Research on Statistical Relational Learning at the University of Washington

Pedro Domingos, Yeuhi Abe, Corin Anderson,[1] AnHai Doan,[2] Dieter Fox, Alon Halevy,
Geoff Hulten, Henry Kautz, Tessa Lau,[3] Lin Liao, Jayant Madhavan, Mausam,
Donald J. Patterson, Matthew Richardson, Sumit Sanghai, Daniel Weld, Steve Wolfman
Department of Computer Science and Engineeering
University of Washington, Seattle, WA 98195-2350
pedrod@cs.washington.edu

Abstract

This paper presents an overview of the research on learning statistical models from relational data being carried out at the University of Washington. Our work falls into five main directions: learning models of social networks; learning models of sequential relational processes; scaling up statistical relational learning to massive data sources; learning for knowledge integration; and learning programs in procedural languages. We describe some of the common themes and research issues arising from this work.

1 Introduction

The machine learning group at the University of Washington is pursuing applications in viral marketing, Web search, adaptive Web navigation, assisted cognition, planning, knowledge integration, and programming by demonstration. In each of these areas, we began with methods that were either statistical but not relational or vice-versa, but the need for statistical relational learning (SRL) rapidly became apparent. As a result, our current focus is both on fundamental issues in SRL that cut across these applications, and on propagating advances in the fundamental issues to the applications. What follows is an overview of these research directions, showing how the need for SRL arose in each application, what fundamental issues we uncovered, what progress we have made, and the wealth of problems that remain for future work.

2 Social Networks

Statistical models of customer behavior are widely used in direct marketing. Typically, these models predict how likely the customer is to buy a product based on properties of the customer and/or the product. We have extended these models by also taking into account the *network of influence* among customers [Domingos and Richardson, 2001; Richardson and Domingos, 2002b]. This takes "word of mouth" effects into account—the fact that a customer's decision to buy is affected by what her/his friends and acquaintances say about the product. This makes it possible to design optimal *viral marketing* strategies, which choose which customers to market to based not only on their likelihood of buying, but also on their likelihood of influencing others to buy, and so on recursively. We mine these models from online sources like collaborative filtering systems and knowledge-sharing sites. We have found experimentally that they can lead to much higher profits than traditional direct marketing.

We have also worked on extending Google's PageRank algorithm for Web search with information about the content of pages [Richardson and Domingos, 2002a]. Instead of a universal PageRank measure, we introduce a query-dependent PageRank, and show how to efficiently pre-compute the necessary information at crawl time. Although superficially very different from the viral marketing problem, this problem is in fact isomorphic to it, with the words on Web pages corresponding to customer attributes, and the links between pages corresponding to social relations among customers. (See also [Chakrabarti *et al.*, 1998].)

Notice that, if we view each customer or Web page as a sample, as is usually done, these models imply that samples are no longer independent. Dependence between samples is perhaps the single most fundamental issue that arises in SRL. Even if a domain contains multiple classes of objects, each with different attributes, if the objects are all independent the joint distribution of their attributes decomposes cleanly into a product of distributions for the individual objects. This is the usual non-relational case, with the sole difference that the probabilities for all objects are not all of the same form. It is particularly remarkable that the space of models that assume sample independence is a minuscule fraction of the space of all possible models. In a sense, once the sample independence assumption is made, all further assumptions made by learning algorithms (e.g., choice of representation) are second-order perturbations.

Early studies of the issue of sample dependence in SRL include [Jensen and Neville, 2002b; 2002a], but the area is still very much in its infancy. We are currently developing general methods for this problem, based on assuming inter-sample dependences that are arbitrary but limited in number (the same type of assumption that Bayesian networks make for inter-variable dependences within a sample).

[1] Current affiliation: Google, Inc.

[2] Current affiliation: University of Illinois at Urbana-Champaign.

[3] Current affiliation: IBM T. J. Watson Research Center.

3 Relational Stochastic Processes

Large Web sites are hard to navigate—finding the information the user is looking for often takes too long, and the user gives up and/or wastes time. A possible way to ameliorate this is to automatically adapt the Web site to the user, by predicting what s/he is looking for [Perkowitz and Etzioni, 1997]. For example, we can add to the current page shortcuts to the five pages the user is most likely to want to see. We initially did this using a simple Markov model with pages as states and links as transitions, but found that, although successful, this approach had significant limitations [Anderson *et al.*, 2001]. Predictions can only be made for pages that the user has visited before (and reliable predictions only for pages that the user has visited multiple times). On large Web sites, this is a vanishingly small fraction of all the pages available. Further, as Web sites change over time, it is not possible to make predictions for new pages when they appear. Finally, generalization across Web sites is not possible: even if the adaptive Web navigation system knows the user often goes from the "Books" page to the "Science Fiction" page at Amazon.com, it cannot infer that s/he is likely to do the same at BarnesAndNoble.com.

To overcome these problems, we introduced *relational Markov models (RMMs)* [Anderson *et al.*, 2002]. RMMs model each page as a tuple in a relation, rather than an atomic state. Different pages can belong to different relations (e.g., pages about books will have different properties from pages about consumer electronics products). The variables in each relation can have hierarchically structured domains (e.g., a hierarchy of categories and subcategories of products). We consider all the abstractions of a page that can be obtained by climbing these hierarchies, and compute transition probabilities for the most informative abstractions. These probabilities are then combined into a "ground-level" prediction using shrinkage [McCallum *et al.*, 1998]. Useful predictions can thus be made for previously unvisited pages, by shrinking to abstractions of them that have been visited before (e.g., "Science Fiction Books").

RMMs are an example of a statistical relational model for a sequential domain. (See also [Friedman *et al.*, 1998; Kersting *et al.*, 2003].) However, they are still a restricted representation, in the same way that hidden Markov models are a restricted form of dynamic Bayesian network (DBNs) [Smyth *et al.*, 1997]. We are currently working on a natural generalization: dynamic probabilistic relational models (DPRMs), which extend PRMs [Friedman *et al.*, 1999] to sequential domains in the same way that DBNs extend Bayesian networks. Most processes in the world involve multiple objects and relations and evolution over time, and DPRMs should therefore be widely applicable. For example, in the viral marketing domain, we can model the spread of a product from customer to customer over time, and optimize our marketing actions at each time step, instead of our initial "one-shot" approach.

A key issue in DPRMs, as in DBNs, is efficient inference. The vastness of relational spaces, where the value of a relational variable can be any object in a given class, makes it particularly thorny. We have extended the particle filtering inference method [Doucet *et al.*, 2001] to the relational domain by Rao-Blackwellising [Murphy and Russell, 2001] relational variables conditioned on propositional ones. Initial results show that this approach is extremely effective [Sanghai *et al.*, 2003]. We are currently working on relaxing the assumptions it requires.

DPRMs are well suited to the problem of probabilistic plan recognition — that is, the task of inferring a person's cognitive state in terms of plans and intentions. The Assisted Cognition Project [Kautz *et al.*, 2003] is using DPRMs to track the behavior of a person suffering from cognitive limitations (such as mild dementia) as they go about their day-to-day activities, in order to provide pro-active help in cases of confusion and cognitive errors. Part of this work involves developing techniques for efficiently encoding hierarchical plan networks.

4 Relational Markov Decision Processes

Factored Markov decision processes (MDPs) have proven extremely successful for solving planning tasks in the presence of uncertainty, but they share the same representational weakness which we discussed in the context of Markov models and DBNs earlier. It is natural, therefore, to extend DPRMs to create relational MDPs (RMDPs). Here, state variables are relational fluents instantiated over a set of domain objects, actions are likewise parameterized, and a reward function specifies how much utility is derived from each action and its outcome. The task is to create a control strategy (called a policy) which will maximize the agent's expected discounted reward.

While it is theoretically possible to expand an RMDP into a traditional (ground) MDP, the resulting MDP is often so large that existing value and policy iteration algorithms are incapable of finding a policy. Previous researchers have proposed symbolic methods for decision-theoretic regression [Boutilier *et al.*, 2001], but these techniques are impractical. Instead, we propose generating first-order policies for RMDPs in a three step process [Mausam and Weld, 2003]. First, we create a number of ground MDPs, by instantiating the RMDP with a small set of representative objects. Second, we solve these traditional MDPs with value or policy iteration. Third, we use first-order regression to generate the high-level policy. Our approach is similar to that of Yoon *et al.* [Yoon *et al.*, 2002], but we consider a much more expressive policy representation.

5 Scaling Up

The "killer apps" of SRL are likely to be in domains where the sources of data are vast and varied. In small domains, propositionalizing the problem at some cost in human labor is often feasible. However, given that the space and time cost of a join are worst-case exponential in the number of relations being joined, in large domains this will generally not be an option. Many relational learners work by propositionalizing parts of the data on the fly (e.g., by adding attributes of related objects to the attributes of the objects of interest), and applying a propositional learner to the result [Dzeroski, 1996]. Doing this efficiently is a key but difficult problem, particularly when the relations involved do not all fit in main memory, and must be read from disk. We are currently addressing this using subsampling techniques in two ways [Hulten *et al.*, 2003]. The first is to minimize the number of tuples that need to be read and joined, while ensuring that the sufficient statistics (and consequently the model) obtained from them is essentially the same that would be obtained from the full database. The second is to minimize the number of tuples that are used in computing an aggregate (e.g., sum, average, count), again ensuring that the result is not significantly different from what we would obtain using all the relevant tuples. This is based on our previous work in applying subsampling techniques to propositional learners [Domingos and Hulten, 2000; Hulten and Domingos, 2002]. Beyond this, we envisage that intelligent control of which tuples a learner looks at, and which join paths it pursues, will be key to scalable SRL. Heuristics for this are thus an important area of research.

6 Knowledge Integration

In traditional learning, data must first be gathered, cleaned, integrated and massaged into a single table. This process typically consumes the majority of the resources of a machine learning project. A key part of the promise of SRL is its potential to reduce or bypass parts of it: a statistical relational learner could in principle gather its own data across multiple sources, including different databases, the Web, etc., as needed for learning. However, to fulfill this potential, SRL must be able to bridge the differences in vocabulary that disparate data sources inevitably exhibit: different ontologies, different names for the same attributes, different representations of the same object, etc. Fortunately, SRL techniques can themselves be applied to help solve this “Babel problem.” Given some manually created mappings between information sources, we can learn generalizations of them that allow us to map new sources automatically. We have done this successfully for relational and XML data [Doan *et al.*, 2001; 2003b] and for Semantic Web ontologies [Doan *et al.*, 2002] for the case of one-to-one mappings, and are currently extending our approach to many-to-one mappings [Doan *et al.*, 2003a]. This approach is based on using a variety of learners to extract different kinds of mapping knowledge, combining their outputs with a meta-learner, and combining the result with different types of constraint, domain knowledge, and user feedback to produce the final mapping.

More generally, SRL lends itself particularly well to knowledge-intensive learning, because it allows input knowledge to be expressed in a rich relational language, and is potentially tolerant of noise in this input. We have designed an architecture for incorporating knowledge from a large number of sources into a learner, which uses SRL techniques to handle inconsistency among sources and high variability in source quality [Richardson and Domingos, 2003a]. Specifically, we use a Bayesian logic program representation [Kersting, 2000], with knowledge-based model construction to extract the Bayesian network required to answer a given query [Ngo and Haddawy, 1997]. Horn clauses with the same consequent are combined using a noisy OR, logistic regression or logarithmic pool. The coefficient of a clause in this combination is effectively the system’s estimate of the quality of the clause, and is estimated from query answers and evidence using the EM algorithm [Koller and Pfeffer, 1997]. We have successfully applied this approach in a printer troubleshooting domain. We are also exploring the use of social network models to form estimates of the quality of knowledge contributed by different users, bootstraping each user’s assessment of the quality of a few others to the entire network of contributors [Richardson *et al.*, 2003].

In general, many different types of knowledge can potentially be integrated into SRL, and we are exploring this spectrum. One such type of knowledge is statements about the dependencies among variables of interest (i.e., about the structure of the Bayesian network representing the joint distribution of these variables). We have developed a method for combining statements from a variety of noisy, inconsistent sources into a single probability distribution over the network structure [Richardson and Domingos, 2003b]. This distribution can then be used as the structure prior in a standard Bayesian network learner. The method is based on postulating a simple generative model for expert statements given the true network, and inverting this using Bayes’ theorem to obtain a distribution over possible networks. Our experiments show that even a small number of noisy sources can be sufficient to obtain high-quality estimates of the structure, and high-performing models as a result. We are currently extending this approach to allow Horn rules as an additional form of noisy, partial knowledge about an underlying probability distribution. Based on our experience in the printer troubleshooting domain, we expect this to be more flexible and effective than the more traditional form of knowledge-based model construction.

7 Learning Procedures

We believe that the goal of SRL should be to learn statistical models of any type of structured information, not just (for example) relational databases or Horn knowledge bases. This includes statistical models of procedures performed by humans, and of programs in procedural languages (e.g., Java, Python, C/C++). We have been pursuing applications in programming by demonstration (PBD), where the learner infers a general procedure from examples of its execution by a user (e.g., changing bibilography from one format into another). We initially approached this in a non-statistical setting, defining version spaces over procedures, and defining a version space algebra to build up complex version spaces from “atomic” ones via operations like union and join [Lau *et al.*, 2003b]. We applied this in the SMARTedit system, which learns text-editing procedures by demonstration. Our experience with this system led us to extend the version space algebra with probability distributions over version spaces, to allow incorporating knowledge from the PBD application designer on which (sub)procedures are more and less likely, and to be more flexible and noise-resistant in recognizing procedures. This can be crucial in arriving at a “best guess” as to what the user’s intentions are in any given interaction. More recently, we have begun to extend this framework to learning programs with a full range of programming constructs [Lau *et al.*, 2003a].

8 Conclusion

This paper presented an overview of recent research on statistical relational learning at the University of Washington. Our work spans applications, fundamental issues, and the interplay between them. Applications we are working on include Web search, Web personalization, viral marketing, assisted cognition, planning, information integration, and programming by demonstration. Fundamental issues we have begun to make progress on include: learning in the presence of interdependencies among samples; modeling stochastic dynamics in relational domains; scaling up; learning across sources with different representations; and extending SRL beyond Horn clauses and relational databases.

Acknowledgments

This research was partly funded by an NSF CAREER Award and an IBM Faculty Partnership Award to the first author, ONR grants N00014-02-1-0408 and N00014-02-1-0932, NASA grant NAG 2-1538, and a gift from the Ford Motor Co.

References

[Anderson *et al.*, 2001] C. Anderson, P. Domingos, and D. Weld. Adaptive Web navigation for wireless devices. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 879–884, Seattle, WA, 2001. Morgan Kaufmann.

[Anderson *et al.*, 2002] C. Anderson, P. Domingos, and D. Weld. Relational Markov models and their application to adaptive Web navigation. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 143–152, Edmonton, Canada, 2002.

[Boutilier *et al.*, 2001] C. Boutilier, R. Reiter, and B. Price. Symbolic dynamic programming for first-order MDPs. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 690–697, Seattle, WA, 2001. Morgan Kaufmann.

[Chakrabarti *et al.*, 1998] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 307–318, Seattle, WA, 1998. ACM Press.

[Doan *et al.*, 2001] A. Doan, P. Domingos, and A. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pages 509–520, Santa Barbara, CA, 2001. ACM Press.

[Doan *et al.*, 2002] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the Semantic Web. In *Proceedings of the Eleventh International World Wide Web Conference*, pages 662–673, Honolulu, HI, 2002. ACM Press.

[Doan *et al.*, 2003a] A. Doan, P. Domingos, and A. Halevy. Learning complex semantic mappings between structured representations. 2003. Submitted.

[Doan *et al.*, 2003b] A. Doan, P. Domingos, and A. Halevy. Learning to match the schemas of data sources: A multistrategy approach. *Machine Learning*, 50:279–301, 2003.

[Domingos and Hulten, 2000] P. Domingos and G. Hulten. Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 71–80, Boston, MA, 2000. ACM Press.

[Domingos and Richardson, 2001] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, San Francisco, CA, 2001. ACM Press.

[Doucet *et al.*, 2001] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[Dzeroski, 1996] S. Dzeroski. Inductive logic programming and knowledge discovery in databases. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 117–152. AAAI Press, Menlo Park, CA, 1996.

[Friedman *et al.*, 1998] N. Friedman, D. Koller, and A. Pfeffer. Structured representation of complex stochastic systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 157–164, Madison, WI, 1998.

[Friedman *et al.*, 1999] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1300–1307, Stockholm, Sweden, 1999. Morgan Kaufmann.

[Hulten and Domingos, 2002] G. Hulten and P. Domingos. Mining complex models from arbitrarily large databases in constant time. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 525–531, Edmonton, Canada, 2002. ACM Press.

[Hulten *et al.*, 2003] G. Hulten, P. Domingos, and Y. Abe. Mining massive relational databases. In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, Acapulco, Mexico, 2003. This volume.

[Jensen and Neville, 2002a] D. Jensen and J. Neville. Autocorrelation and linkage cause bias in evaluation of relational learners. In *Proceedings of the Twelfth International Conference on Inductive Logic Programming*, Sydney, Australia, 2002. Springer.

[Jensen and Neville, 2002b] D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 259–266, Sydney, Australia, 2002. Morgan Kaufmann.

[Kautz *et al.*, 2003] H. Kautz, O. Etzioni, D. Fox, D. Weld, and L. Shastri. Foundations of assisted cognition systems. Technical Report CSE-03-AC-01, Department of

Computer Science and Engineering, University of Washington, Seattle, WA, 2003.

[Kersting *et al.*, 2003] K. Kersting, T. Raiko, S. Kramer, and L. De Raedt. Towards discovering structural signatures of protein folds based on logical hidden Markov models. In *Proc. 8th Pacific Symposium on Biocomputing*, Kauai, HI, 2003.

[Kersting, 2000] K. Kersting. *Bayesian Logic Programs*. PhD thesis, University of Freiburg, Freiburg, Germany, 2000.

[Koller and Pfeffer, 1997] D. Koller and A. Pfeffer. Learning probabilities for noisy first-order rules. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 1316–1321, Nagoya, Japan, 1997. Morgan Kaufmann.

[Lau *et al.*, 2003a] T. Lau, P. Domingos, and D. Weld. Learning programs from traces using version space algebra. 2003. Submitted.

[Lau *et al.*, 2003b] T. Lau, S. Wolfman, P. Domingos, and D. Weld. Programming by demonstration using version space algebra. *Machine Learning*, 2003. To appear.

[Mausam and Weld, 2003] Mausam and D. Weld. Solving relational MDPs with first-order machine learning. In *Proceedings of the ICAPS-2003 Workshop on Planning under Uncertainty and Incomplete Information*, Seattle, WA, 2003.

[McCallum *et al.*, 1998] A. McCallum, R. Rosenfeld, T. Mitchell, and A. Y. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 359–367, Madison, WI, 1998. Morgan Kaufmann.

[Murphy and Russell, 2001] K. Murphy and S. Russell. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, pages 499–516. Springer, New York, 2001.

[Ngo and Haddawy, 1997] L. Ngo and P. Haddawy. Answering queries from context-sensitive probabilistic knowledge bases. *Theoretical Computer Science*, 171:147–177, 1997.

[Perkowitz and Etzioni, 1997] M. Perkowitz and O. Etzioni. Adaptive Web sites: An AI challenge. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 16–21, Tokyo, Japan, 1997. Morgan Kaufmann.

[Richardson and Domingos, 2002a] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in PageRank. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1441–1448. MIT Press, Cambridge, MA, 2002.

[Richardson and Domingos, 2002b] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 61–70, Edmonton, Canada, 2002. ACM Press.

[Richardson and Domingos, 2003a] M. Richardson and P. Domingos. Building large knowledge bases by mass collaboration. 2003. Submitted.

[Richardson and Domingos, 2003b] M. Richardson and P. Domingos. Learning with knowledge from multiple experts. 2003. Submitted.

[Richardson *et al.*, 2003] M. Richardson, R. Agrawal, and P. Domingos. Building the Semantic Web by mass collaboration. 2003. Submitted.

[Sanghai *et al.*, 2003] S. Sanghai, P. Domingos, and D. Weld. Dynamic probabilistic relational models. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 2003. Morgan Kaufmann.

[Smyth *et al.*, 1997] P. Smyth, D. Heckerman, and M. I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9:227–269, 1997.

[Yoon *et al.*, 2002] S. Yoon, A. Fern, and R. Givan. Inductive policy selection for first-order Markov decision processes. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, Edmonton, Canada, 2002. Morgan Kaufmann.