Structural Knowledge Discovery in Chemical and Spatio-Temporal Databases

Ravindra N. Chittimoori, Jesus A. Gonzalez and Lawrence B. Holder
Department of Computer Science and Engineering
University of Texas at Arlington
Box 19015, Arlington, TX 76019-0015
{chittimo,gonzalez,holder}@cse.uta.edu

Most current knowledge discovery systems use only attribute-value information. But relational information between objects is also important to the knowledge hidden in today's databases. Two such domains are chemical structures and domains where objects are related in space and time. Inductive Logic Programming (ILP) discovery systems handle relational data, but require data to be expressed as a subset of first-order logic. We are investigating the application of the graph-based relational discovery system SUBDUE (Cook, Holder, & Djoko 1996) in structural domains. Input to SUBDUE is a graph with labeled vertices and directed or undirected labeled edges. SUBDUE performs a beam search of the space of all possible subgraphs of the input graph. The search is guided by the minimum description length (MDL) principle, looking for subgraphs (substructures) with many instances that can be used to compress the original data and represent structural knowledge.

We applied SUBDUE to the task of identifying structural patterns that distinguish carcinogenic and non-carcinogenic chemical compounds available from the National Toxicology Program (ntp-server.niehs.nih.gov). Each atom in a compound is represented as a vertex with directed edges to other vertices, where the edge labels specify whether the vertex is the atom name, type or partial charge. Bonds between atoms are represented as undirected edges between the vertices. We divided the data into a training set (268 compounds) and a testing set (30 compounds).

SUBDUE found a substructure containing a bromine atom that occurred in 134 of the 143 carcinogenic training compounds and in only 24 of the 125 noncarcinogenic training compounds. This same substructure was found in 15 of the 19 carcinogenic testing compounds and in only 4 of the 11 noncarcinogenic testing compounds. The results are similar to those of ILP systems like PROGOL (Srinivasan *et al.* 1997). We are experimenting with a new concept-learning version of SUBDUE that finds substructures compressing the positive data without compressing the negative data. Preliminary results show that the new version is competitive with the predominantly concept-learning ILP systems.

We have applied SUBDUE to two spatio-temporal domains: the Aviation Safety Reporting System (ASRS) database (olias.arc.nasa.gov/ASRS) and the Earthquake database (wwwneic.cr.usgs.gov). The ASRS database consists of a set of reports containing 74 fields describing an incident that might affect aviation safety. Each record in the earthquake database consists of 35 fields describing the seismic event. In both databases an event is represented by a vertex with attribute-labeled directed edges to vertices labeled with that attribute's value. The data was augmented with *near_in_distance* and *near_in_time* relational edges between such events. We empirically selected the distance and time thresholds to not overload the graph with spatio-temporal information, but to still bias SUBDUE during the search.

In the ASRS domain SUBDUE found a substructure relating similar events of type *damage* using the *near_in_distance* relation, suggesting that such incidents are localized and recommending further investigation in that region. In the earthquake database SUBDUE found a substructure relating two earthquakes, whose epicenters were at the same depth of 33km, using a *near_in_distance* relation. Our collaborator, Dr. Burke Burkart of the UT Arlington geology department, says this pattern suggests a fault in the area.

Experimental results show that SUBDUE is able to discovery relevant structural knowledge in a graphical representation of real-world structural domains like chemical toxicity and in databases augmented with spatio-temporal relations like the ASRS and earthquake databases. We will continue analysis of these and other structural domains, comparing performance to competing ILP systems and domain-specific approaches. Source code for the SUBDUE system is available at http://cygnus.uta.edu/subdue.

References

Cook, D. J.; Holder, L. B.; and Djoko, S. 1996. Scalable discovery of informative structural concepts using domain knowledge. *IEEE Expert* 11(5).

Srinivasan, A.; King, R. D.; Muggleton, S. H.; and Sternberg, M. J. E. 1997. Carcinogenesis predictions using ILP. In *Proceedings of the Seventh International Conference on Inductive Logic Programming*, 273–288.