Trading Expressivity for Efficiency in Statistical Relational Learning

Ph.D. Thesis Abstract

Niels Landwehr
Department of Computer Science, Katholieke Universiteit Leuven
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
niels.landwehr@cs.kuleuven.be

ABSTRACT

Statistical Relational Learning (SRL) is concerned with building statistical models for relational data. While SRL approaches have shown much potential in complex real-world application domains, their computational complexity remains a major issue and often limits their practical applicability. This thesis is concerned with relatively *simple* yet *efficient* SRL techniques. We show how expressivity and generality can be traded for efficiency by *restricting model complexity* and developing *special-purpose inference and learning algorithms* that take advantage of such restrictions, as well as by *tailoring models to specific application domains*.

1. INTRODUCTION

Statistical relational learning (SRL) combines state-of-the-art statistical modeling with relational representations [2; 1]. It thereby promises to provide effective machine learning techniques for domains that cannot adequately be described using a *propositional* representation, that is, a representation in which examples are described by a fixed set of attributes. Driven by new applications in which data is structured, interrelated, and heterogeneous, this area of machine learning has recently received increasing attention. Example domains for SRL include structure-activity prediction for molecules in bioformatics, classification of web pages based on the surrounding hyperlink structure, or the analysis of data from the interaction of several (human or artificial) agents, as in social networks or multiplayer games. Such domains are visualized in Figure 1.

However, combining statistical modeling and relational representations also poses new challenges. There is a trade-off between the expressivity of a machine learning formalism and its computational efficiency, as a higher expressivity entails a larger search space during inference and learning. Propositional machine learning techniques are at one end of this trade-off, while approaches that combine the full power of statistical and relational learning are at the other end. We present a collection of simple SRL techniques that focus on computational efficiency rather than maximum expressivity, and thereby occupy an intermediate position in the outlined expressivity-efficiency trade-off. Such systems are useful for application domains where more powerful SRL approaches cannot be applied because of their prohibitive computational complexity. Starting from well-established SRL techniques such as propositionalization and knowledge-based model contruction, we have developed efficient SRL approaches for different learning settings, such as (probabilistic) classification and sequence modeling.

2. THREE APPROACHES TO EFFICIENCY IN SRL

The approaches we develop can be roughly grouped into three categories: *dynamic propositionalization* techniques, which improve more traditional (static) propositionalization; *Markov models for relational sequences*, which restrict knowledge-based model construction techniques to a sequential and fully observable setting; and *application-specific SRL approaches*, which gain computational efficiency at the cost of generality by tailoring models to particular application domains.

One of the simplest approaches for solving classification problems in SRL is propositionalization [3]. In propositionalization, the original relational data is mapped to a propositional (attribute-value) representation by employing a set of relational features. Such relational features capture a structural property of the relational examples (such as an aromatic ring structure in a molecule), and represent it as a binary or numeric attribute. Standard statistical machine learning techniques can then be applied on the resulting attribute-value representation. A disadvantage of traditional propositionalization is that the construction of the feature set is decoupled from the statistical modeling, and thus the constructed feature set is typically not optimal for the statistical model that will eventually be used. We introduced *dynamic propositionalization* approaches, in which the construction of the feature set and the statistical modeling are tightly coupled [6; 8]. This often yields more accurate statistical-relational models with smaller feature sets. Computational efficiency can be preserved by integrating feature set construction and the training of the statistical classifier.

Another interesting setting in SRL is learning from sequential data. Such data can in principle be handled using general and powerful SRL frameworks based on knowledge-based model construction. However, these frameworks are not specifically tailored to the sequential case, and their generality comes at a computational price. Instead, we proposed two simple and efficient sequential models by extending the well-known Markov model framework to the relational case. By assuming a Markov property and fully ob-

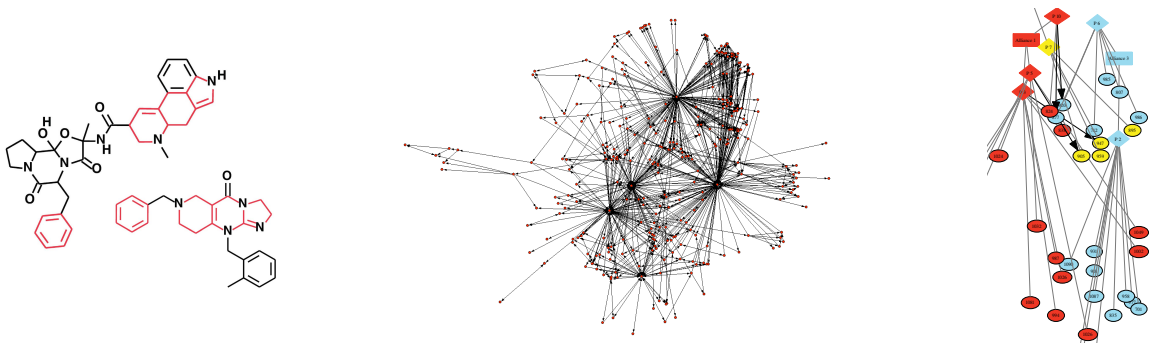

Figure 1: Three example domains for SRL: structure-activity relationship for molecules (left), hyperlink structure of the world wide web (center), graph structure relating players and game artifacts in a massively multiplayer online game (right).

servable data, the resulting learning and inference tasks become significantly easier than in the general case. Sequence elements can be either individual relational facts or complete relational state descriptions. An example domain for the first setting is a log of communication events from a mobile phone user, where a fact $call(outgoing, 233111, busy)$ indicates that the user has tried to reach the phone number 233111 but the line was busy. To deal with this setting we proposed r-grams, which extend simple n-gram models to the relational case [5]. In r-grams, grams can be generalized both by shortening the gram and by relational generalization, that is, by introducing variables as in $call(outgoing, X, busy)$. Variables and unification is used to share information between consecutive sequence elements.

An example domain for the second setting are massively multiplayer online games, in which game states are characterized by several players, game artifacts, and their relationship (see Figure 1). Such games feature complex interactions of players in social networks, and can be seen as models for real-world environments in which humans interact. For this setting, we proposed a simple model that defines a distribution over sequences of relational state descriptions called CPT-L [9], based on the well-known causal probabilistic logic *CP-logic*. In contrast to standard CP-logic, CPT-L is specifically tailored to sequences, and assumes that data is fully observable. This allows for much more efficient inference and learning compared to general CP-logic.

A different way of trading expressivity for efficiency in SRL is to specialize systems to a particular application domain. The general idea is to first model a relational domain in a general-purpose SRL framework, ground the model in the application domain, and then develop special-purpose algorithms for efficient inference and learning in this particular ground model. We explored this approach in two important real-world application domains: population-based haplotype reconstruction from genotype data, and human activity recognition from RFID sensor data. The resulting systems compare favorably to state-of-the-art solutions in terms of effectiveness and efficiency [7; 4].

3. CONCLUSIONS

Statistical relational learning has shown wide application potential, but computational complexity issues limit the pratical applicability of many approaches. In this thesis we have proposed ways of improving computational efficiency in SRL at the cost of expressivity and generality. The resulting systems are applicable to a wide range of significant real-world problems involving structured data and uncertainty.

Ph.D. Dissertation Committee

Prof. Dirk Vandermeulen (chairman), Prof. Luc De Raedt (promoter), Prof. Hendrik Blockeel, Prof. Johan Suykens, Prof. Paolo Frasconi, and Prof. David Page.

---

4. REFERENCES

[1] L. De Raedt. *Logical and Relational Learning*. Springer-Verlag, 2008.

[2] L. Getoor and B. Taskar, editors. *Statistical Relational Learning*. MIT press, 2007.

[3] S. Kramer, N. Lavrac, and P. Flach. Propositionalization Approaches to Relational Data Mining. In *Relational Data Mining*, pages 262–291. Springer, 2001.

[4] N. Landwehr. Modeling interleaved hidden processes. In *Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, June 5-9*, 2008.

[5] N. Landwehr and L. De Raedt. r-grams: Relational Grams. In *Proc. of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India*, 2007.

[6] N. Landwehr, K. Kersting, and L. De Raedt. Integrating Naïve Bayes and FOIL. *Journal of Machine Learning Research*, 8:481–507, 2007.

[7] N. Landwehr, T. Mielikäinen, L. Eronen, H. Toivonen, and H. Mannila. Constrained Hidden Markov Models for Population-based Haplotyping. *BMC Bioinformatics*, 8 (Suppl 2), 2007.

[8] N. Landwehr, A. Passerini, L. De Raedt, and P. Frasconi. kFOIL: Learning Simple Relational Kernels. In *Proceedings of the 21st National Conference on Artificial Intelligence July 16-20, 2006, Boston, MA, USA*, 2006.

[9] I. Thon, N. Landwehr, and L. De Raedt. A Simple Model for Sequences of Relational State Descriptions. In *Proceedings of the 19th European Conference on Machine Learning, Antwerp, Belgium, September 15–19*, 2008.