

## Statistical Relational Learning at U Penn

### Alexandrin Popescul

Computer and Information Science  
University of Pennsylvania  
Philadelphia, PA 19104  
popescul@cis.upenn.edu

### Dean P. Foster

Department of Statistics  
University of Pennsylvania  
Philadelphia, PA 19104  
foster@gosset.wharton.upenn.edu

### Lyle H. Ungar

Computer and Information Science  
University of Pennsylvania  
Philadelphia, PA 19104  
ungar@cis.upenn.edu

*We do statistical relational learning by incrementally extracting data from a relational database, and computing features of that data which are then used in a classical discriminative statistical model component. Candidate features for the model are generated by a structured search in the space of relational database queries and selected using statistical information criteria. The structuring of the search space is inspired by techniques in inductive logic programming (ILP), but the use of statistical modeling relaxes the necessity of limiting the search space to logical expressions. We use a rich feature space that includes clusters, which can be generated incrementally and used to augment the basic relational schema. Current areas of research include determining optimal model selection criteria for use in this setting where an infinite sequence of features can be incrementally generated and the use of intelligent search heuristics to focus search on more promising subspaces.*

A growing number of machine learning applications of high interest involves the analysis of data which is both noisy and is of complex relational structure. This dictates a natural choice in such domains: the use of statistical rather than deterministic modeling and relational rather than propositional representation [Popescul *et al.*, 2002]. Classical statistical learners provide powerful modeling component but are often limited to a “flat” file propositional domain representation where potential features are fixed-size attribute vectors. Often the manual process of preparing such attributes is costly and not obvious when more complex regularities are involved. We are developing a methodology which combines the strengths of classical statistical models with the higher expressivity of features automatically generated from a relational database.

Our interest in statistical relational learning developed while working on modeling in CiteSeer<sup>1</sup>, an online digital library of computer science papers. CiteSeer contains a rich set of relational data, including citation information, the text of titles, abstracts and documents, author names and affiliations, conference or journal names. Applications we have addressed include: i) prediction in social networks, e.g. link prediction: given two papers estimate whether they cite each other [Popescul and Ungar, 2003], ii) document classification, modeling of more complex features than traditional word counts improves classification accuracy [Popescul *et al.*, 2003]. We are planning to apply statistical relational learning in bioinformatics domains, in particular for prediction of protein-protein interactions.

Figure 1 highlights the main components of our learning setting. Two main processes—relational feature generation and statistical modeling—are coupled into a single loop. Knowing which features have been selected by the statistical modeler allows the query generation component to guide its search, focusing on promising subspaces of the feature space.

Our statistical relational learning approach has several key features which distinguish it from either pure probabilistic modeling or inductive logic programming.

- We assume an application domain in which there are many entities connected by many relations (e.g. a patient database in a hospital), in which complex features (e.g. a set of patients clustered by the similarity of the symptoms, and treated by doctors working in the same clinic) are highly predictive of outcomes of interest (e.g. expected stay in the hospital). In such areas, it is generally not feasible to build a large generative model (e.g. a PRM) of the world, and a more focussed exploration of the space of possible relations is needed.
- Our search in the query space is an instance of propositionalization, as proposed in the inductive logic programming community; however, the use of statistics rather than logic allows the formulation of rich feature spaces, extending far beyond boolean-valued features. This richer search space can include statistical summaries or aggregates, more

---

<sup>1</sup><http://citeseer.org/>

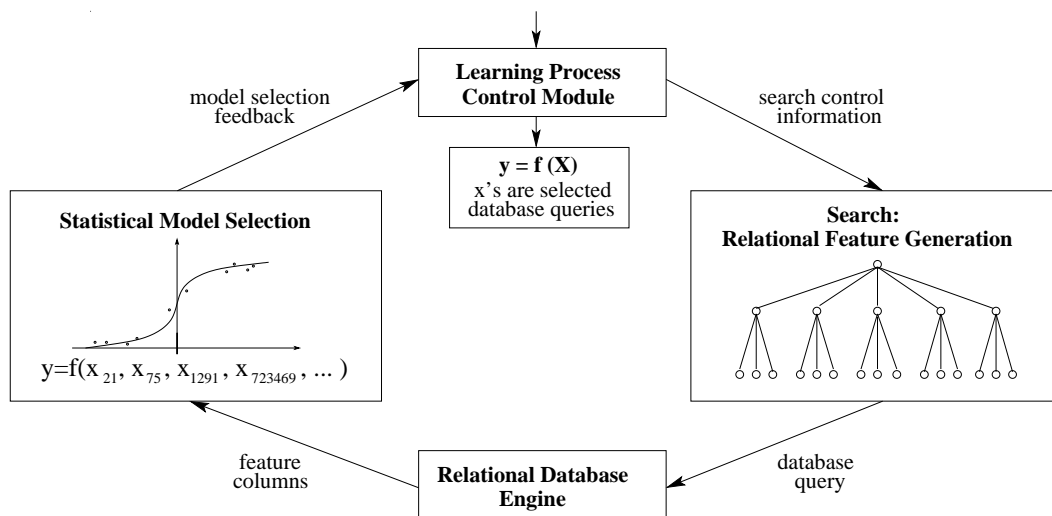


Figure 1: Learning process diagram. The search in the space of database queries involving one or more relations produces feature candidates one at a time to be considered by the statistical model selection component. The process results in a statistical model where each selected feature is the evaluation of a database query encoding a predictive data pattern in a given domain.

expressive substitutions through nesting of intermediate aggregates (e.g., how many times does this publication cite the most cited author in conference to which it was submitted?) A key question is how best to define the search space and how to control the search space complexity and search space bias.

- We use clustering to extend the set of relations generating new features. Clusters improve modeling of sparse data, improve scalability, and produce richer representations [Foster and Ungar, 2002]. New clusters can be derived using the same features used in the statistical modeling. For example, one can cluster words based on co-occurrence in documents, giving “topics”, or authors based on the number of papers they have published in the same venues, giving “communities.” Once clusters are formed, they represent new relationships (e.g. `on_topic_3(paper1798)` or `in_community_5(author7)`), which can be added to the relational database schema, and then used interchangeably with the original relations.
- Learning takes place with an exponential number of potential feature candidates, only relatively few of which are expected to be useful. Feature selection methods recently derived by statisticians give promising results for handling this potentially infinite stream of features with only a finite set of observations.
- Our formulation supports sophisticated procedures for determining which subspaces of the query space to explore. Intelligent search techniques which combine the relational structure of the data, feedback from the feature selection algorithm, and other information such as sampling from feature subspaces to determine their promise will help scale to truly large problems.
- We use relational database management systems (RDMSs) and SQL rather than Prolog. Most real data lie in RDMSs, which have specified schema and meta-information which we can use. RDMSs also incorporate decades of work on optimization, providing better scalability.

## References

- [Foster and Ungar, 2002] Dean Foster and Lyle Ungar. A proposal for learning by ontological leaps. In *Proceedings of Snowbird Learning Conference*, Snowbird, Utah, 2002.
- [Popescul and Ungar, 2003] Alexandrin Popescul and Lyle H. Ungar. Statistical relational learning for link prediction. In *Proc. of the Workshop on Learning Statistical Models from Relational Data at IJCAI-2003*, 2003.
- [Popescul et al., 2002] Alexandrin Popescul, Lyle H. Ungar, Steve Lawrence, and David M. Pennock. Towards structural logistic regression: Combining relational and statistical learning. In *Proc. of the Workshop on Multi-Relational Data Mining at KDD-2002*, 2002.
- [Popescul et al., 2003] Alexandrin Popescul, Lyle H. Ungar, Steve Lawrence, and David M. Pennock. Statistical relational learning for document mining. Computer and Information Sciences, University of Pennsylvania, 2003. <http://www.cis.upenn.edu/~popescul/publications.html>.