Integrating Ontological Prior Knowledge into Relational Learning

Stefan Reckow RECKOW@MPIPSYKL.MPG.DE
Max Planck Institute of Psychiatry, Proteomics and Biomarkers, 80804 Munich, Germany

Volker Tresp VOLKER.TRESP@SIEMENS.COM
Siemens AG, Corporate Research & Technology, 81739 Munich, Germany

Abstract

Ontologies represent an important source of prior information which lends itself to the integration into statistical modeling. This paper discusses approaches towards employing ontological knowledge for relational learning. Our analysis is based on the *IHRM* model that performs relational learning by including latent variables that can be interpreted as cluster variables of the entities in the domain. We apply our approach to the modeling of yeast genomic data and demonstrate that the inclusion of ontologies as prior knowledge in relational learning can lead to significantly improved results and to better interpretable clustering structures.

1. Introduction

One of the great challenges in Bayesian learning is to find appropriate ways to include prior knowledge. Traditionally, prior knowledge is introduced via the specification of prior parameter distributions. Despite the wide success of this approach, parameter distributions are often not very intuitive for the domain expert. Furthermore, even in case that the expert has sophisticated prior knowledge, a statistician might find it difficult to formalize this prior knowledge as prior parameter distributions to be included in statistical modeling.

A quite different effort to formalize prior knowledge is practiced in knowledge engineering where typically many experts are asked to agree on an ontology. An ontology is a data model that represents a set of concepts within a domain and the relationships between those concepts. Typical constructs are subclass hierarchies, type constraints for relations and even more sophisticated rule-based constraints. A class can be subclass of one or several parent classes. If an instance is known to belong to a given class it is also a member of all its ancestral classes. This implies that constraints propagate from top to down: if a constraint is true for a class it is also true for all its offspring classes.

It should be clear that an ontology is an invaluable source of information also for machine learning. There might be some statistical dependencies that apply to all objects in a domain and some dependencies which only apply to members of a particular class and all its subclasses. In this paper we analyze how ontological prior knowledge can be integrated into machine learning. Since most domains for which ontologies have been developed are relational we apply ontology supported learning to the recently developed *Infinite Hidden Relational Model (IHRM)*. The IHRM can be considered as relational (soft-)clustering and showed excellent predictive performance in previous experiments.

2. Statistical Relational Learning and the IHRM Model

In statistical relational learning one needs to agree on a language for describing a model. Our preference is the DAPER model (Heckerman et al., 2004), which is based on the entity relationship (ER) model. The ER model has been developed as a graphical representation of a relational database structure. The DAPER model includes directed arcs indicating direct probabilistic dependencies. The *Hidden Relational Model* (HRM) is a particular DAPER model with a uniform dependency structure: For each entity a latent variable is introduced. The latent variables are the parent nodes of all attribute nodes and all relational nodes. The details are best illustrated using a concrete example. In the following sections we illustrate the HRM

Presented as extended abstract at the NIPS SISO workshop, Whistler, Canada, 2008.

in a movie recommendation system. By employing a Dirichlet Process prior, the *Infinite Hidden Relational Model* (IHRM) generalizes the HRM to include an infinite number of states in the latent variables. For more details about HRM and IHRM see (Xu et al., 2006; Kemp et al., 2006).

2.1. Hidden Relational Models

Figure 1 shows the structures for a movie recommendation system. It shows the DAPER model with entity classes User, Movie and relation class Like. In addition there are User Attributes, Movie Attributes and Relation Attributes R. The *ontological concepts* are simply additional Movie Attributes, the details of which are discussed in a later section. Directed arcs indicate direct probabilistic dependencies. In the Hidden Relational Model (HRM), for each entity a latent variable is introduced, in the example denoted as Z^u and Z^m. They can be thought of as unknown attributes of the entities and are the parents of both entity attributes and relationship attributes. The underlying assumption is that if the states of the latent variables were known, both attributes and the relational attribute R can be well predicted.

For the sake of clarity figure 1 omits the parameters and priors, but we will shortly describe them in the following (note that many alternative parameterizations are also possible). Assume that Z^u has K^u states, and that $\pi^u = (\pi^u_1, \ldots, \pi^u_{K^u})$ are multinomial parameters with $P(Z^u = k) = \pi^u_k$ ($\pi^u_k \geq 0, \sum_k \pi^u_k = 1$). The multinomial parameters are drawn from a Dirichlet prior with $\pi^u \sim \mathrm{Dir}(\cdot|\alpha^u_0/K^u, \ldots, \alpha^u_0/K^u)$. In the experiments all user attributes are assumed to be discrete and independent given Z^u. Thus a particular user attribute A^u with S states is a sample from a multinomial distribution with $P(A^u = s|Z^u = k) = \theta^u_{s,k}$ and

$$(\theta^u_{1,k}, \ldots, \theta^u_{S,k}) \sim G^u_0 = \mathrm{Dir}(\cdot|\beta^{u*}_1, \ldots, \beta^{u*}_S).$$

It is also convenient to re-parameterize the Dirichlet parameters as $\beta^u_0 = \sum_{s=1}^{S} \beta^{u*}_s, \beta^u_s = \beta^{u*}_s/\beta^u_0$ for $s = 1, \ldots, S$, and $\beta^u = (\beta^u_1, \ldots, \beta^u_S)$. In the application, we assume a neutral prior with $\beta^u_s = 1/S$, which represents our prior belief that the multinomial parameters should be equal. β^u_0 is a parameter indicating how strongly we believe that the prior distribution should be true. Similarly we can define the parameters for the Movie class and the relationship class Like. Note, that for the relationship attribute R, $K^u \times K^m$ parameter vectors ϕ are generated.

2.2. Infinite Hidden Relational Models

The latent variables can be interpreted as cluster assignments where the number of their states correspond to the number of clusters, which is typically unknown in advance. It thus makes sense to allow an arbitrary number of latent states by using a Dirichlet process mixture model. This permits the model to decide itself about the optimal number of clusters. For our discussion it is sufficient to say that we obtain an IHRM model by simply letting the number of states, K^u and K^m, approach infinity. Although a model with infinite numbers of states and parameters cannot be represented, it has been shown that sampling in such model is elegant and simple. A single parameter α_0 is known to determine the tendency to either use a large or small number of states in the latent variables. Learning and inference in the IHRM is based on a Gibbs sampler using the Chinese Restaurant Process. For more details, please consult (Xu et al., 2006; Kemp et al., 2006).

3. Integrating Ontological Prior Knowledge into the IHRM

The lower part of figure 1 shows the inclusion of an imaginary movie ontology with boolean concept variables B_k. Let $par(B_k)$ denote the set of parent concepts of B_k. Also let $par(B_k) = 1$ stand for the fact that all parents of B_k have state equal to 1 and $par(B_k) \neq 1$ indicate that at least one of the parents is in state equal to 0.

In databases, annotation of an item with respect to a given attribute is typically as specific as possible. GO annotations (Gene Ontology Consortium, 2006) for example refer to the most specific biological processes a gene participates in, since the ontological structure provides implicit annotation with all parent concepts. Making those explicit is what we call *ontological enhancement* of the data. A simple preprocessing of the data ensures that

$$\forall B_k : B_k = 1 \Rightarrow par(B_k) = 1.$$

This enriched feature representation reflects the full information from the ontology.

In the following we present two approaches that differ in the way, dependencies among the ontological concepts are handled in the model. In our first approach we simply treat all ontological concepts as additional independent attributes of the corresponding entity as depicted in figure 1(a). We assume that $P(B_k = 1|Z^m = l) = \xi_{k,l}$, $P(B_k = 0|Z^m = l) = 1 - \xi_{k,l}$

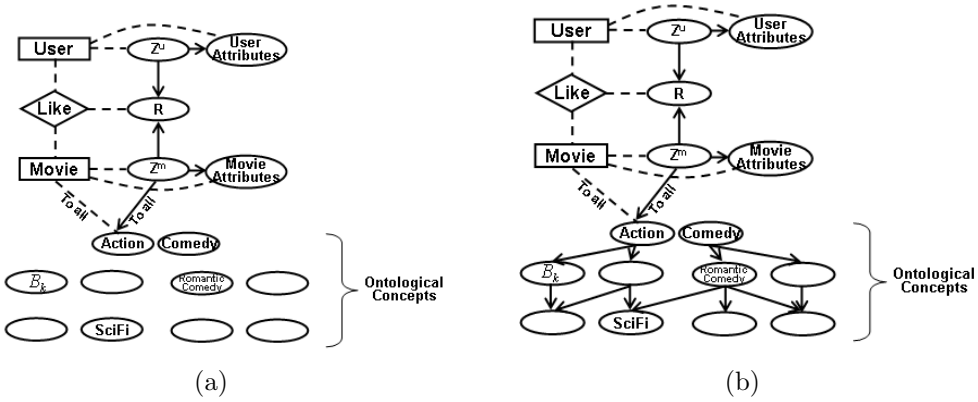


Figure 1. The IHRM model for the recommendation example integrating a hypothetical movie ontology.

and thus the dependencies can be modeled as

$$P(\{B_k\}|Z^m = l) = \prod_{k|B_k=1} \xi_{k,l} \prod_{k|B_k=0} (1 - \xi_{k,l}).$$

Treating concepts as independent ignores the constraints that our ontological enhancement imposed on the attributes. Our second approach takes these constraints into account by defining that $P(B_k = 1|par(B_k) \neq 1, Z^m = l) = 0$. Now we have that $P(B_k = 1|par(B_k) = 1, Z^m = l) = \xi_{k,l}$, $P(B_k = 0|par(B_k) = 1, Z^m = l) = 1 - \xi_{k,l}$ and the probability for a concept pattern breaking the ontological constraint is equal to zero. The resulting dependency structure is shown in figure 1(b). Now we have

$$P(\{B_k\}|Z^m = l) = \prod_{k|B_k=1, par(B_k)=1} \xi_{k,l} \prod_{k|B_k=0, par(B_k)=1} (1 - \xi_{k,l}).$$

Note, that the only difference is that if the ontological constraints are obeyed, the ontological concepts whose parent concepts are not all equal to one drop out of the equations.

4. Experiments on Genomic Data

For the experiments we used 1000 genes from the *Comprehensive Yeast Genome Database (CYGD)* (Guldener et al., 2005). The genes were randomly selected out of the set of all genes/proteins having known interactions in the *Database of Interacting Proteins (DIP)* (Xenarios et al., 2000) and at least one known annotation for the *complex* feature. Additional features included *chromosome*, *structural class*, *phenotype* and *function*. *Interaction* is treated as the binary relational attribute linking pairs of genes.

4.1. Ontological information

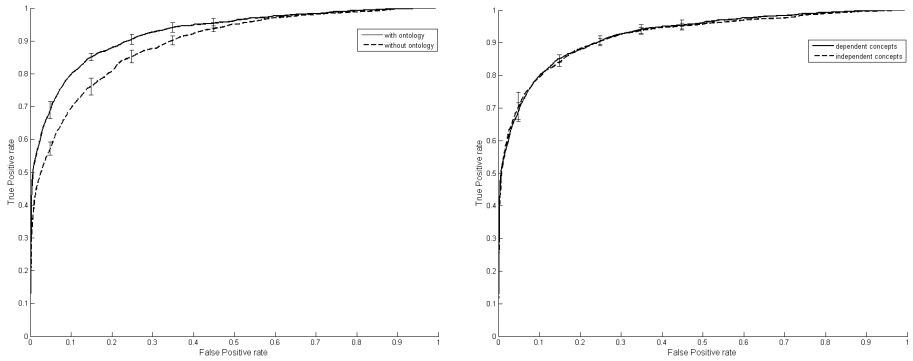
The *complex* annotation scheme is an ontologically organized set of attributes refering to molecular complexes a protein may form with others to perform certain higher order tasks. It is hierarchically structured from quite general complexes to more specific ones on 5 levels. The top level comprises 69 different concepts. Annotations of this feature are very sparse, so we chose our data set in a way, that every protein has at least 1 complex annotation.

4.2. Clustering

The beneficial effect of the ontological enhancement becomes apparent in the clustering of the genes. By enriching the information some drawbacks of the Dirichlet process mixture model can be diminished. In all our IHRM experiments without ontological information we observed the formation of one single extraordinarily big cluster, which obviously collects all data points which are too similar to be separated. Additionally there appear many singleton clusters containing a single element which may be too different from the rest to be assigned to any populated cluster. When using the ontological information during clustering, the size of the huge cluster reduces significantly, distributing the genes to other existing clusters and all singeltons disappear. The ontology thus helps to assign the items to suitable clusters.

4.3. Predictive Performance

To investigate the effects of the improved clustering we evaluated the IHRM's performance in predicting the *function* attribute from the remaining features. We conducted 5-fold cross-validation and plotted the averaged ROC-curves to visualize classifier performance. ROC-curves were averaged vertically, which gives rise



(a) (b)

Figure 2. ROC curves visualizing classification performance.

to confidence intervals at fixed classification thresholds.

At first we investigated the benefits of the complex ontology assuming independent ontological concepts, given the state of the latent variable. Model learning was performed twice and differed in the usage of the *complex* attribute in the following way. First, we only used the *complex* annotations corresponding to the lowest (i.e. most specific) class in the ontology neglecting the ontological information about the parents. In contrast, for the second run we added to each assigned *complex* feature the annotations for all parents of that feature and their parents, respectively. By the ontological enhancement, the features got more expressive and therefore more helpful for clustering. Note, that we did not take dependencies among the concepts into account, yet and handled every feature independently. The results of the two experiments can be seen in figure 2 (a). The classifier, using ontological information clearly outperforms the other in the critical region near the upper left corner. The error bars denote 95% confidence intervals at selected thresholds.

The next experiment examined, how modeling of the dependencies within the ontological concepts affects prediction. An experiment with the same setting was performed, but dependencies between the ontological concepts were explicitely modeled as described in section 3. In this experiment, however, we couldn't see an improvement over the independent concepts. In Figure 2 (b) it is made clear, that the two classifiers show no mentionable difference. We suspect, that this is due to the extreme sparsity of the complex attribute, which prevents the dependency modeling to have a wide effect.

Additional experiments, where we let α range from 1 to 100 proved the final clustering to be quite stable. The insensitivity to variations of this parameter, which controls the number of clusters, indicates that the model is robust when there is a true underlying cluster structure that can be discovered.

5. Conclusions

We have developed a concept for integrating domain ontologies as prior knowledge into relational machine learning. Using a genomic data set we have shown that the integration of the ontology has lead to more meaningful clustering structures and to better predictive performance. We expect that a growing number of domain ontologies will be developed in the future and it seems quite useful to integrate them in machine learning problems. A current area of interest concerns the integration of medical ontologies into learning-based medical decision support systems.

References

Gene Ontology Consortium (2006). The gene ontology (go) project in 2006. *Nucleic Acids Research, 34, Database Issue*, D322–D326.

Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N., van Helden, J., Lemer, C., Richelles, J., Wodak, S. J., Garcia-Martinez, J., Perez-Ortin, J. E., Michael, H., Kaps, A., Talla, E., Dujon, B., Andre, B., Souciet, J. L., De Montigny, J., Bon, E., Gaillardin, C., & Mewes, H. W. (2005). Cygd: the comprehensive yeast genome database. *Nucleic Acids Research, 33*, D364–D368.

Heckerman, D., Meek, C., & Koller, D. (2004). *Probabilistic models for relational data* (Technical Report MSR-TR-2004-30). Microsoft.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada,

T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., & Eisenberg, D. (2000). Dip: the database of interacting proteins. *Nucleic Acids Res*, *28*, 289–291.

Xu, Z., Tresp, V., Yu, K., & Kriegel, H.-P. (2006). Infinite hidden relational models. *Proc. 22nd UAI*.