that are present in the majority of the samples. Table 7 shows improved performance with this reduced feature set, but a deeper look at the results (Table 8 – row 1 'Imbalanced data') suggest a bias toward predicting "present."

Attempting to test this, our third experiment under-sampled the "present" instances in each training set to create an equal balance between the two classes. The overall results are slightly worse, but the treatments are more balanced (Table 8 – row 2 'Balanced data').

Table 8. *Test error by treatment*

Description	Presence treatment	Absence treatment	Eviction treatment
Imbalanced data	30.8%	46.2%	41.5%
Balanced data	46.2%	46.2%	36.9%

Conclusion

We proposed, described, and evaluated a methodology for using machine learning to assist in the analysis of a complex biological domain that has implications for the global carbon cycle and climate change. This is a novel domain for machine learning, and one that we believe can benefit from the powerful predictive modeling tools it enables. Although we did not achieve the accuracies that might be expected in other fields, the strength of our predictions were stronger than many commonly observed in such complex ecological systems using traditional statistical tools (Crowther et al. 2013). More importantly, we have gained considerable insights into how machine learning can be used to help with this important research question.

The hyper-diverse and complex nature of natural ecological communities calls for the use of machine learning approaches to generate meaningful predictions about the functioning of natural ecosystems under current and future global change scenarios. Given the importance of controlled, laboratory-based mesocosm experiments for disentangling the complex relationships in ecosystems, it is important that we develop a machine learning methodology that can be applied to such experiments. We have shown that traditional 10-fold cross validation does not work appropriately for our mesocosm-based data, and we have proposed, evaluated, and demonstrated the appropriateness of our mesocosm-based N-fold cross validation method.

Limitations

The small number of samples in the data is one limitation of our study. Unfortunately, the cost and labor-intensive nature of the data collection process makes a large sample size nearly impossible.

Another limitation is that the current research ignores the biological taxonomy that groups species into higher order-ranks (*i.e.*, genus, family, order, class, phylum, kingdom). This is a limitation because the actual top-down effect might be more clearly seen when the data is aggregated at higher levels in the taxonomy. The reported research only looked for patterns at the OTU (species) level.

Future Work

The testing method we established will serve as the basis for future machine learning explorations with this and other similar data. Additional future work includes similar machine learning analyses in other related ecosystem studies as well as following up on the biological implications of patterns our machine learning work uncovers. We will also continue to explore various techniques in our effort to determine the best methodology for analyzing this domain.

References

Bardgett, R.D. 2005. *The Biology of Soil*. Oxford: Oxford University Press.

Bishop, C. M. 1995. *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Breiman, L. 2001. Random forests. *Mach. Learn.*, 45(1): 5–32.

Burton, T. M.; and Likens, G. E. 1975. Salamander populations and biomass in the Hubbard Brook Experimental Forest, New Hampshire. *Copeia,* 3: 541–546.

Cohen, W. W. 1995. Fast effective rule induction. In Proceedings of the Twelfth International Conference on Machine Learning, 115–123.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Mach. Learn.*, 20(3): 273–297.

Cover, T.,; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE T. Inform. Theory.,* 13(1): 21–27.

Crowther, T.; Boddy, W.L.; Jones, H. 2011. Outcomes of fungal interactions are determined by invertebrate grazers. *Ecol. Lett.,* 14: 1134–1142.

Crowther, T. W.; Stanton, D. W. G.; Thomas, S. M.; et al. 2013. Top-down control of soil fungal community composition by a globally distributed keystone consumer. *Ecology,* 94(11): 2518–2528.

Davic, R. D; and Welsh, H.H. Jr. 2004. On the ecological role of salamanders. *Annu. Rev. Ecol. Evol. Syst*. 35: 405–434.

Hairston, N. G. Sr. 1987. *Community ecology and salamander-guilds*. Cambridge Univ. Press, New York, NY.

Kampichler, C., Džeroski, S. and Wieland, R., 2000. Application of machine learning techniques to the analysis of soil ecological data bases: relationships between habitat features and Collembolan community characteristics. *Soil Biol. Biochem.*, 32(2): 197–209.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI,* 14(2): 1137–1145.

Kohavi, R.; and John, G. H. 1997. Wrappers for feature subset selection. *Artif. Intell.*, 97(1): 273–324.

Lekberg, Y., Gibbons, S. M. and Rosendahl, S., 2014. Will different OUT delineation methods change interpretation of arbuscular mycorrhizal fungal community patterns?. *New Phytol*, 202: 1101-1104

Mitchell, T. M. 1997. *Machine learning*. New York: McGraw Hill.

Pace, M. L.; Cole, J. J.; Carpenter, S. R.; et al. 1999. Trophic cascades revealed in diverse ecosystems. *Trends Ecol. Evol.,* 14(12): 483–488.

Petranka, J. W. 1998. *Salamanders of the United States and Canada*. Smithsonian Institution Press, Washington, DC.

Quinlan, J. R. 1986. Induction of decision trees. *Mach. Learn*. 1(1): 81–106.

Rognes T.; Flouri T.; Nichols B.; et al. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*., 4: e2584.

Schloss P. D.; Westcott S. L.; Ryabin T.; et al. 2009. Introducing mothur: open-source, platform-independent, community- supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.,* 75(23): 7537–7541.

Shurin, J. B.; Borer, E. T.; Seabloom, E. W.; et al. 2002. A cross-ecosystem comparison of the strength of trophic cascades. *Ecol. Lett.,* 5(6): 785–791.

Van der Heijden, M. G. A.; Bardgett, R.; and Van Straalen, N.M. 2008. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol. Lett.,* 11(3): 296–310.

Walker, D.M.; Lawrence, B.R.; Esterline, D.; et al. 2014. A metagenomics- based approach to the top- down effect on the detritivore food web: a salamander's influence on fungal communities within a deciduous forest. *Ecol. Ecol.*, 4(21): 4106–4116.

Wardle, D. A. 2006. The influence of biotic interactions on soil biodiversity. *Ecol. Lett.* 9(7): 870–886.