Synergistic Machine Learning: Collaboration and Topology Exploitation in Dynamic Environments

NSF Robust Intelligence Project Report for Year 3, June 2009 to May 2010

People

Current Contributors:

- Terran Lane (PI), Univ. of New Mexico (UNM) Associate Professor of Computer Science
- Kiri Wagstaff (co-PI), JPL Researcher in Computer Science
- Rick Aster, New Mexico Institute of Mining and Technology (NMT) Professor of Geophysics
- Ashley Davies, JPL Researcher in Volcanology
- Hunter Knox, NMT Ph.D. student in Geophysics
- Ben Yackley, UNM Ph.D. student in Computer Science

Collaborators and Past Contributors:

- Blake Anderson, UNM Ph.D. student in Computer Science
- Julie Coonrod, UNM Professor of Civil Engineering
- Eduardo Corona, UNM student (now Ph.D. student at NYU)
- Marie desJardins, Univ. of Maryland, Baltimore County (UMBC) Associate Professor of Computer Science
- Eric Eaton, UMBC Ph.D. student in Computer Science (now at Lockheed Martin)
- Jillian Green, California State Univ. Los Angeles undergraduate student in Computer Science
- Phil Kyle, NMT Professor of Geochemistry
- Joshua Neil, UNM Ph.D. student in Statistics / Los Alamos National Lab
- Umaa Rebbapragada, Tufts Univ. Ph.D. student in Computer Science
- Alex Roper, California Institute of Technology undergraduate student in Computer Science (now Ph.D. student at Univ. of Michigan)
- Sushmita Roy, UNM Ph.D. student in Computer Science (now at the Broad Institute of MIT and Harvard)
- Curtis Storlie, UNM Professor of Statistics
- Maggie Werner-Washburne, UNM Professor of Biology

Activities

In this year, our project advanced on both core directions of the project: topological learning and collaborative learning.

1. Exploiting Network Topology

This year, we extended our work on rapid scoring for network topology identification. Previously, we showed that our method for approximating the scores of network structures provided fast and accurate estimation of posterior graph likelihoods. We had also started on structure search. This year, we fleshed out our work on graph structure identification. We developed a Gaussian process regressor for graph structures that exploits the topology of the "meta-graph" space (the graph over the space of all possible graphs). We used the resulting approximator in a structure search mechanism, showing 1-2 orders of magnitude speedup over searches that do exact scoring. We showed that our search also often finds results that are better than the exact score-based search. We showed that this counterintuitive effect arises from a smoothness imposed on the space by the GP regressor, as a function of our choice of meta-graph topology. This work is currently under review at UAI, and we are preparing a JMLR paper on it. Terran gave talks on this work at the Universidad Complutense de Madrid (UCM), Universidad Autonama de Madrid (UAM), Universidad Polytechnica de Madrid (UPM), the Functional Imaging Lab at University College London (UCL), and Los Alamos National Lab (LANL).

We also extended this work in two directions. On one front, we are developing topological learning methods for other classes of combinatorial optimization problems. Specifically, we have developed algorithms for partial factor graph structure identification and for permutation modeling. Like our previous work, we choose a projection from our combinatorial sets (random variable groupings or permutations) into a topological space. In this case, we found embeddings of both sets onto the surface of hyperspheres in Euclidean space. We can then manipulate that space using continuous tools, such as continuous probability distribution functions or continuous basis functions. For example, we can represent a probability distribution over the set of permutations using a von Mises-Fisher distribution on the hypersphere. Or we can model an arbitrary (smooth) function on this set using spherical harmonic functions (Bessel functions). We have worked out and implemented the algorithms for these, and empirical investigation is currently under way.

On another front, we are developing a variant of multi-task learning for Bayesian network structure search. We are looking at learning condition-dependent graphs. That is, there are a set of k condition variables, and every combination of conditions (2^k) is associated with a graph, such as a Bayesian network. But the graphs are not completely independent; related conditions induce structural dependencies among the BNs (such as shared graph structure). This scenario arises, for example, in sensor networks, where conditions could be different categories of events to be detected. It is also relevant in medical modeling and diagnosis, where conditions are things like different diseases, age, gender, race, drug treatments, and so on. Using our 'meta-graph' strategy, we developed a technique for simultaneously identifying multiple, related, condition-dependent graphs. We have showed that our method finds better Bayesian networks than either searching for all graphs independently or than a standard multi-task learning algorithm that does not use the meta-graph structure. These results are currently under review at UAI.

This work also led to a number of spin-off papers concerned with graph structure identification for genomic data.

2. Collaborative Learning

This year, we tackled the other major component of collaborative learning, which was to extend collaborative systems to also work with clustering algorithms. Our prior work focused on how classifiers could exchange information to improve their performance. In the summer of 2009, Jillian Green implemented a collaborative clustering system during her internship at JPL. To move from classifiers to clustering, we formulated a different way for learners to share information. While classifiers can naturally generate and learn from individual data labels, clustering systems cannot. Instead, we enabled the clusterers to generate and learn from pair-wise information, in the form of must-link and cannot-link constraints. These constraints have been used extensively over the past decade to inform semi-supervised clustering methods. However, the constraints almost always come from an oracle. When using constraints to share information collaboratively, they are much less reliable, and we observed interesting behavior on a variety of data sets. We also investigated methods for "cleaning" or otherwise improving the reliability of the learned constraints, by pruning inconsistent or contradictory constraints. We are currently in the process of writing up this work for publication. Kiri gave talks/colloquia on this work at the University of Utah (November 2009), Pomona College (April 2010), and the University of New Mexico (April 2010).

The other main advance in our experimental work was that we moved from applying collaborative learning methods only to UCI data sets to applying them to seismic and acoustic data collected by the Mount Erebus Volcano Observatory. This data is publicly available (http://www.iris.edu/data/), but we benefited greatly from the assistance of our colleagues at the New Mexico Institute of Mining and Technology (NMT), who are directly involved in the collection and formatting of the data. They helped us interpret and convert the data files, and they provided labels for ~40 observed events to separate them into "eruption" and "icequake" categories. Recently, they provided a much larger list of 20,000 (unlabeled) events from all of 2007 and 2008, which we are eager to analyze.

To aid in evaluating our collaborative learning algorithms, Umaa Rebbapragada (Tufts University) assembled some artificially linked multi-view data sets from standard UCI data. She was able to carefully control the difficult of each such assembled data set, which is useful in determining which strategies are more or less robust to different problems.

3. Integrating orbital and ground observations

Ashley Davies (JPL) has obtained orbital infrared observations of Mt. Erebus using the Hyperion instrument on the EO-1 spacecraft. He has been investigating ways to compare these observations (thermal␣ output) with the ground-based observations that the Mt. Erebus Volcano Observatory provides (seismic and acoustic). He has identified the final piece of information needed to do a reliable comparison, which is a reliable estimate of the size of the volcano's lava lake, which likely changes over time. We are investigating whether periodic observations from a video camera stationed just above the lava lake can aid in that size estimate. Ultimately, we aim to combine Hyperion, MODIS, and ASTER data (all orbital) with the ground-based data. If there is a strong correlation, then that derived relationship could help extrapolate to estimates of seismic activity on Io, which is also volcanically active but for which we have only orbital observations.

4. Coordination with NMT

We held a second project meeting at the New Mexico Institute of Mining and Technology (NMT) on November 23, 2009. We exchanged updates about progress in developing new machine learning and data analysis methods and updates from NMT about the deployed sensor network at Mt. Erebus and their own analysis of the data. Their research goals center on understanding activity inside the volcano, and we seek for our analysis methods to find new ways to shed light on that subject. Two promising avenues are 1) automatically separating events into eruption and non-eruption classes, and 2) clustering events into smaller groups of similar event types, which may aid in the discovering of new kinds of activity.

Findings

For this project year, our findings are:

1. Topological Learning

- Fast structure search for Bayesian networks that achieves 1-2 orders of magnitude improvement over traditional search methods.
- Determined that the high-quality results of the topological structure search comes from imposing smoothness on the space of possible graph structures.
- New learning algorithms for combinatorial structures (random variable clusters and permutations) that work by embedding the combinatorial sets into continuous topologies (hyperspheres) and developing probability distributions on those topologies.
- A multi-task learning framework for condition-specific Bayesian networks.
- Spin-off results for rapid graph identification for genomic and␣ biological networks.

2. Collaborative Learning

- Collaborative classification improved performance in classifying eruptions and icequakes in Mt. Erebus data from 70% to 80%.
- Collaborative clustering improved clustering performance on Mt. Erebus data by 150%.
- To date, “cleaning” constraint sets to remove contradictory constraints has not had any impact on performance. We are still investigating this.
- To properly compare the orbital and ground observations of Mt. Erebus activity, we need a good estimate of the size of the lava lake.

Publications

Wagstaff, K.L., Kocurek, M., Mazzoni, D., and Tang, B. (2010), "Progressive Refinement for Support Vector Machines", *Data Mining and Knowledge Discovery*, 20(1):53-69. DOI: 10.1007/s10618-009-0149-y.

Roy, S., Martinez, D., Platero, H., Lane, T., and Werner-Washburne, M. (2009), "Exploiting Amino Acid Composition for Predicting Protein-Protein Interactions", *PLoS ONE*, 4(11):e7813. DOI: 10.1371/journal.pone.0007813.

Roy, S., Plis, S., Werner-Washburne, M., and Lane, T. (2009), "Scalable learning of large networks", IET Systems Biology, 3(5):404-413. DOI: 10.1049/iet-syb.2008.0161.

Roy, S., Lane, T., and Werner-Washburne, M. (2009), "Learning structurally consistent undirected probabilistic graphical models", In *Proc. of the Twenty-Sixth International Conference on Machine Learning*, p. 905-912.

Training and Development

We are training two students at the UNM site:

Blake Anderson is a Ph.D. student in the Department of Computer Science. He is focusing on exploiting the topology of the combinatorial graph structure space to perform fast model identification. His work has led to a conference paper currently under review at the European Conference on Machine Learning (ECML). He is also working with our New Mexico Tech collaborators on the sensor network data from the MEVO observatory, attempting to exploit sensor network topology to learn high-quality discriminative models of seismic events.

Ben Yackley is a Ph.D. student in the Department of Computer Science. His work focuses on efficient estimation of graph structure objective functions by building Laplacian-based extimators on graph structure space. His early work in that area led to a NIPS publication in 2008. He is currently working on methods for efficient Bayesian model averaging in combinatorial graph spaces. Extensions of this work will form the core of his dissertation. In addition, Ben is developing his own teaching skills toward a career in a teaching-oriented university position.

Umaa Rebbapragada is a Ph.D. student at Tufts University in Computer Science. She has continued to develop and evaluate the collaborative machine learning methods that came out of her 2008 summer internship at JPL. Co-PI Kiri Wagstaff joined Umaa's dissertation committee, and Umaa plans to defend on June 14, 2010. Further, she has accepted a permanent position as a researcher at JPL, to start in the fall of 2010. Her work on this project continues to provide interesting new directions, and she plans to officially re-join the project when she arrives at JPL.

We have also had two undergraduate students involved in this project at JPL. Alex Roper (summer 2007) contributed to our initial work with structured, remote sensing data, with the goal of predicting crop yield from orbital data. He graduated from the California

Institute of Technology in June of 2009 and, after taking a year off, has now been accepted to the University of Michigan as a Ph.D. student and will start there this fall.

Jillian Green completed a summer internship with this project in 2009. She implemented the collaborative clustering system and was involved in designing experiments, evaluating performance, and writing up the results for publication. She graduated from California State University, Los Angeles, in late 2009 with her B.S. in Computer Science.