

Aggregation and Concept Complexity in Relational Learning

Claudia Perlich and Foster Provost

NYU Stern School of Business

44 West 4th Street

New York, NY 10012, U.S.A.

cperlich@stern.nyu.edu, fprovost@stern.nyu.edu

Traditional, flat-table learning algorithms have been scrutinized for many years from many angles, and are well understood with respect to their applicability and expressive power. Much less can be said about relational learning approaches. One main reason is the lack of clarity of the space of possible relational concepts.

We believe that a workshop that brings together people with a variety of perspectives on relational learning represents a unique opportunity to develop a formal characterization of relational concepts. Such formalization would contribute to research on relational learning in a number of ways.

1. It would provide a framework for the theoretical analysis and comparison of relational learners with respect to their ability to express and learn certain concepts.
2. It could thereby provide guidance in the choice of a method for a particular domain.
3. It would be a valuable tool for addressing the issue of model and representation complexity.
4. It would help to clarify the problem domain that a practitioner/research effort is targeting.
5. It would help to specify the scope of generalization of scientific claims.

We argue for a hierarchy of increasingly more general relational concept classes, conveniently placing standard, single-table (“propositional”) learning algorithms as learning concepts in the most specialized class. Some other, more general, cases of relational concepts have been examined already, for example attribute hierarchies (Almuallim et al., 1995). The most general (and complex) class would capture global concepts that include the entire relational structure and all object attributes.

We expect learners developed for more specialized concept classes to apply broadly within the class, but of course to be suboptimal (biased) when a more general class is required. Nevertheless, it may be that the biased learning method actually performs better than an unbiased counterpart, for some problems. This is in direct analogy to what has been observed time and time again in machine learning. For example, linear models (including naïve Bayes) learn a more restricted concept class than does tree induction. However, they are extremely useful. In some

situations, linear models are preferable to tree induction even for problems where the true, underlying model is represented better by a tree (Domingos & Pazzani, 1996; Perlich, Simonoff, & Provost, 2003). Relational learning systems (e.g., ILP systems) often perform suboptimally on purely propositional tasks, even when in principle they are capable of representing the true concept.

Furthermore, given the extreme complexity of higher-level relational concept classes, it is likely that modeling approaches designed for lower levels will have broader expressive power within the lower-level class. For example, ILP systems can represent very complex relational concepts. However, they cannot take full advantage of statistical properties like relational autocorrelation (Jensen and Neville 2000), which can be extremely useful for modeling in relational domains.

A well-designed concept-class hierarchy will also facilitate a “bi-directional search” approach to relational learning research. Although we are not aware of it being framed as such, this type of approach already is being taken. Some researchers are asking how ILP systems can be extended to deal more robustly with more specialized concept classes, and recently there has been a surge of interest in generalizing propositional algorithms (e.g., Bayesian networks, decision trees, logistic regression, and naïve Bayes). A concept-class hierarchy will clarify and to some degree quantify the nature of extensions of propositional learning methods. For example, these extensions of propositional methods typically cannot learn concepts as general as can ILP systems. However, they may be more robust at learning concepts at the lower level. Research results along these lines are much more satisfying than saying “my relational Foo algorithm is better than FOIL on the domains I’ve tested.”

An Aggregation-based framework

Moving upward from propositional algorithms to algorithms that can learn more complex relational concepts, we must address two main issues: (i) how to explore related objects in secondary tables and (ii) how to aggregate the bags of related objects that are found. We conjecture that the formalization of relational concepts must (at least) reflect these two main issues:

- The definition of “relatedness”

- The aggregation of bags of related objects

We have proposed (Perlich and Provost 2003) a first formalization of a relational concept hierarchy that is increasing in the complexity of aggregation with respect to: the number of related objects, the number of object attributes, dependencies between object attributes, number of different object types, and the locality of the concept.

Results using the framework

The paper also reports an empirical study that compares experimentally several different aggregation approaches and assesses the generalization ability with respect to the maximum complexity that the approach could express. In particular we compared (in order of increasing complexity) simple propositional methods without background knowledge, the somewhat more complex case of an abstraction hierarchy without aggregation, simple single-attribute aggregation (most common categorical value, counts, proportions), complex single-attribute aggregation using target-dependent set distances, and logic-based dependent attribute aggregation. The relative performances suggest an optimal aggregation complexity level (target-dependent attribute aggregation) above which the performance decreases.

We conclude from our results so far that the expressive power (and performance) of a relational learning algorithm is strongly related to the aggregation methods applied. Increasing the complexity of the relational aggregation shows significant improvements in the generalization performance. In fact, the best method we tested is a transformation-based approach that uses aggregation to construct a single-table learning task from relational data, and then takes advantage of the sophistication of traditional propositional learners.

Further Implications and Work

Having identified aggregation as a major driver of generalization performance suggests that more efforts should be made to dissect complex relational learning systems into their components and to identify the source of generalization power.

As a research field we should work our way up stepwise through increasing complexity, drawing from the knowledge and experience of lower levels, rather than jumping to very expressive model classes that (currently) suffer from massive search problems.

Additional outstanding tasks are the

- improvement of the hierarchy of relational concepts,
- exploration of “relatedness” and alternative methods of selecting related objects,
- identification of other components of concepts and of algorithms (besides exploration and aggregation),

- undertaking of more comparative studies and acquiring additional benchmark datasets.

Acknowledgements

We are grateful to Jeff Simonoff and Sofus Macskassy for valuable comments and discussions.

This work is sponsored in part by the Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory, Air Force Materiel Command, USAF, under agreement number F30602-01-2-585.

References

- Almuallin, H; Y. Akiba, and S. Kaneda (1995). On handling tree-structured attributes in decision tree learning. *In Proceedings of the Twelfth International Conference on Machine Learning 1995*. 12-20. Morgan Kaufmann.
- Domingos, P. and P. Michael (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning 29(2-3)*, 103-130
- Jensen, D. and J. Neville (2002). Autocorrelation and Linkage Cause Bias in Evaluation of Relational Learners. *In Proceedings of The Twelfth International Conference on Inductive Logic Programming (ILP 2002)*. Springer-Verlag
- Perlich, C. and F. Provost (2003). Aggregation-based Feature Invention and Relational Concept Classes. *In Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining(KDD-2003)*
- Perlich, C., F. Provost, and J. Simonoff (2003). Tree Induction vs. Logistic Regression: A Learning-curve Analysis. *To appear in the Journal of Machine Learning Research*. Preliminary version: *CeDER Working Paper #IS-01-02*, Stern School of Business, New York University