STATEMENT OF RESEARCH INTERESTS

The general focus of my research is on data-driven modeling for business applications. I will use the term "model estimation" and "modeling" to refer to approaches from a variety of fields including statistics, econometrics, machine learning, and data mining. Besides my current technical work, which extends traditional predictive modeling techniques to complex and in particular multi-relational data (ubiquitous in business domains), I am also interested in model estimation and evaluation for financial domains (Reisz and Perlich 2003), text classification for behavioral research, system development for decision support, and comparative studies on the applicability and relative performance of various modeling methods (Perlich et al. 2003).

My dissertation explores the particular task of predictive modeling from multi-relational data. Most business data are stored in multi-relational databases with complex relationships between tables. In order for modeling technologies to be applied successfully, such complex, multi-table data must undergo laborious manual preprocessing, which requires skill and luck to be effective. My thesis work develops a theoretical framework and a prototype system for automating this "relational feature construction". My prototype implements a novel methodology for automated feature construction using density estimates and distances as aggregates (Perlich and Provost 2003) that was derived from a Bayesian modeling perspective. This implementation of a transformation-based framework shows superior generalization performance on a variety of complex and noisy application domains including direct marketing for online retailing, citation-based document classification (Perlich 2003a), medical diagnostics, default prediction for bank loans, customer classification for life insurance, and terrorist identification. Additional theoretical work explores the implications of relational modeling assumptions made by different relational learning approaches on the expressive power of the resulting models and the maximum concept complexity that can be expressed. My large-scale empirical comparison across application domains investigates the apparent tradeoff between the expressive power (size and complexity of the considered model space) and robustness (in terms of generalization performance) to noise.

More technically, my research interests concentrate on foundations and applications of modeling and fall into three interrelated categories. I will first introduce those categories generally and will then discuss my particular interests in more detail.

1. **Applicability and Relative Performance:** One of the most challenging tasks in machine learning has been to characterize under what circumstances one method outperforms another on real domains. The relative performance of model estimation methods is a function of the interactions of domain characteristics and the particular model estimation fundamentals. There is a large body of theoretical work on those interactions in the field of econometrics, judging the quality of an estimator in terms of its bias and variance. In the case of nonparametric model estimation techniques, deriving a strong theoretical bound is often impossible and the relative performance has to be investigated in empirically studies.
2. **Fundamentals:** Foundations of induction are shared across all disciplines involved in model estimation. These foundations include: distance measures, input representations, optimization and search, appropriateness of cost functions, noise and uncertainty, model bias and expressive power, bias-variance tradeoff, and performance and error measures.
3. **System Development and Applications**: Results from categories 1 and 2 can guide the development of learning architectures and strategies for modeling increasingly complex domains, addressing issues of feature construction, choice of representation, integration of multiple modeling approaches, and hybrid models. Applications add further requirements including: usability by domain experts (including a high degree of automation), reliability across tasks, and interpretability of models.

APPLICABILITY AND RELATIVE PERFORMANCE

It is a challenging task to characterize the circumstances under which one method will outperform another on a real domain, and it has become even more pressing as researchers introduce new modeling methods. Not only can this knowledge increase our understanding, by highlighting strengths and weaknesses suggesting further improvements, but it also opens the door for practitioners to choose among the many different model estimation approaches. Theoretical analysis of methods and their underlying assumptions can often provide guidelines (Perlich and Provost 2003b) but are limited with regard to generalization behavior if the underlying assumptions are violated in unknown ways. One important requirement for empirical method comparison is the ability to characterize a task and domain with respect to the driving factors of relative model class performance. Two topics in this category of applicability and relative performance that I have been particularly interested in are the interactions of data representation and model biases and the interactions of task difficulty (inherent uncertainty and noise), training-set size, and the model specific bias-variance tradeoff (Perlich et al. 2003). We have found that linear models outperform tree-based classifiers if there is a high level of inherent uncertainty. Small training-set sizes further enhance this effect. The interaction of model bias and data representation is of significant importance for the task of feature construction. A large part of my dissertation focuses on this issue for multi-relational models. Note that those interactions are often driven by the implicit definition of similarity in a particular model class.

FUNDAMENTALS OF INDUCTION

Similarity: The justification of predictive modeling rests on the assumption that induction is possible, and in particular that objects that are similar will exhibit similar target values. The essence of learning is the identification of properties that make objects similar with respect to a specific target. One important observation is that the notion of similarity depends on the task. In my dissertation, I have focused and on similarities of complex objects in relational databases. The main challenge in a relational domain is to assess similarity with respect to links between objects and sets of related objects. The first step of my work was to develop relational notions of similarity that are general enough to support a general-purpose system architecture, in order to learn a variety of similarities for different relational prediction tasks, and that is at the same time precise enough to capture the particular semantics of the task (for maximal prediction performance). The analysis of ACORA's generalization performance across domains suggests that the flexibility of using a variety of similarity measure is one significant source of reliability. Beyond the scope of my dissertation, I have been interested in graph and time-series similarity measures.

Performance measures for estimation of class-membership probability: The judgment of data mining research is always dependent on the rigor of the evaluation. Besides the rigor of the methodology, the choice of an appropriate performance measure is crucial. The question of performance measures for probability estimation is of particular interest due of the mismatch between the target observations (e.g., binary) and the predictions (continuous [0,1]). Performance measures are relevant not only in the evaluation but also during model estimation to guide the parameter optimization or structural search. It is possible to derive theoretically appropriate cost functions using maximum likelihood from assumptions about the statistical nature of noise. However, little guidance exists if those assumptions are consistently violated. This emphasizes the need for better understanding and better assessment of the true empirical noise. Methodology evaluations in a research setting pose additional requirements. In particular, invariance to specific domain properties such as class priors and scaling allows for a higher degree of abstraction and generalization of findings across domains. I have recently developed a framework that categorizes a variety of performance measures with respect to their statistical properties and assumptions. Most measures focus on the model's ability to decrease variance, whereas other focus on the relative precision of the conditional expected value of the target. I am currently interested in the preferential

selection of models using different performance measures controlling for different statistical models of noise.

Noise and uncertainty: The evaluation of model estimation approaches typically is relative to some lower-bound baseline. However, our ability to model a phenomenon also has an upper bound: the level of inherent uncertainty. From a machine learning perspective, the point of interest is how well the perfect model could possibly perform on a given problem. There are a number of theoretical concepts (i.e., Bayes rate), but it is very difficult to distinguish empirically between a prediction error due to inherent noise or due to a shortcoming of the model. The reason for this difficulty is that the identification of noise requires assumptions either about statistical characteristics of the noise or about the true underlying relationship. I am currently investigating theoretically and empirically how to measure the "true noise" (or degree of inherent uncertainty) for different types of noise. I have reasons to believe that the particular characteristics of the "noise model" interact strongly with model quality across different modeling methodologies. Those interactions are very valuable for the design of automated machine learning methods.

SYSTEM DEVELOPMENT AND APPLICATIONS

System design not only puts many of these concepts under empirical scrutiny, but it adds a number of additional requirements. The design of my dissertation prototype ACORA (Perlich and Provost 2002) follows a number of application-driven guidelines: modular design in order to allow for extension and adaptations to particular domains, a high degree of automation to allow for efficient use by domain experts, support of different degrees of parameter specification to support different user groups. The system demonstrates favorable performance across a number of noisy business domains and has been adopted for behavioral research projects at the Stern School of Business.
In addition, I am collaborating on a number of interdisciplinary research projects in (mostly) financial domains were my particular focus is on modeling and evaluation techniques (e.g., Reisz and Perlich 2003).

REFERENCE:

Perlich, C. and F. Provost. 2002. "A Modular Approach to Relational Data Mining," American Conference on Information Systems (AMCIS) 2002

Perlich, C., F. Provost, and J. Simonoff. 2003. "Tree Induction vs. Logistic Regression: A Learning Curve Analysis," Journal of Machine Learning Research 4 (2003) 211-255

Perlich, C. and F. Provost. 2003a. "Aggregation-Based Feature Invention and Relational Concept Classes," Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003), 167-176

Perlich, C. 2003. "Citation-Based Document Classification," Workshop on Information Technology and Systems (WITS) 2003

Perlich, C. and F. Provost. 2003b. "Aggregation and Concept Complexity in Relational Learning," Workshop on Learning Statistical Models from Relational Data (SRL), at IJCAI 2003

Reisz, S.A. and C. Perlich. 2003. "Temporal Resolution of Uncertainty and Corporate Debt Yields: An Empirical Investigation," Fourth round, Journal of Business