DISSERTATION ABSTRACT

Modern businesses collect and store vast amounts of data on business transactions, accounting, and customers. The relational data format provides the required flexibility to represent different entities and relations between them to capture the reality of complex business domains adequately. However, there is a significant discrepancy between the relational data representation as used for storage and the constraints imposed by common model estimation techniques (e.g. linear regression, tree induction). These constraints in particular limit the input format to a fixed number of independent variables. Data with multiple entity types (e.g. customers and transactions) with one-to-n relationships between them (every customer has a different number of transactions) traditionally require the manual construction of independent attributes using summary statistics and aggregates such as recency, frequency, and average price. This manual process becomes not only increasingly more time consuming as the complexity of the domain increases, but also involves the significant loss of information. In particular categorical fields with a large number of possible values such as product types are often discarded, since there are no adequate aggregation methods that could reduce the dimensionality sufficiently to guarantee model generalization.

My dissertation addresses the task of automating predictive modeling from multi-relational data. The prototype of "Automated Construction of Relational Features," (henceforth ACORA), implements a novel, domain-independent methodology for automated feature construction using class-conditional density estimates and distances for aggregation that are derived from a Bayesian modeling perspective. Conditional density distances are particularly suitable for categorical fields with many possible values, since they compress the entire distribution into a single distance measure. The prototype provides a number of different distance measures including likelihood, cosine, Euclidean, Mahalanobis, and sum of absolute error to match the particular characteristics of an application domain. This data-driven and task-specific feature construction retains the predictive information and constructs dominantly discriminative features. A large-scale empirical evaluation of ACORA's generalization performance on classification and probability estimation tasks shows superior performance over alternative aggregation methods across a variety of complex and noisy application domains including direct marketing for online retailing, citation-based document classification, medical diagnostics, default prediction for bank loans, customer classification on life insurance, and terrorist identification. Additional theoretical work explores the implications of aggregation assumptions made by different feature construction approaches on the expressive power of the resulting models and the maximum concept complexity that can be expressed. The large-scale empirical comparison across domains investigates further the apparent tradeoff between the expressive power (size and complexity of the considered model space) and robustness (in terms of generalization performance) to noise.