

Statistical Relational Learning: Four Claims and a Survey

Jennifer Neville, Matthew Rattigan, David Jensen

Knowledge Discovery Laboratory, Department of Computer Science, University of Massachusetts,
140 Governors Drive, Amherst, MA 01003 USA
{jneville | rattigan | jensen } @cs.umass.edu

Statistical relational learning (SRL) research has made significant progress over the last 5 years. We have successfully demonstrated the feasibility of a number of probabilistic models for relational data, including probabilistic relational models, Bayesian logic programs, and relational probability trees, and the interest in SRL is growing. However, in order to sustain and nurture the growth of SRL as a subfield we need to refocus our efforts on the science of machine learning — moving from demonstrations to comparative and ablation studies. We will outline four assertions that are implicit to SRL research but which have been only minimally evaluated. We hope to stimulate discussion as to how, as a community, these claims can be addressed in future research.

1 Introduction

In the hopes of generalizing the results of recent research from the statistical relational learning (SRL) community, we surveyed twenty recent SRL papers. From the papers studied we were able to distill four implicit claims that underlie much of the current SRL research. We present an examination of those claims in the context of the papers surveyed.

We chose twelve of the papers as a representative sample for the purposes of this discussion. Each paper chosen describes and evaluates a discriminative, probabilistic relational model. A descriptive list of the selected models and papers appears in Table 1.

The purpose of this paper is to stimulate a discussion of the scientific methods that will help to illustrate and evaluate the relative merits of the different models and their frameworks.

2 Relational vs. propositional

Claim: Models learned from both intrinsic and relational information perform better than those learned from intrinsic information alone, and are therefore worth the added complexity.

This is an implicit claim of relational learning in general. We expect that predictive information exists in relationships among instances, and that this information can be used to reduce model bias. However, decreasing bias often results in increased variance (Friedman 1997). This is a very real concern for relational learning algorithms that are faced with an exponential explosion in the size of the model space.

The simplest way to evaluate this claim is to record model performance using *intrinsic* data, a subset of the data where relational information is removed. By this we mean data where the instances are objects in isolation, and the only information available are the attributes intrinsic to those objects as individuals. Popescul, Ungar, Lawrence, and Pennock (2003) use this approach when evaluating their models on citation data, comparing models learned on information intrinsic to documents alone with those learned from both intrinsic and citation information. Getoor, Segal, Taskar and Koller (2001) use an alternative approach, including results from a baseline propositional model learned on intrinsic data. This technique is also employed in four other papers. See figure 2 for details.

More than half of the papers surveyed included some comparative intrinsic analysis, and the results vary considerably across models and datasets. For example, when using relational features Neville, Jensen, Gallagher, and Fairgrieve (2003) found marked improvement in model performance on two datasets, but no significant gain on a third. We believe that this type of analysis is important baseline for determining whether the inclusion of relational information is of any benefit, and if so whether the additional model complexity is warranted. Although preliminary analysis along these lines is a common component of SRL research, we feel that more explicit and directed experimentation is needed to fully justify the use of SRL models for relational datasets.

3 Probabilistic vs. deterministic

Claim: Probabilistic relational models offer significant advantages over deterministic relational models in relational domains.

Table 1: Statistical relational learning models surveyed

Model	Description	Selective	Generative	Reference
RVS	relational vector-space model	No	no	Bernstein, Clearwater, and Provost, 2003
FOIL-PILFS	relational learner w/statistical predicate invention	Yes	no	Craven and Slattery, 2001
Maccent	maximum entropy model with clausal constraints	Yes	no	Dehaspe, 1997
SNM	Markov random field for social networks	No	no	Domingos and Richardson, 2001
BLP	Bayesian logic programs	yes	yes	Kersting and De Raedt, 2002
1BC2	first-order naive Bayesian classifier	no	no	Lachiche and Flach, 2002
RBC	relational Bayes classifier	no	no	Neville, Jensen, Gallagher and Fairgrieve, 2003
RPT	relational probability trees	yes	no	Neville, Jensen, Friedland and Hay, 2003
SLR	structural logistic regression	yes	no	Popescul, Ungar, Lawrence, and Pennock, 2003
NBILP-R	naive Bayes classifier with ILP features	no	no	Pompe and Kononenko, 1995
PRM	probabilistic relational model	yes	yes	Getoor, Segal, Taskar and Koller, 2001
RMN	relational Markov network	no	no	Taskar, Abbeel and Koller, 2002

Research in relational learning has investigated deterministic models for many years (e.g. Muggleton & De Raedt 1994, Lavrac & Dzeroski 1994). Recent efforts have shifted the focus towards a probabilistic setting. We outline a number of advantages of probabilistic models below, but we feel that discussion of the strengths and weaknesses of each technique is worth exploring in greater detail. Discussion along these lines is necessary to come to a general understanding of the range and applicability of SRL models.

One strength of probabilistic models is the ability to evaluate how these models will perform over a range of class and cost distributions (Provost and Fawcett, 1997). Classification tasks involving complex relational data often have varying levels of misclassification costs as well as uncertain class distributions. Since deterministic models do not associate a level of confidence with their classifications, it is difficult to estimate their behavior in these domains.

Another advantage of probabilistic models is their suitability to real-world analysis tasks. Since these models generate meaningful, continuous probability scores, they lend themselves to an iterative, hierarchical approach to analysis. As Bernstein, Clearwater, and Provost (2003) point out, “scores may be most useful as feature constructors in other, more complicated systems.” It is therefore crucial to evaluate the probabilities produced in SRL models quantitatively; unfortunately, none of the papers we surveyed perform this type of evaluation. Secondly, probability scores allow us to rank instances in order of certainty. This is of great use to real-world analysts who have limited time to investigate “positive” instances, as confidence scores allow an analyst to prioritize instances rather than treat all members of a predicted class equally.

Finally, probabilistic models are in general more suited to learning with relational data than deterministic ones. Due to their complexity, relational datasets are often noisy, which can be troublesome for deterministic models (Popescul et al. 2003). Furthermore, the advantage of working with relational data may be lost without the use of probabilistic models. For example, Craven and Slattery (2001) found in the text classification domain that “learned rules will not be dependent on the presence or absence of specific key words as a conventional relational method. Instead, the statistical classifiers in its learned rules consider the weighted evidence of many words.”

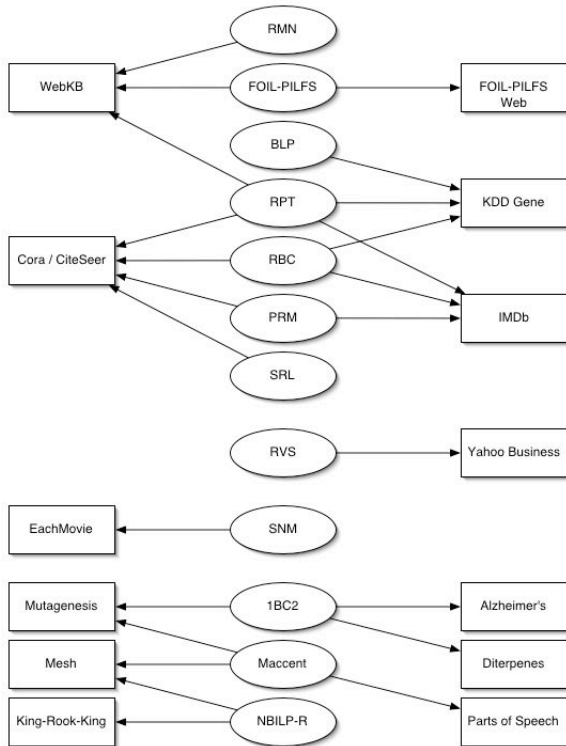


Figure 1: SRL models and evaluation datasets.

4 Heterogeneous data

Claim: SRL algorithms learn accurate models of structured data.

Most conventional classification techniques assume data instances are recorded in homogeneous structures. Relational data however, often have complex structures that are difficult to model in propositional form. For example, information about actors, directors and producers may be useful when building a model of movie success but each movie has a different number of related entities. This variety results in examples with diverse structure — some movies may have 10 actors, and others may have hundreds. The ability to deal with heterogeneous data instances is a defining characteristic of relational learning algorithms.

The relational learning community has developed a number of models that can handle heterogeneous data. For example, Lachiche and Flach (2002) extend conventional naive Bayes classifiers to handle heterogeneous instances and Deshape (1997) extends conventional maximum entropy models to use a richer first-order logic format.

Each of the 12 papers surveyed introduces a different model for this purpose. However, few of these papers evaluate the effects of heterogeneity on the learned models. Some of our recent work has examined how particular characteristics of relational data affect the statistical inferences necessary for accurate learning (Jensen & Neville 2002, Jensen, Neville & Hay 2003). Specifically, we have shown that concentrated linkage combined with high autocorrelation can lead to feature selection bias if models are constructed naively. Also, we have shown that degree disparity can lead to spurious correlations in aggregated features, resulting in overly complex models with excess structures.

These characteristics of relational data can greatly complicate efforts to construct good statistical models. Only selective models are vulnerable to the particular biases mentioned above, but 7 of the models surveyed do some form of selection while learning. It is difficult to evaluate models for unidentified biases; however, comparative studies among the various SRL algorithms should help to uncover these biases. In particular, detailed comparisons of selective and non-selective model performance may help to uncover additional biases. Figure 2 depicts the 12 SRL models with links to the various models compared to during evaluation. The paucity of outlinks speaks for itself.

We have only begun to explore the effects of data characteristics on model learning. While many relational models outperform propositional models on the same datasets, the relational models may not be living up to their full potential. Further investigation of the complexities of relational data will help to identify sources of potential bias and correcting for these biases will unleash the full power of SRL models.

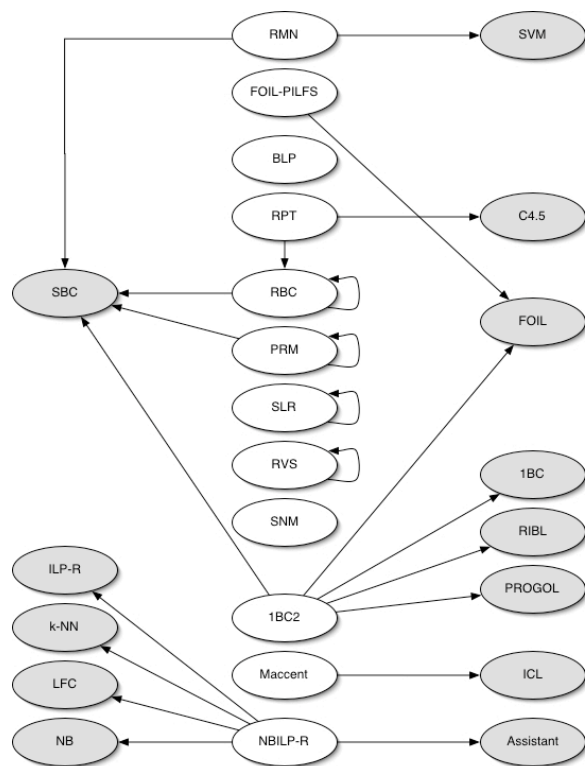


Figure 2: SRL models and evaluation models. Self-loops indicate ablation comparisons.

5 Interdependent data

Claim: SRL algorithms learn accurate models of dependent data instances.

Independence of instances is a deeply buried assumption of traditional machine learning methods that is contradicted by many relational datasets. For example, in scientific literature datasets there are dependencies among papers written by the same author and in web datasets there are dependencies among pages linked to by the same document. The structure of complex relational data such as these presents a unique opportunity for improving the accuracy of statistical models. If two objects are related, inferring something about one object can aid inferences about the other.

In our analysis of relational data, we have encountered many examples of dependencies that could be exploited to improve learning. For example, in analysis of the 2001 KDD Cup data we found that the proteins located in the same place in a cell (e.g., mitochondria or cell wall) had highly autocorrelated functions (e.g., transcription or cell growth). Such autocorrelation has been identified in other domains as well. For example, fraud in mobile phone networks has been found to be highly autocorrelated (Cortes, Pregibon & Volinsky 2001). The topics of authoritative web pages are highly autocorrelated when linked through directory pages that serve as “hubs” (Kleinberg 2001).

Table 2: Characteristics of data and sampling approach used for evaluation

Model	Data connectivity	Sampling approach
RVS	one large connected component	not mentioned
FOIL-PILFS	disjoint large graphs	leave-one-graph-out cross validation
Maccent	disjoint small graphs	leave-k-graph-out cross validation
SNM	one large connected component	sample by time
BLP	one large connected component	transduction
1BC2	disjoint small graphs	leave-k-graph-out cross validation
RBC	one large connected component	subgraph sampling
RPT	one large connected component	subgraph sampling
SLR	one large connected component	transduction
NBILP-R	disjoint small graphs	leave-k-graph-out cross validation
PRM	large conn comp / disjoint graphs	transduction / leave-one-graph-out cv
RMN	disjoint large graphs	leave-one-graph-out cross validation

Many of the models surveyed do not attempt to exploit dependencies among relational instances. More than half of the algorithms are designed to learn models relational datasets with independent, heterogeneous instances (i.i.d. relational data) where any dependencies among instances are ignored.

Inductive logic programming (ILP) models have been capable of representing dependencies among instances for years, albeit only extreme (deterministic) dependencies (Lavrac & Dzeroski 1994). However, it is only recently that statistical models have been developed to exploit the dependencies in relational data. For example, Kersting and De Raedt (2002) combine ILP with Bayesian networks to integrate probabilities into logic programs and model the dependencies among proteins in a cell. Getoor et al. (2001) use probabilistic relational models (PRMs) to model the the dependencies among hyperlinked web pages. Taskar, Abbeel and Koller (2002) use conditional Markov networks to model the same domain. Domingos and Richardson (2001) represent market entities as social networks and develop Markov random field models to model the influence in purchasing patterns throughout the network. Bernstein, Clearwater and Provost (2003) outline a relational vector-space model that uses autocorrelation to identify the group membership of linked entities.

Statistical models capable of collective classification across a network of instances are a relatively new phenomenon. It is unclear how to effectively evaluate the performance of these models. In what context do we expect to be using these models in the real world? Will we be applying the model to a completely new graph or do we expect new instances to arrive temporally related to the existing (training set) graph. Answers to this question should help to develop sampling methods to get an unbiased estimate of model performance.

Furthermore, how should we sample from a large connected graph? Table 2 outlines the characteristics of datasets examined by each of the models along with the sampling approach that was chosen. There are four approaches to sampling currently in use; more work is

needed to determine which of these approaches is appropriate for a particular learning task.

6 Conclusions

Although the SRL community has successfully demonstrated the feasibility of a number of probabilistic models for relational data, there is much work to be done in order to begin generalizing the range and applicability of the various models. We have presented four claims for discussion with the purpose of advancing the science of SRL as well as machine learning in general.

References

- Bernstein, A., S. Clearwater, and F. Provost. The relational vector-space model and industry classification. CDeR working paper #IS-03-02, Stern School of Business, New York University, 2003.
- Cortes, C., D. Pregibon, and C. Volinsky. Communities of Interest. *Proceedings of the Fourth International Symposium on Intelligent Data Analysis*, 2001.
- Craven, M. and S. Slattery. Relational learning with statistical predicate invention: Better models for hypertext. *Machine Learning Journal* 43:97-119, 2001.
- Dehaspe, L. Maximum entropy modeling with clausal constraints. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, pages 109-124. Springer-Verlag, 1997.
- Domingos, P., M. Richardson. Mining the Network Value of Customers. *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*, pp. 57-66, 2001.
- Friedman, J. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55-77, 1997.

Getoor, L., E. Segal, B. Taskar and D. Koller. Probabilistic Models of Text and Link Structure for Hypertext Classification. Proceedings of the IJCAI01 Workshop on Text Learning: Beyond Supervision, 2001.

Jensen, D., J. Neville and M. Hay. Avoiding bias when aggregating relational data with degree disparity. Proceedings of the 20th Int. Joint Conf. on Machine Learning, to appear.

Kersting, K., L. De Raedt. Basic principles of learning Bayesian logic programs. Technical Report No. 174, Institute for Computer Science, University of Freiburg, Germany, June 2002.

Kleinberg, J. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46:604-632, 1999

Lachiche, N. and P. Flach. 1BC2: a True First-Order Bayesian Classifier. Proceedings of the 12th International Conference on Inductive Logic Programming, 2002.

Lavrac, N. and Dzeroski. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, 1994.

Muggleton, S. editor. *Inductive Logic Programming*. Academic Press, 1992.

Neville, J., D. Jensen, B. Gallagher, R. Fairgrieve. Simple Estimators for Relational Data. University of Massachusetts, Technical Report 03-04, 2003.

Neville, J., D. Jensen, L. Friedland, M. Hay. Learning relational probability trees. Proceedings of the 9th International Conference on Knowledge Discovery & Data Mining, to appear.

Popescul, A., L. Ungar, S. Lawrence, D. Pennock, Statistical relational learning for document mining. Submitted, 2003.

Pompe, U. and I. Kononenko. Naive Bayesian classifier within ILP-R. Proceedings of the 5th International Workshop on Inductive Logic Programming, pages 417-436, 1995.

Provost, F. and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. Proceedings of the 3rd International Conference on Knowledge Discovery & Data Mining, pages 43-48, 1997.

Taskar, B., P. Abbeel and D. Koller. Discriminative probabilistic models for relational data. Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, 2002.