Summary 8: Learning Systems of Concepts with an Infinite Relational Model

Susan Cherry

This paper presents a nonparametric Bayesian model that discovers systems of related concepts. It illustrates how simple relational systems can be acquired by unsupervised learning. For example, suppose that a domain includes several types (in medicine, a type could be the set of words that appear in a medical chart). Domain theory attempts to specify the types of entities that exist in a set and the likely relationships between them. The model presented in this paper assumes that each entity belongs exactly to each cluster and simultaneously discovers the clusters and relationships between clusters best supported by the data. This model does not require the number of clusters to be fixed in advanced.

The paper begins by describing the infinite relational model (IRM). To motivate the model, assume that we have a single type in the data, which is people. A relation $likes(i, j)$ indicates whether person i likes person j. The goal of this model is to organize entities into clusters that relate to each other in predictable ways. The IRM is attractive because it can handle arbitrarily complex systems of attributes, entities, and relations. To allow the IRM to discover the number of clusters, the authors suggest a prior that assigns some probability mass to all possible partitions of the data. They choose a distribution over partitions induced by a CRP. Data is generated from the clusters through the following generative mode: $z|\gamma\ CRP(\gamma)$, $\eta(a, b)|\beta\ Beta(\beta, \beta)$, $R(i, j)|z, \eta\ Bernoulli(\eta(z_i, z_j))$. Inference is carried out using MCMC methods to sample from the posterior on cluster assignment. We want to discover the best representation for each data set and we can search for the best partition z by running hill climbing from an initial configuration where a single cluster is used for each type.

Next the paper discusses related work and how the IRM differs from previous models. Compared to much of the previous work, the IRM has the ability to handle arbitrary collections of relations which may each take on any number of arguments. It's a framework that can be applied to data sets with qualitatively different forms. The IRM also has the ability to learn increasingly complex representations as it encounters more data. Finally, most of the previous approaches discover only clusters of objections. The IRM also focuses on clustering predicates. The IRM's attempt to cluster entities, features, and relations is a step towards finding patterns at all levels.

The paper concludes by applying the IRM model to several different models. First, the authors generate synthetic data to explore the IRM's ability to infer the number of clusters in each type. When the data are clean, the IRM accurately recovers the true number of clusters. Performance only begins to suffer when the data becomes extremely noisy, suggesting nonparametric Bayesian methods are useful for relational problems. Next, the authors test clustering objects and features. The IRM is applied to an animal-feature matrix. The authors find that the IRM performed better than the infinite mixture model at sorting the animals into groups. The authors conclude by discussing learning ontologies and learning kinship systems from Australian tribes. Again, the IRM performs better than the IMM. Finally, the text the IRM on a political dataset with multiple types and relations. Once again, the IRM performs well by dividing the countries into reasonable groups.