Graph-Based Hierarchical Conceptual Clustering in Structural Databases

Istvan Jonyer, Lawrence B. Holder and Diane J. Cook
University of Texas at Arlington
Department of Computer Science and Engineering
Box 19015, Arlington, TX 76019-0015
{jonyer|holder|cook}@cse.uta.edu
URL: http://cygnus.uta.edu/subdue/clustering

Introduction

Cluster analysis has been studied and developed in many areas for a wide variety of applications. The purpose of applying clustering to a database is to gain better understanding of the data, in many cases through revealing hierarchical topologies. We are working on extending the Subdue structural knowledge discovery system with clustering functionalities. Past works related to ours are an incremental approach called Cobweb [Fisher 1987], and its extension, Labyrinth [Thompson & Langley 1991], that can represent structured objects using a probabilistic model.

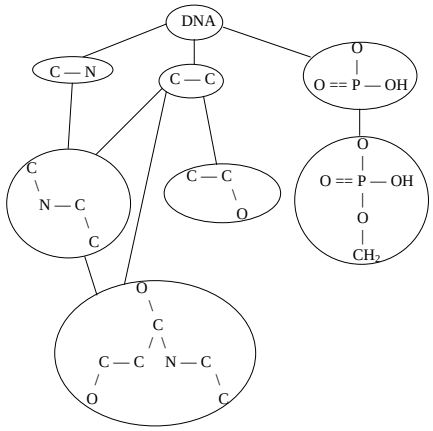
Conceptual Clustering Using Subdue

Subdue [Holder and Cook 1993] is a knowledge discovery system that can deal with structured data by working on their graph representation. This includes vertex and edge labels, as well as directed and undirected edges, where objects and data usually map to vertices, and relationships and attributes map to edges.

Subdue's discovery algorithm discovers interesting, repetitive substructures in the input graph, which is used by our new Graph-Based Hierarchical Conceptual Clustering (GBHCC) algorithm. This algorithm builds a *classification lattice*, versus a tree suggested by other work. We have found that in structured domains the strict tree representation is inadequate. GBHCC begins with an empty lattice and calls Subdue to find a substructure S that maximally compresses the input graph G according to our Minimum Description Length heuristic. If S achieves some compression of G, then S is added to the lattice and used to compress the graph G. S becomes the definition of a cluster. The compressed graph is passed again to Subdue to find another substructure. This iterative approach on successively more compressed graphs allows Subdue to find new substructures defined in terms of previously discovered substructures. Therefore, when substructures are added to the lattice, their parents may include other, non-root nodes in the lattice. This approach allows the discovery of conceptual clustering hierarchies of the database.

To illustrate Subdue's strength–the ability to work with structured data–we present a task that involves describing a DNA sequence by clustering. To represent the DNA as a graph, atoms and small molecules are mapped to vertices, and bonds are represented by undirected edges. The edges are labeled according to the type of bond, single or double. A portion of the lattice generated is shown in the figure.

The lattice closely resembles a tree, with the exception that two nodes (bottom-left) have two parents. The lattice describes 71% of the DNA sequence. As the figure shows, smaller, more commonly occurring compounds are found first that compose the first level of the lattice. These account for more than 61% of the DNA. Subsequently identified clusters are based on these smaller clusters that are either combined with each other, or with other atoms or molecules to form a new cluster. The second level of the lattice extends the conceptual clustering description such that an additional 7% of the DNA is covered. Future work on Subdue will continue discovery of hierarchical clusterings in real-world domains, and both objective and expert-based comparisons to other clustering systems.



References

Fisher, D. H. Knowledge Acquisition Via Incremental Conceptual Clustering, *Machine Learning.* Kluwer, The Netherlands, 1987.

Holder, L. B. and D. J. Cook. Discovery of Inexact Concepts from Structural Data. In *IEEE Transactions on Knowledge and Data Engineering*, Volume 5, Number 6, 1993, pages 992-994.

Thompson, K. and P. Langley. Concept formation in structured domains. In Fisher, D.H., & Pazzani, M. (Eds.), *Concept Formation: Knowledge and Experience in Unsupervised Learning,* chap. 5. Morgan Kaufmann Publishers, Inc., 1991.