Machine Learning and Ecosystem Informatics: Challenges and Opportunities

Thomas G. Dietterich

Oregon State University, Corvallis OR 97331, USA
tgd@cs.orst.edu
http://web.engr.oregonstate.edu/~tgd

Abstract. Ecosystem Informatics is the study of computational methods for advancing the ecosystem sciences and environmental policy. This talk will discuss the ways in which machine learning—in combination with novel sensors—can help transform the ecosystem sciences from small-scale hypothesis-driven science to global-scale data-driven science. Example challenge problems include optimal sensor placement, modeling errors and biases in data collection, automated recognition of species from acoustic and image data, automated data cleaning, fitting models to data (species distribution models and dynamical system models), and robust optimization of environmental policies. The talk will also discuss the recent development of The Evidence Tree Methodology for complex machine learning applications.

1 Introduction

There are many different ways of conducting scientific research. At one extreme—which we might call "science-in-the-small"—individual scientists formulate hypotheses, perform experiments, gather data, and analyze that data to test and refine their hypotheses. This approach provides profound scientific understanding, but it tends to yield slow progress, because each individual scientist can only study a small collection of phenomena limited in time and space. At the other extreme—which we might call "science-in-the-large"—automated instruments collect massive amounts of observational data, which are then analyzed via machine learning and data mining algorithms to formulate and refine hypotheses. This approach to science can lead to rapid progress, but because it is driven by data rather than hypotheses, it tends not to result in deep causal understanding. Progress can be both fast and deep if we can combine these two approaches to scientific research.

Until the early 1990s, molecular biology was conducted exclusively as science-in-the-small. But the development of automated sequencing methods and algorithms for proteins, DNA sequences, and ultimately whole genomes has permitted the rapid development of science-in-the-large. We are now witnessing a strong and healthy interaction between these two forms of research that is producing rapid progress.

Z.-H. Zhou and T. Washio (Eds.): ACML 2009, LNAI 5828, pp. 1–5, 2009.

Our planet is facing numerous challenges, including global climate change, species extinctions, disease epidemics, and invasive species, that require the development of robust, effective policies based on sound scientific understanding. However, in most cases, the scientific understanding is lacking, because the ecosystem sciences are still in their infancy. Ecology is still dominated by science-in-the-small. Individual investigators collect observational data in the field or conduct controlled experiments in order to refine and test hypotheses. But there is relatively little science-in-the-large. The goal of our research at Oregon State University is to develop novel computer science methods to help promote science-in-the-large in the ecosystem sciences to address critical policy questions.

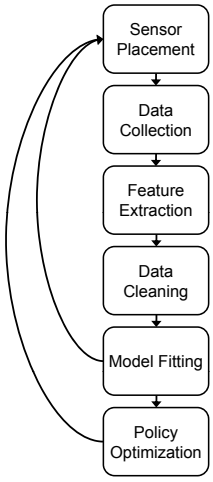

Fig. 1. The Ecosystem Informatics Pipeline

Figure 1 shows a conceptual model of the information processing pipeline for ecological science-in-the-large. Each box corresponds to a computational problem that requires novel computer science (and often machine learning) research to solve. Let us consider each problem in turn.

- **Sensor Placement.** Many novel sensors are being developed including wireless sensor nodes and fiber-optic-based distributed sensors. The first decision that must be made is where to place these sensors in order to most effectively collect data. This problem can be formulated statically, or it can be considered dynamically, as a form of active learning in which the sensors are moved around (or new sensors are added) based on information collected from previous sensors. Among some of the objective functions that must be considered are (a) maximizing the probability of detecting the phenomenon (e.g., detecting a rare or endangered species), (b) improving model accuracy

(as in active learning), (c) improving causal understanding, and (d) improving policy effectiveness and robustness (which is related to exploration in reinforcement learning).

- **Data Collection.** The process of data collection can introduce biases and errors into the data. For example, the Cornell Lab of Ornithology runs a large citizen-science project called ebird (`www.ebird.org`) in which amateur bird watchers can fill out bird watching checklists and upload them to a web site. In this case, humans are the "sensors", and they may introduce several kinds of noise. First, they introduce sampling bias because they tend to go bird watching at locations near their homes. Second, even if a bird is present at a location, they may not detect it because it is not singing and it is hidden in dense foliage. Third, the humans may misidentify the species and report the wrong species to ebird.org. Machine learning algorithms need to have ways of dealing with these problems.
- **Feature Extraction.** It is almost always necessary to transform the raw data to extract higher-level information. For example, image data collected from cameras must be processed to recognize animals and classify them according to species. In some applications, such as counting the number of endangered wolves or bears, it is important to recognize individual animals so that they are not counted multiple times. A related problem is to track individuals as they are detected by multiple instruments over time. At Oregon State, we have been working on identifying the species of arthropods.
- **Data Cleaning.** As large numbers of inexpensive sensors are placed in the environment, the quantity of data increases greatly, but the quality of that data decreases due to sensor failures, networking failures, and other sources of error (e.g., recognition failures in image or acoustic data). This gives rise to the need for automatic methods for data cleaning. This is an important area for machine learning research.
- **Model Fitting.** The computational task of fitting models to data is a core problem in machine learning with many existing algorithms available. However, ecological problems pose several novel challenges. One problem is the simultaneous prediction of the spatio-temporal distribution of thousands of species. A simplified view of species distribution modeling is that it is simple supervised learning, where the goal is to learn $P(y|x)$, where x is a description of a site (elevation, rainfall, temperature, soil, etc.) and y is a boolean variable indicating whether a particular species is present or absent there. However, we are often interested in the presence/absence of thousands of species, and these species are not independent. Species can be positively correlated (e.g., because they have similar environmental requirements or because one of them eats the other) or negatively correlated (e.g., because they compete for the same limited resources). While each species could be treated as a separate boolean classification problem, it should be possible to exploit these correlations to make more accurate predictions. This is a form of very-large-scale multi-task learning. Existing methods are unlikely to scale to handle thousands of species.

Another machine learning challenge is to predict the behavior of species. For example, many bird species are migratory. Ecologists need models that can predict when the birds will migrate (north or south), what paths they will take, and where they will stop. It may be possible to formulate this problem as a form of structured prediction problem.

A third novel machine learning challenge is to fit dynamical systems models to populations of single or multiple species. This involves fitting nonlinear differential equations to observations, which is a problem that has received very little attention in the machine learning community. Such models can exhibit exponential increases and decreases as well as chaotic behavior, so this presents formidable statistical challenges.

Figure 1 shows an arrow from model fitting back to sensor placement. This is intended to capture the opportunity to apply active learning methods to improve sensor placement based on fitted models.

– **Policy Optimization.** In many cases, the models that are fit to data become the inputs to subsequent optimization steps. Consider, for example, the problem of designing ecological reserves to protect the habitat of endangered species. The goal is to spend limited funds to purchase land that will ensure the survival (and even the recovery) of endangered species. An important challenge here is to develop solutions that are robust to the errors that may exist in fitted models. Can we develop ways of coupling optimization with model fitting so that the solutions are robust to the uncertainties introduced throughout the data pipeline?

  The arrow leading from policy optimization back to sensor placement suggests one possible solution to this challenge—position more sensors to collect more data to reduce the uncertainties in the fitted models.

2 Summary Remarks

Machine Learning has the potential to transform the ecosystem sciences by enabling science-in-the-large. Although many problems in ecology are superficially similar to previously-studied problems (e.g., active learning, density estimation, model fitting, optimization), existing methods are not directly applicable or do not completely solve the ecological problems.

The keynote talk will describe three instances of this. First, standard methods for generic object recognition do not provide sufficient accuracy for recognizing arthropods. The talk will describe two novel machine learning methods that can solve this problem and that also work well for generic object recognition tasks. Second, standard methods for multi-task learning do not suffice to jointly predict thousands of species. The talk will provide evidence that joint prediction is important, but no known method can currently solve this problem. Third, the talk will describe a spatio-temporal Markov decision problem for managing forests to reduce catastrophic forest fires. In principle, this can be solved by existing dynamic programming algorithms. But in practice, the size of the state and action spaces makes existing algorithms completely infeasible.

I urge everyone to work on these interesting research problems. Given the ecological challenges facing our planet, there is an urgent need to develop the underlying science that can guide policy making and implementation. Ecology is poised to make rapid advances through science-in-the-large. But ecologists can't do this alone. They need computer scientists to accept the challenge and develop the novel computational tools that can make these advances a reality.