Structure Learning for Statistical Relational Models

Jennifer Neville
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
jneville@cs.umass.edu

Many data sets routinely captured by businesses and organizations are relational in nature yet over the past decade most machine learning research has focused on "flattened" propositional data. Propositional data record the characteristics of a set of homogeneous and statistically independent objects; relational data record characteristics of heterogeneous objects and the relations among those objects. Examples of relational data include citation graphs, the World Wide Web, genomic structures, fraud detection data, epidemiology data, and data on interrelated people, places, and events extracted from text documents. Relational data offer unique opportunities to boost the accuracy of learned models and improve the quality of decision-making if the algorithms can learn effectively from the additional information the relationships provide.

Relational data representations greatly expand the range and applicability of machine learning techniques, but the greater expressive power of relational representations produces new statistical challenges. The data often have irregular structures and complex dependencies that contradict the assumptions of conventional modeling techniques. First, algorithms for propositional data assume that the data instances are recorded in homogeneous structures (a fixed set of fields for each object), but relational data instances are usually more varied and complex. For example, molecules have different numbers of atoms and bonds, and web pages have different numbers of incoming and outgoing links. The ability to generalize across heterogeneous data instances is a defining characteristic of relational learning algorithms. Second, algorithms designed for attribute-value data assume the data are independent and identically distributed. Relational data, on the other hand, have dependencies both as a result of direct relations (e.g., hyperlinked pages) and through chaining of multiple relations (e.g., pages linked through the same directory page). Recent work has shown that these dependencies can be exploited to improve classification accuracy if inferences about related data instances are made simultaneously (Chakrabarti, Dom, & Indyk 1998; Neville & Jensen 2000; Taskar, Abbeel, & Koller 2002; Macskassy & Provost 2003; Jensen, Neville, & Gallagher 2004). However, the dependencies among instances also complicate efforts to learn accurate statistical models because the traditional assumption of instance independence is violated (Jensen & Neville 2002; 2003).

There are a number of relational models that are used for *individual inference*, where inferences about one instance are not used to inform the inferences about related instances (e.g., Perlich & Provost 2003, Popescul *et al.* 2003, Neville *et al.* 2003). These approaches consider relational instances represented as independent, disconnected subgraphs (e.g., molecules). Such models can represent and reason with the complex relational structure of a single instance. However, they do not attempt to model the relational structure among instances—thus removing the need (and the opportunity) for simultaneous *collective inference*. Joint relational models (Getoor *et al.* 2001; Taskar, Abbeel, & Koller 2002; Neville & Jensen 2004), on the other hand, model the interconnections among relational instances that are represented as one large graph. These approaches are able to exploit *relational autocorrelation* to improve classification performance by estimating joint probability distributions over the entire graph and *collectively* inferring the labels of related instances (Jensen, Neville, & Gallagher 2004). Autocorrelation is a statistical dependency between the values of the same variable on related entities and is a common characteristic of many relational data sets. It is easy to see how autocorrelation could be used to improve the predictions of statistical models in relational domains. For example, consider the problem of automatically predicting the topic of a scientific paper (e.g., neural networks, reinforcement learning, genetic algorithms). One simple method for predicting topics would look at papers in the context of their citation graphs. It is possible to predict a given paper's topic with high accuracy based on the topics of neighboring papers because there is high autocorrelation in the citation graph—papers tend to cite other papers with the same topic.

Research on joint relational models has focused primarily on knowledge representation and inference—there has been little attention paid to the challenges and opportunities that are unique to learning in relational domains. For example, probabilistic relational models (PRMs) (Getoor *et al.* 2001) extend Bayesian networks to support reasoning in complex relational domains by defining a template of typed dependencies, tying parameters across object of the same type, and aggregating over heterogeneous sets of attribute values.

Relational Markov networks (RMNs) (Taskar, Abbeel, & Koller 2002) extend Markov networks in a similar manner, with typed clique templates, parameter tying, and aggregations over probabilities rather than values. Within these extended representations, conventional propositional learning techniques are used for both parameter estimation and structure learning.

This approach is a good first step but there are a number of limitations to using propositional learning techniques in a relational setting. First, dependencies among instances, combined with varying relational structure, will increase the variance of parameter estimates (Jensen & Neville 2002). This inefficiency in parameter estimation will cause the learning algorithms to overfit and may result in biased structure learning. One solution is to use a non-selective model, that is, a model that doesn't do feature selection. However, for relational tasks, which are likely to have a large number of features, this lack of selectivity will make the model more difficult to interpret and will increase the computational burden for inference. Second, the search space is potentially much larger than in propositional domains, making conventional generate-and-test approaches to structure learning intractable. When the data are propositional, the possible dependencies are exponential in the number of attributes. When the data are relational, the possible dependencies are exponential in the number of attributes *and the number of instances*. Third, restrictions to the model space to make structure learning tractable, generally involve a change in representation that presents a *chicken and egg* problem. Representation can impair the ability to learn important knowledge, but knowing the right representation often requires just that knowledge. To date there has been little work that has explored this issue.

This work will consider in depth the issue of structure learning in statistical relational models. There are three aspects to developing accurate and efficient structure learning techniques: (1) hypothesis testing, (2) structuring the search space, and (3) efficient search of the space. We propose to develop accurate hypothesis testing techniques that will adjust for the widely varying structure and dependencies among instances in relational data. In addition, we will investigate approaches to structuring the search space, offering alternatives to the current naive approach of general-to-specific ordering, which will break down as the data complexity grows. Specifically, we will consider a view-learning approach, which uses predicate invention to decouple the search space into a set of biased abstractions, allowing the search to consider a wider range of dependencies. Finally, we propose to develop more efficient search techniques for model selection. By interleaving structure learning, parameter estimation, and inference, we will constrain the search using the structure of data in a bottom-up fashion.

Our work to date has identified a number of key characteristics in relational data (e.g., heterogeneous relational structure, autocorrelation), their adverse effects on learning, and initial strategies to adjust for these characteristics (Jensen & Neville 2002; 2003). This work will be extended to analyze hypothesis testing techniques used in learning directed and undirected joint models, and to develop accurate structure learning techniques for these models. In addition, we are currently investigating abstractions that decouple structure and attribute dependencies with latent groups as a means to structure the search space. This work will be combined with our investigation of the properties of aggregation techniques (Neville, Jensen, & Gallagher 2003) and approximation techniques (Neville & Jensen 2004) to develop accurate and efficient search techniques for relational domains.

This work is positioned to extend the range, applicability, and performance gains of joint statistical relational models. A sufficient number of relational representations have been developed to start generalizing their characteristics and extending their performance. The models have been successfully applied in a number of domains, including the World Wide Web, genomic structures, and citation graphs, but progress is hindered by the complexity of learning and inference. If we do not focus on techniques for structure learning in relational models, we risk learning models with uninterpretable structure, poor generalization, and inefficient performance. The benefits of accurate and efficient structure learning are two-fold. First, in small datasets, where we have the computational resources to apply current learning and inference techniques, improved structure learning will increase the accuracy of the models. Second, in large datasets, where learning and inference are computationally intensive, if not intractable, improved structure learning will make relational modeling both practical and feasible.

References

Chakrabarti, S.; Dom, B.; and Indyk, P. 1998. Enhanced hypertext categorization using hyperlinks. In *SIGMOD-1998*, 307–318.

Getoor, L.; Friedman, N.; Koller, D.; and Pfeffer, A. 2001. Learning probabilistic relational models. In *Relational Data Mining*. Springer-Verlag.

Jensen, D., and Neville, J. 2002. Linkage and autocorrelation cause feature selection bias in relational learning. In *ICML-2002*.

Jensen, D., and Neville, J. 2003. Avoiding bias when aggregating relational data with degree disparity. In *ICML-2003*, 274–281.

Jensen, D.; Neville, J.; and Gallagher, B. 2004. Why collective inference improves relational classification. In *SIGKDD-2004*.

Macskassy, S., and Provost, F. 2003. A simple relational classifier. In *KDD-2003 Workshop on Multi-Relational Data Mining*.

Neville, J., and Jensen, D. 2000. Iterative classification in relational data. In *AAAI-2000 Workshop on Learning Statistical Models from Relational Data*.

Neville, J., and Jensen, D. 2004. Dependency networks for relational data. In *ICDM-2004*.

Neville, J.; Jensen, D.; Friedland, L.; and Hay, M. 2003. Learning relational probability trees. In *ACM SIGKDD-2003*, 625–630.

Neville, J.; Jensen, D.; and Gallagher, B. 2003. Simple estimators for relational Bayesian classifers. In *ICDM-2003*.

Perlich, C., and Provost, F. 2003. Aggregation-based feature invention and relational concept classes. In *KDD-2003*.

Popescul, A.; Ungar, L.; Lawrence, S.; and Pennock, D. 2003. Statistical relational learning for document mining. In *ICDM-2003*.

Taskar, B.; Abbeel, P.; and Koller, D. 2002. Discriminative probabilistic models for relational data. In *UAI-2002*.