Research Statement
Li-Ping Liu
EECS, Oregon State University

1 General Introduction

My strong interest in artificial intelligence has led me to my graduate study of machine learning. I am excited to exploit the power of data and solve problems that cannot be solved by manual analysis only. My current research focuses on machine learning in computational sustainability. Working with real applications ensures that my research has practical value. However, it does require me to dig deep enough to identify new machine learning problems and discover new solutions, so my research also advances the field of machine learning. My strong skills in math and programming have helped me to become a good researcher. During my graduate study, I have published ten scientific papers, most of which are in top conferences in machine learning and data mining.

1.1 Masters Degree

During my graduate study in the LAMDA group in Nanjing University, my research focused on accelerating the feature extraction step in classification. My first project used a cascading feature extraction process to avoid extracting all features during label prediction. The work showed that only a small fraction of features need to be extracted to achieve good prediction accuracy. The second effort sought to accelerate an online version of Linear Discriminant Analysis. Both works were published in ICDM (ICDM, 2008)(ICDM, 2009). My master’s thesis won the “Jiangsu Provincial Outstanding Masters Thesis Award”.

1.2 PhD Study

I came to Oregon State University in 2010 and joined Prof. Thomas Dietterich’s research group. My research overall focuses on machine learning problems in computational sustainability, and it investigates three important problems in the field: detecting bird species by classifying their songs, modeling migrating species, and designing bioreserves for endangered species.

In the first problem of bird song classification, I formulated the problem as an instance of the superset label learning problem. This problem is similar to the multiclass classification problem. The only difference is that instead of being told the true label on each training instance, we are given a set of labels, one of which is the correct label. For this problem, I proposed a new probabilistic learning model (NIPS, 2012) that improves the classification performance over previous models. I coded this model into a well-designed package, which was used by many other researchers. To theoretically analyze my method, I then proved a PAC-learnability result for superset label learning problem for the realizable case (ICML, 2014).

In the bird migration project, our objective is to recover the transition model, assumed to be a Markov chain model, of migrating birds. Each state of the Markov chain specifies the position of a single bird on the map. In the migration process, we can observe noisy counts of birds at different states, and we need to infer birds’ transition counts—sufficient
statistics of model—to recover the model parameter. The data is modeled with a Collective Graphical Model (CGM), which is developed by Dan Sheldon. The CGM model is defined at the population level instead of at the level of individual birds, and exact inference with this model is NP-hard. To circumvent this problem, I develop an approximate inference method for the CGM model. The CGM distribution of bird transition counts is approximated with a Gaussian distribution, and then the exact inference can be performed with the Gaussian distribution (ICML, 2014). Theoretically, my approximation is guaranteed to be correct by the Central Limit Theorem. Moreover, my inference method improves inference accuracy and speed over previous inference methods for CGM.

Currently, I am working on the problem of reserve design, in which we need to decide which land parcels to purchase to protect a species of interest. The difficulty is that, instead of knowing presence/absence of the species on all land parcels, we only have a small training set of land parcels that are labeled with presence/absence. Traditional methods decompose this type of problem into two separate steps: predicting the distribution of the species with supervised learning models, and optimizing the selection against the prediction. This strategy discards the uncertainty in the predictive model and leads to suboptimal reserve designs. I propose to integrate learning and decision making in a combined optimization problem, which is a Mixed-Integer Programming problem. I develop a convex relaxation of the problem and solve it. Initial results show that the solution obtained from this direct approach is better than those from traditional methods in both theory and practice.

I have also participated several other research projects: fitting non-parametric occupancy-detection models (AAAI, 2011), active learning from outlier data (SDM, 2012), active learning from bird acoustic data (ECML, 2015), and time series analysis of sales data (INFORMS Workshop, 2015).

2 Research Proposal

In the future, I plan to continue my research in machine learning and computational sustainability. There are still many important research directions to be explored with the data related to birds. First, I would like to continue my work on probabilistic modeling of bird migration and integrate more evidences to my model. Besides bird checklists from birders, radar data is also an evidence that shows the density and direction of flight of birds during major migration nights. I plan to extend the current model to include the radar data in hopes of improving the model accuracy in both inference and prediction. Second, I intend to use active learning techniques to suggest places for birders so that we can collect more informative bird reports. The current bird data is very sparse in some areas for various reasons, such as low population density and lack of accesses. The learning model will get much more supervision if we can collect more data in such areas. The Cornell Lab of Ornithology has developed various incentive mechanisms (games, competitions) to encourage birds to visit such areas. My method will suggest the best way to deploy those mechanisms. To better exploit birder capacity, I plan to consider both the data needs and birders’ interests.

I am also interested in the broad problem of learning from sensor data. In many applications, data is collected from multiple sensors of the same or different types, and we need to jointly model the data to get the most knowledge from it. At the same time, sensors often produce a large quantity of data, which poses a speed requirement for the learning system. The two problems exist in many applications. For example, environmental data can be collected by different sensors such as temperature sensors and sound recorders, and the data heterogeneity presents a challenge to joint modeling. In the application of brain-computer
interaction, sensors can generate massive data in high speed, and it often requires real-time processing of the data. I plan to approach these two problems in the following directions. First, I propose to use deep learning techniques to extract human-understandable features from raw sensor data. Second, I would like to try ensemble of models with each model learning from the data collected from one sensor. I plan to explore the two research directions with ecological data and sensor data from healthcare systems.

I would also like to study probabilistic graphical models in the background of big data. While probabilistic models are powerful in modeling various data, they encounter many difficulties when applied to big data. Models with big data need to be complex to capture enough factors, and at the same time, they are required to run fast. One research direction is to accelerate model learning and inference with parallel computing. Parallelization not only is an engineering problem but also poses a question for the machine learning research: how to parallelize learning and inference at the fundamental level? I am willing to study this topic with my experience in probabilistic graphical models and MapReduce programming.