Knowledge-Enhanced Discovery System (KEDS): Incorporating Background Knowledge for Scientific Discovery

NSF ITR Project Report covering September 2009 to August 2010

Project website: http://maple.cs.umbc.edu/keds/

People

Current Contributors:

- Marie desJardins (PI), University of Maryland, Baltimore County (UMBC) Associate Professor of Computer Science
- Kiri Wagstaff (co-PI), JPL Researcher in Computer Science
- Jennifer Adamshick, UMBC undergraduate student in Computer Science
- Peter Hamilton, UMBC Master's student in Computer Science
- James MacGlashan, UMBC Ph.D. student in Computer Science
- Lan Mei, high school student
- Eisha Nathan, high school student

Collaborators and Past Contributors:

- Sugato Basu, Google Research Senior Research Scientist
- Ian Davidson, University of California, Davis, Associate Professor of Computer Science
- Stuart Schwartz, UMBC Center for Urban Environmental Research and Education (CUERE) Senior Scientist
- Craig Cambias, UMBC undergraduate student in Computer Science (grad. 2005)
- Ryan Carr, UMBC undergraduate student in Computer Science (grad. 2007)
- Eric Eaton, UMBC Ph.D. student in Computer Science (grad. 2009)
- Gregory Handy, UMBC undergraduate student in Computer Science (grad. 2010)
- Natalie Podrazik, UMBC undergraduate student in Computer Science (grad. 2006)
- JC Montminy, UMBC Master's student in Computer Science (grad. 2008)
- Jake Thompson, UMBC undergraduate student in Computer Science (grad. 2008)
- Brandon Wilson, UMBC Master's student in Computer Science (grad. 2008)
- Kevin Winner, UMBC undergraduate student in Computer Science (grad. 2010)
- Qianjun Xu, UMBC Ph.D. student in Computer Science (grad. 2006)

Activities

Our research activities this year focused on 4 major areas: (1) preference learning and modeling for subset selection; (2) active feature acquisition techniques; (2) developing methods for handling non-random missing data in large real-world data sets; (3) applications of machine learning to real-world application domains; and (4) integrating HTN planning with reinforcement learning.

1. Subset selection: We published a journal paper summarizing our methods for modeling, learning, and applying user preferences over sets of objects. The paper appeared in the Journal of Experimental and Theoretical Artificial Intelligence (JETAI) in November, 2009. We applied these methods to music collections and images collected by Mars rover field tests.

2. Active feature acquisition: We published a conference paper (at SDM-10) on Confidence-based Feature Acquisition (CFA), which acquires missing feature values in an inductive learning setting. In this problem, costs are associated with obtaining individual feature vales for each data instance in the training and test data. A common scenario is that of medical diagnosis, in which the feature values are the results of different tests that can be performed to gather information about a patient's condition. Some tests, such as obtaining weight, height, or blood pressure, have minimal or zero associated cost (and are typically performed for all patients), whereas other tests that require lab work, centrifuges, or expensive reagents may involve significant cost. Taking into account a user's desired confidence level for prediction accuracy, the CFA method builds a model that minimizes the cost of training and using the classifier.

We are currently extending this work to use intelligent feature selection heuristics, and are studying the effect of different cost distributions on CFA's performance. We are also looking at applying CFA to new data sets, such as observations collected by a radio telescope. In this domain, while the raw data streams in at a fixed observational cost, there are several derived features that provide far more predictive power but which require significant computational time to compute. Keeping up with the data flow in a real-time system constrains how many of these features can be computed, and a technique such as CFA is needed to make good decisions about how to allocate the available computational time so as to achieve the best performance.

3. Handling non-random missing data: We have developed an ensemble-based machine learning technique that can learn from data sets with significant amounts of missing data, where the data is not missing at random. (That is, the attributes that are missing are correlated with each other, and potentially with the values of other attributes in an instance.) During Summer 2010, we will be implementing this approach with the goal of publishing the work this fall or winter.

4. Applying machine learning to real-world application domains: We are using machine techniques to make predictions in environmental science (the effect of land use and land cover on biological activity in a complex ecosystem; an Earth/space science application (the use of satellite observations of cropland to predict the end-of-season crop yield); and a medical domain (multiple sclerosis). In these projects, we have been exploring the use of the techniques for integrating background knowledge that we have developed over the course of the award, in the context of probabilistic modeling of complex physical systems.

5. Integrating HTN planning with reinforcement learning: James MacGlashan, a graduate student funded on the project, is exploring this area for his dissertation research.

His objective is to create an agent that learns hierarchical sills that can solve progressively more complex problem types. The learned skills will also be incorporated into the agent's planning knowledge so that it can plan at a higher level of abstraction. James presented this work at the ICAPS Doctoral Consortium in September 2009. He passed his preliminary exam and advanced to candidacy in March 2010, and will present his work at the AAAI Doctoral Consortium this year.

Contributions

The focus of our research has been the incorporation of background knowledge into various types of learning/discovery techniques, including preference learning, clustering, and classification. The ability to incorporate different kinds of background knowledge into scientific discovery techniques will allow domain experts to interact more effectively with discovery systems.

The key contributions to the field for this year include (1) Confidence-Based Feature Acquisition (CFA), the first technique for cost-sensitive feature acquisition when feature values are missing at both training and test time, (2) a framework for selecting previous learning tasks to transfer to new learning problems, and several inductive transfer algorithms. The problems of active learning (acquiring missing feature values and instance labels) and transfer learning are of great interest to the machine learning community, as it moves towards more complex real-world applications, (3) an ensemble-based approach for handling missing data in classification learning, and (4) the Skill Bootstrapping framework for integrating HTN planning and reinforcement learning to acquire abstract skills in complex domains.

We anticipate that the techniques we have developed will lead to useful discovery tools for scientists in various disciplines. Our first application problem from planetary science was how to best select a subset of Mars rover images to simultaneously accommodate bandwidth limitations and mission science goals. Our recent work with text features promises to greatly extend the applicability of our subset selection methods to apply to other domains such as news article selection and automatic bibliography construction. The IVC (Interactive Visual Clustering) work can be applied to a variety of domains, allowing users to discover patterns (clusters) within data sets without making strong a priori assumptions about the nature of the desired clusters. The Test-Cost Sensitive Regression permits regression methods to be extended to cases where feature acquisition costs can be modeled, improving predictive performance while minimizing data gathering costs.

The most obvious application of our Cost-Sensitive Feature Acquisition (CFA) method is in the medical diagnosis domain, where each feature (e.g., a test result for a patient) may have a different associated cost, and there is a desire to only request tests that are necessary to determine the patient's diagnosis and improve their expected outcome. However, this technology can benefit scientific investigations in many other fields, such as Mars rover exploration, where the features are obtained from a variety of different instruments, with different power consumption and integration time costs.

Our three new application areas (predicting biological activity based on land use and land cover predicting crop yields from satellite data, and diagnosis and prediction of disease progression in multiple sclerosis patients) can benefit from the entire range of techniques we have developed. Gathering additional data in these areas may require additional or redirected resources, so cost-sensitive methods are critical. Interactive techniques for clustering the data may help to provide insight into the dynamics and structure of the domains. Learning transfer will permit us to apply models that are learned in one geographical region, or for one type of organism or crop, to a different region or organism.

Our Cost-sensitive Feature Selection (CFA) method provides an entirely new approach to data analysis in the radio astronomy domain. The state of the art involves an exhaustive computation of derived features to detect new, transient sources. It is known that this simply will not scale to next-generation ratio telescope arrays, such as the Square Kilometer Array (composed of more than 3,000 telescopes) that will begin construction in the next few years. Software advances to enable the selective computation of only the most relevant, cost-effective features will make scaling to SKA-level data volumes possible. We will soon begin testing this concept using radio data collected by the Parkes Observatory in Australia, with synthetic pulses injected into the data to provide controlled tests with known sources. Ultimately, an approach such as CFA could be a game-changing technology for radio arrays and other instruments generating massive data volumes.

Publications

Wagstaff, K.L., desJardins, M., and Eaton, E. "Modeling and Learning User Preferences Over Sets," *Journal of Experimental and Theoretical Artificial Intelligence*, doi:10.1080/09528130903119336, 2009.

Eaton, E. and desJardins, M. "Set-based boosting for instance-level transfer," *Working Notes of the ICDM-2009 Workshop on Transfer Mining*, 2009.

MacGlashan, J. and desJardins, M. "Hierarchical skill learning for high-level planning," *Working Notes of the ICML/UAI/COLT Workshop on Abstraction in Reinforcement Learning*, 2009.

desJardins, M., MacGlashan, J., and Wagstaff, K.L. "Confidence-based feature acquisition to minimize training and test costs," *Proceedings of the SIAM Conference on Data Mining*, p. 514-524, 2010.

Miner, D., MacGlashan, J., and desJardins, M. "A game playing system for use in computer science education," *Proceedings of The 23rd International Florida Artificial Intelligence Research Society (FLAIRS)*, p. 305-310, 2010.

Training and Development

James MacGlashan, who was supported on this award first as an undergraduate researcher, is now in the Ph.D. program and supported as a graduate research assistant. He worked on the IVC user study and has been supporting our work on cost-sensitive feature acquisition and environmental modeling. James is exploring the integration of HTN planning and reinforcement learning for his dissertation research, and advanced to candidacy in May 2010. He has published several papers, including work on IVC, CFA, his dissertation research, and a web-based system for teaching students about game theory in multi-agent systems.

Kevin Winner, who was an undergraduate researcher in 2009-10, graduated in May 2010 and is continuing for his M.S. at UMBC.

Eric Eaton, who worked on the development of preference learning and reuse of learned models via knowledge transfer, received his Ph.D. in May 2009. He is currently a Senior Computer Scientist at Lockheed Martin's Advanced Technology Laboratories.

Outreach and Service Activities

- Marie desJardins was invited to present a CRA-W Distinguished Lecture at Waterloo University in Ontario, Canada. While at Waterloo, she also participated on a career panel for undergraduates and graduate students.
- Marie desJardins served on a panel on "Turning Failures into Lessons Learned" organized by WISE-Grad (Women in Science and Engineering Graduate Students) at UMBC in November 2009.
- Marie desJardins gave a guest lecture on "Academia and Artificial Intelligence" at the UMBC Honors Forum in September 2009.
- Marie desJardins was appointed an Honors Faculty Fellow at UMBC for a 2-year term beginning in Fall 2010.
- Marie desJardins is serving as the Faculty Advisor for the 2010 Workshop for Women in Machine Learning, to be collocated with NIPS in December 2010.
- Kiri Wagstaff accepted a 3-year appointment as an Action Editor for the Machine Learning journal.
- Kiri Wagstaff served as a reviewer and a career panelist for the 2009 AAAI/SIGART Doctoral Consortium. Dr. desJardins also served as a reviewer for this event.
- Kiri Wagstaff and Marie desJardins both served as reviewers and mentors for the 2010 AAAI/SIGART Doctoral Consortium. Dr. Wagstaff also served as a career panelist at this event.
- Marie desJardins and Kiri Wagstaff served on the organizing committee for the First Symposium on Educational Advances in Artificial Intelligence (EAAI-2010), and co-organized a workshop on Teaching and Mentoring as part of this symposium.
- Marie desJardins is a member of the Imagine of Computing Task Force.
- Marie desJardins served as a program committee member for AAAI-10, ICML-10, and AAMAS-10 (senior PC).
- Marie desJardins was invited to serve as an Associate Editor for the Journal of Artificial Intelligence Research, and continued her appointments as Associate Editor of the Journal of Autonomous Agents and Multiagent Systems and Editorial Board Member for AI Magazine.

- Marie desJardins served as a reviewer for NSF proposal review panels in July 2009 and March 2010.
- Kiri Wagstaff served as an organizer for the 2010 Conference on Intelligent Data Understanding (CIDU).
- Kiri Wagstaff served as a co-editor for a special issue of Machine Learning on Machine Learning in Space.
- Kiri Wagstaff served as a reviewer for Pattern Recognition, Machine Learning, IEEE Transactions on Knowledge and Data Engineering, Data Mining and Knowledge Discovery, and the 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings.
- James MacGlashan and several other UMBC students (some of whom are supported on other NSF projects) published a paper at FLAIRS-2010 on a Game Theory Playing website for research and education. The website allows users to create game-theoretic games, such as the Iterated Prisoner's Dilemma, to submit the code for game-playing agents, and to create tournaments in which different users' agents compete against each other in the various games. This website could be used by other researchers to test new algorithms against a wide range of existing methods, or by educators to teach students AI and multi-agent systems concepts.