Discovering Substructures in the Chemical Toxicity Domain

Ravindra N. Chittimoori, Lawrence B. Holder and Diane J. Cook

Department of Computer Science Engineering
University of Texas at Arlington
Box 19015 (416 Yates St.), Arlington, TX 76019-0015
Email: {chittimo, holder, cook}@cse.uta.edu
Phone: (817)-272-2596
Fax: (817)-272-3784

Introduction

The researcher's ability to interpret the data and discover interesting patterns within the data is of great importance as it helps in obtaining relevant SARs [Srinivasan et al.], for the cause of chemical cancers (e.g., Progol identified a primary amine group as a relevant SAR for the cause of chemical cancers [Srinivasan et al. 1997]). One method for interpreting and discovering interesting patterns in the data is the identification of common substructures within the data. These substructures should be capable of compressing the data and identifying conceptually interesting substructures that enhance the interpretation of data. This identification also helps in simplifying the data by replacing instances of the substructure with a pointer to the newly discovered substructure. The subsequent iterations of the discovery and replacement process construct a hierarchical description of the structural data in terms of discovered substructures.

Discovering substructures for identifying relevant SARs has been a prominent area of application for knowledge discovery systems e.g., the Progol study. In this research we are using the Subdue system [Cook and Holder 1994] to identify SARs in chemical data. Subdue discovers interesting substructures in structural data based on the Minimum Description Length (MDL) principle [Rissanen 1989]. Subdue discovers substructures that compress the original data and represent structural concepts in data.

This paper is organized as follows. The next section explains in detail the current domain (chemical toxicity). The next section discusses the mechanism by which the knowledge discovery system Subdue extracts molecular descriptions for attaining relevant SARs. The methodologies used by the domain specialist to represent the data and the preliminary results are discussed in brief. The final sections talk about the conclusions and the future work in this area.

Chemical Toxicity Domain

The datasets in the NTP database contain information about more than 300 chemical compounds that are either carcinogenic or noncarcinogenic. Primarily there are 298 chemical compounds whose carcinogenicity is known. This comprises the training set of the Subdue program. The training set is further divided to provide learning to Subdue. There are 69 compounds whose carcinogenicity is not known. This comprises the experimental set of the Subdue program. The information in these sets relates to the molecular structures of the compounds, and includes the atoms, bonds and domain specific knowledge about various groups like alcohol, amine, amino, benzene, ester, ether, ketone, methanol, methyl, nitro, phenol and sulfide. The representation also contains information about the compound test results (+/-) on the various properties of carcinogenicity like Ames test, Chromex, Chromaberr, Drosophilia, Mouse-Lymph, Salmonella Assay. The aim of this research project is to obtain SARs despite the diversity present among the compounds.

Overview of SUBDUE

The Subdue system discovers the substructures in the databases that compress the original data and represent structural concepts in the data. The best substructure is found after multiple passes by replacing the previously discovered substructures in each pass. A substructure is a connected subgraph within the graphical representation. The discovery system represents structural data as a labeled graph. Objects in the data map to vertices or small subgraphs in the graph, and relationships between objects map to directed or undirected edges in the graph. This graphical representation serves as input to Subdue (e.g., see Figure 1). The algorithm begins with the substructure matching a single vertex in the graph. The algorithm selects the best substructure in each iteration and incrementally expands the instances of a substructure. 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. These new substructures become candidates for further expansion. This algorithm searches for the best substructure until all possible substructures have been considered or the total amount of computation exceeds a given limit. Evaluation of each substructure is determined by how well the substructure compresses the description length of the concerned database.

Methodology

The training set is further divided into learning set and testing set. The learning set comprises approximately 90% of the 298 chemical compounds whose carcinogenicity is known. The learning set is further subdivided into positive and negative examples. Subdue is applied to the positive (cancerous) and the negative (non-cancerous) examples separately and the best substructures are identified in each of these training sets. The resultant best substructures from each of the two training sets (positive and negative) are compared. The substructures that occur in the positive examples but not in the negative examples are identified. These identified substructures are used as the pattern indicative of cancerous activity. This learning of Subdue is applied on the testing set, which contains compounds whose carcinogenicity is known, and the results are compared.

The toxicity of the chemicals in the experimental set can be determined by the following approaches. One approach is to apply Subdue individually to the compounds in the experimental set and record the best substructure in each of the compounds. Based on the judgement of the domain specialist (comparing the best substructure returned by Subdue with the substructures identified from the training set) the compound in the experimental set is determined to be carcinogenic or noncarcinogenic. Presently we are using the approach mentioned above in identifying the carcinogenicity of compounds in the experimental set.

The second approach is to include substructures identified in the training set as predefined substructures for Subdue in its search on the experimental set. Subdue will first search the input graph of the compound for instances of the predefined substructures, using inexact graph matches. Instances that match within the inexact match threshold are subsequently expanded. The domain specialist determines the carcinogenicity or noncarcinogenicity of a compound in the experimental set depending on how well the predefined substructure helped in compressing the description length of the compound.

The third approach is to check if the discovered substructure SAR appears anywhere in the compound to be classified. Once unique SARs are discovered, the presence of only one substructure might be enough evidence to predict carcinogenicity.

The input to the Subdue program is the graphical representation of all the chemical compounds. Each of the atoms in a compound is represented as a vertex and the bonds between the atoms are represented as undirected edges between the vertices. Domain knowledge is incorporated into Subdue to guide the discovery process. Various groups like methyl, benzene, amino etc., each are represented as a vertex in each compound and have directed edges to all the atoms in a compound, which participate in the group. Properties like Salmonella assay and Ames test are each represented as a vertex in each compound and have directed edges to all the atoms in the compound with the string label on each of these edges specifying whether the compound tested positive or negative on this property. Figure 1 below shows a sample graph. To capture the diversity present in the atoms (atom name, atom type, and partial charge), each of the atoms is represented as a separate node with directed edges to the name of the atom (n), type (t) and partial charge (p).

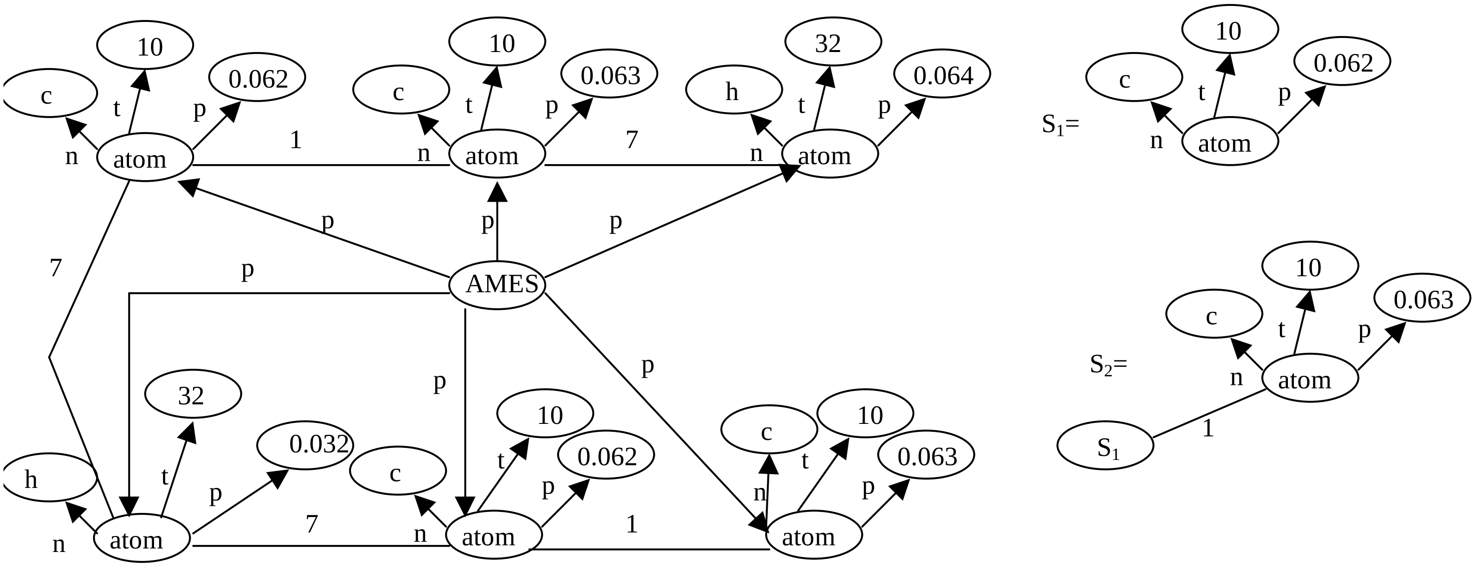

Figure 1: Results of Subdue on part of chemical compound

Subdue discovers substructure S_1 in the compound. S_1 when used to compress the sample graph further finds substructure S_2. Subdue generates a similar hierarchical description of structures with such repeated applications. Subdue can be directed towards giving less importance to certain substructures by specifying the appropriate parameters.

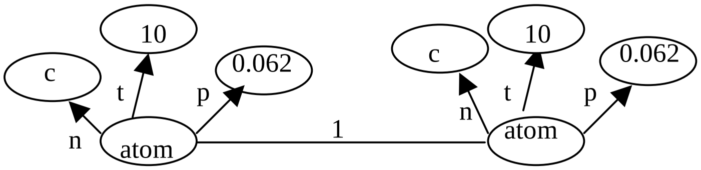

Figure 2: Substructure S_3 .

Results

Subdue has been successful in discovering small substructures. Subdue discovered a substructure S_3 with 8 vertices as in figure 2. The above substructure was discovered by Subdue in 134 of 143 compounds which are positive on the cancerous activity and only in 24 of the 125 compounds which are negative on the cancerous activity in the learning set. Based on this learning, Subdue was applied on testing set. Subdue discovered the same substructure in 15 of the 19 compounds which are positive and 4 of the 11 compounds which are negative in the testing set. Efforts are being made in guiding Subdue to discover more complex substructures that might help in relating a compound with carcinogenicity. The usefulness of applying approaches two and three mentioned in the methodology section is currently under research. Substructures discovered using approach one on the training set are included as predefined substructures for Subdue. We are optimistic that relevant SARs that indicate carcinogenic activity can be identified by Subdue.

Conclusions

The prediction of carcinogenicity and the modeling of diverse chemical compounds is of unquestionable importance. The data mining algorithms capable of handling the increasing structural component of today's databases can achieve this. Subdue, a data mining algorithm, is specifically designed to discover interesting and repetitive patterns within the data that relates molecular structure to cancerous activity.

In this paper, the methodologies of representing the chemical toxicity domain are discussed at length. Subdue is presently in an experimental phase. The initial results of Subdue are explained, and an effort is made to explain the eventual capability of Subdue to discover a pattern that distinguishes carcinogenic and noncarcinogenic compounds.

Future research aims at describing the possible relationships between molecular structure of a compound on the one hand, and biological and toxicological processes on the other. Making use of parallel and distributed resources can significantly improve the run-time performance of data-intensive and compute-intensive discovery programs such as Subdue. We are currently evaluating the benefits of applying a parallel version [Galal, Cook and Holder 1997] of Subdue on the chemical toxicity domain.

References

[Srinivasan et al.] A.Srinivasan, S.H.Muggleton, R.D.King and M.J.E.Sternberg. The Predictive Toxicology Evaluation Challenge. http://www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/

[Cook and Holder 1994] D.J.Cook and L.B.Holder. Substructure Discovery Using Minimum Description Length and Background Knowledge. In *Journal of* Artificial *Intelligence Research,* Volume 1, pages 231-255, 1994.

[Rissanen 1989] J.Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company, 1989.

[Srinivasan et al. 1997] A. Srinivasan, R.D. King, S.H. Muggleton and M.J.E. Sternberg. Carcinogenesis predictions using ILP. In the *Proceedings of the Seventh International Workshop on Inductive Logic Programming,* pages 273-287. 1997.

[Galal, Cook and Holder 1997] G.Galal, D.J.Cook and L.B.Holder, Improving Scalability in a Knowledge Discovery System by Exploiting Parallelism. In the *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining,* pages 171-174. 1997.