Substructure Discovery Using Minimum Description Length Principle and Background Knowledge

Surnjani Djoko

Department of Computer Science and Engineering
University of Texas at Arlington
Box 19015, Arlington, TX 76019
djoko@cse.uta.edu

Abstract

Discovering conceptually interesting and repetitive substructures in a structural data improves the ability to interpret and compress the data. The substructures are evaluated by their ability to describe and compress the original data set using the domain's background knowledge and the minimum description length (MDL) of the data. Once discovered, the substructure concept is used to simplify the data by replacing instances of the substructure with a pointer to the newly discovered concept. The discovered substructure concepts allow abstraction over detailed structure in the original data. Iteration of the substructure discovery and replacement process constructs a hierarchical description of the structural data in terms of the discovered substructures. This hierarchy provides varying levels of interpretation that can be accessed based on the goals of the data analysis.

The structural data is represented as a labeled graph. A substructure is a connected subgraph within the graphical representation. An instance of a substructure in an input graph is a set of vertices and edges from the input graph that match, graph theoretically, to the graphical representation of the substructure. The substructures are evaluated by their ability to describe and compress the original data set using the domain's background knowledge and the minimum description length (MDL) of the data. Once interesting substructures are discovered, they can be replaced by a single representative node in the original graph, and can be used as part of another substructure definition in a hierarchy of discovered structures.

The minimum description length principle states that the best theory to describe a set of data is the theory which minimizes the description length of the entire data set. The minimum description length of a graph is defined to be the number of bits necessary to completely describe the graph. The theory that best accounts for a collection of data is the one that minimizes I(S) + I(G|S), where S is the discovered substructure, G is the input graph, I(S) is the number of bits required to encode the discovered substructure, and I(G|S) is the number of bits required to encode the input graph G with respect to S.

Although the principle of minimum description length is useful for discovering substructures that maximize compression of the data, scientists often employ knowledge or assumptions of a specific domain to the discovery process. To make the discovery process more powerful across a wide variety of domains, the background knowledge have been added to guide the discovery process. This background knowledge is entered in the form of rules for evaluating substructures. Because only the most-favored substructures are kept and expanded, these rules control the discovery process of the system.

For example, in the CAD circuit domain, circuit components can be classified according to their passivity. A component which is not passive is said to be active. The active components are the main driving components. Identifying the active components is the first step in understanding the main function of the circuit. The component rule assigns relatively higher values to the active components, and assigns lower values to the passive components. Once the active components are selected, attention can be focused on the passive components. Similarly, the loop analysis rule favors subcircuits containing loops. Since the components in the closed path are generally a part of the subcircuit or the subcircuit itself. Furthermore, the component complexity rule prefers minimum number of distinct component in the substructure.

The approch has also been applied to the domains of chemical compound analysis, scene analysis, CAD circuit analysis, and analysis of artificially-generated graphs. The results demonstrate the applicability and significance of the approcah in the above domains.

References

J. R. Quinlan and R. L. Rivest. Inferring decision trees using the minimum description length principle. *Information and Computation*, 80:227-248, 1989.

P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman. Autoclass: A bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, 54-64, 1988.