Finding frequent substructures in chemical compounds

Luc Dehaspe
Dept. of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
B-3001 Heverlee, Belgium
Luc.Dehaspe@cs.kuleuven.ac.be

Hannu Toivonen
Rolf Nevanlinna Institute &
Dept. of Computer Science
P.O. Box 4, FIN-00014
University of Helsinki, Finland
Hannu.Toivonen@rni.helsinki.fi

Ross Donald King
Dept. of Computer Science
The Univ. of Wales, Aberystwyth
Penglais, Aberystwyth, Ceredigion
SY23 3DB, Wales, United Kingdom
rdk@aber.ac.uk

Abstract

In this paper we apply data mining to the problem of predicting chemical carcinogenicity. We discover queries in first order logic that succeed with respect to a sufficient number of examples; the goal being to find common substructures and properties in chemical compounds, and in this way to contribute to scientific insight. This approach contrasts with previous machine learning research on this problem, which has mainly concentrated on predicting the toxicity of unknown chemicals. Our contribution to the field of data mining is the ability to discover useful frequent patterns that are beyond the complexity of association rules or their known variants. Background knowledge has an essential role here, unlike in most data mining settings for the discovery of frequent patterns.

Introduction

The toxicology evaluation problem. In this paper we apply a data mining method to the problem of predicting whether chemical compounds are carcinogenic or not. A large percentage of cancers are linked to environmental factors such as exposure to carcinogenic chemicals (estimated as high as 80%). Very few compounds have been fully tested for carcinogenesis as the process is very expensive and time consuming. Better computer-based methods are therefore valuable.

The National Toxicity Program of the U.S. National Institute for Environmental Health Sciences conducts standardized bioassays of chemicals on rodents, in order to estimate their carcinogenetic effects on humans. Assays of published but untested chemicals will be completed by Predictive Toxicology Evaluation PTE project (Bristol, Wachsman, & Greenwell 1996) during this year. These cases offer a possibility for true blind trials in carcinogenicity prediction research. The prediction of rodent chemical carcinogenesis was launched at IJCAI '97 as a research challenge for artificial intelligence (Srinisavan *et al.* 1997).

Rather than competing with expert chemists in classifying chemicals to carcinogenic or otherwise, our goal was to discover frequent patterns that would aid chemists – and data miners seeking predictive theories – to identify useful substructures for carcinogenicity research, and so contribute to the scientific insight. This can be contrasted with previous machine learning research in this application, which has mainly concentrated on predicting the toxicity of unknown chemicals (Srinisavan *et al.* 1997; Kramer, Pfahringer, & Helma 1997). We believe that a repository of frequent substructures and their frequencies would be valuable for chemical (machine learning) research.

Contribution to data mining. The task of discovering recurrent patterns has been studied in a variety of data mining settings. In its simplest form, known from association rule mining (Agrawal *et al.* 1996), the task is is to find all frequent itemsets, i.e., to list all combinations of items that are found in a sufficient number of examples. A prototypical application example is in market basket analysis: find out which products tend to be sold together.

Our contribution to the field of data mining is in considering the discovery of useful frequent patterns that are far more complex than association rules or their known variants. We discover queries in first-order logic that succeed with respect to a sufficient number of examples. Such patterns are out of the reach of simple transformations to frequent itemsets. We present an attempt for knowledge discovery in structured data, where patterns reflect the one-to-many and many-to-many relationships of several tables. Background knowledge, represented in a uniform manner, has an essential role here, unlike in most data mining settings for the discovery of frequent patterns.

Datalog concepts. We use DATALOG to represent both data and patterns. In DATALOG, a *term* is defined as a constant symbol, written in lowercase, or a variable, written with initial uppercase. A *logical atom* is an m-ary predicate symbol followed by a bracketed m-tuple of terms. A *definite clause* is a universally quantified formula of the form $B \leftarrow A_1, \ldots, A_n$ $(n \geq 0)$, where B

and the A_i are logical atoms. This formula can be read as "B if A_1 and ... and A_n". If $n = 0$, a definite clause is also called a *fact*. A (deductive) DATALOG *database* is a set of definite clauses. A formula $\leftarrow A_1, \ldots, A_n$ without a conclusion part is called a *denial*. Such a formula can also be viewed as a (PROLOG) *query* ?- $A_1, \ldots, A_n$: (the resolution based derivation of) the answer to a given query with variables $(X_1, \ldots, X_m)$ binds these variables to terms $(a_1, \ldots, a_m)$, such that the query succeeds if each X_i is replaced by a_i. This binding is denoted by $(X_1/a_1, \ldots, X_m/a_m)$. Due to the nondeterministic nature of the computation of answers, a single query Q may result in many bindings. We will refer by $answerset(Q, D)$ to the set of all bindings obtained by submitting query Q to a DATALOG database D.

Data and background knowledge. The DATALOG database for the carcinogenesis problem was taken from *http://www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/*. The set we have used contains 337 compounds, 182 (54%) of which have been classified as carcinogenic and the remaining 155 (46%) otherwise.

Each compound is basically described as a set of atoms and their bond connectivities, as proposed in (King *et al.* 1996). The atoms of a compound are represented as DATALOG facts such as *atom(d1,d1_25,h,1,0.327)* stating that compound *d1* contains atom *d1_25* of element *h* and type *1* with partial charge *0.327*. For convenience, we have defined additional view predicates *atomel*, *atomty*, and *atomch*; e.g., *atomel(d1,d1_25,h)*. Bonds between atoms are defined with facts such as *bond(d1,d1_24,d1_25,1)*, meaning that in compound *d1* there is a bond between atoms *d1_24* and *d1_25*, and the bond is of type *1*. There are roughly 18500 of these atom/bond facts to represent the basic structure of the compounds.

In addition, background knowledge contains around 7000 facts and some short DATALOG programs to define mutagenic compounds, genotoxicity properties of compounds, generic structural groups such as alcohols, connections between such chemical groups, tests to verify whether an atom is part of a chemical group, and a family of structural alerts called *Ashby* alerts (Ashby & Tennant 1991).

Representation of substructures. The target patterns or substructures are expressed as DATALOG queries. For instance, ?- *atomel(C,A,c), methyl(C,S), occurs_in(A,S)* is a pattern representing a carbon atom A that occurs in a methyl structure S within compound C.

Related work. Related problems in structure discovery in molecular biology have been considered, e.g., in (Wang *et al.* 1997; Kramer, Pfahringer, & Helma 1997; King *et al.* 1996; King & Srinivasan 1996). Substructure discovery and the utilization of background knowledge have been discussed in (Djoko, Cook, & Holder 1995). Discovery of logical patterns, similar to DATALOG queries, has been considered in (De Raedt & Dehaspe 1997) and in the context of metaqueries (Shen *et al.* 1996).

In data mining, related problems in the area of discovering frequent patterns include association rules (Agrawal *et al.* 1996), and a family of problems discussed in more general in (Mannila & Toivonen 1997).

In (Dehaspe & Toivonen 1999) we discuss the relationship of inductive logic programming (ILP) to frequent pattern discovery, and relate data mining problems to ILP. The logical setting for substructure discovery is based on the learning from interpretations paradigm introduced in (De Raedt & Džeroski 1994).

Frequent substructure discovery

Discovery task. Intuitively, the problem we consider is the following: given the above data on chemical compounds and their structures and properties, find recurrent compound substructures and properties. Since the properties are also a result of the structure of a compound, for the rest of the paper we just talk collectively about (sub)structure discovery.

This problem is an instance of the generic problem of finding all potentially interesting sentences (Mannila & Toivonen 1997). Given a database $\mathbf{r}$, a class $\mathcal{L}$ of sentences (patterns), and a selection predicate q which is used for evaluating whether a sentence $Q \in \mathcal{L}$ defines a potentially interesting pattern in $\mathbf{r}$. The task is to find the theory of $\mathbf{r}$ with respect to $\mathcal{L}$ and q, i.e., the set $Th(\mathcal{L}, \mathbf{r}, q) = \{Q \in \mathcal{L} \mid q(\mathbf{r}, Q) \text{ is true}\}$. In (Dehaspe & Toivonen 1999), this framework has been used to formulate the task of frequent query discovery in DATALOG. We now define frequent substructure discovery as a special case of frequent query discovery.

Definition 1 (Frequent substructure discovery) *Assume*

- $\mathbf{r}$ *is a* DATALOG *database of chemical compounds, their structures and properties, as described above,*
- $\mathcal{L}$ *is a set of substructures expressed as* DATALOG *queries* ?- $A_1, \ldots, A_n$, *where each logical atom* A_i *concerns some structural property of the compounds, as described above,*
- $q(\mathbf{r}, Q)$ *is true if and only if the frequency of query* $Q \in \mathcal{L}$ *with respect to* $\mathbf{r}$ *is at least equal to the frequency threshold specified by the user.*

The task is to find the set $Th(\mathcal{L}, \mathbf{r}, q)$ *of all frequent substructures.*

We next define what frequency exactly means in this setting.

Definition 2 (Substructure frequency) *Given* $\mathbf{r}$ *and* $Q \in \mathcal{L}$ *as above, a relation* keypred(C), *where* keypred *is a predicate name not used in* Q *or* $\mathbf{r}$, *and* C *is the key variable used in* Q *to refer to the compound name, the (absolute) frequency of query* Q *w.r.t.* $\mathbf{r}$ *is*

$$|answerset(\text{?- keypred(C)}, \mathbf{r} \cup \{\text{keypred(C)} \leftarrow Q\})|,$$

i.e., the number of bindings of the C *variable with which the query Q is true in* r*, i.e., the number of compounds in which substructure Q occurs.*

Query extensions. Once frequent substructures and their frequencies are discovered, probabilistic rules, called query extensions, can be produced, much like in the case of association rules. In terms of the DATALOG concepts introduced above, a *query extension* R is an expression of the form $A_1,\ldots,A_k \rightsquigarrow A_{k+1},\ldots,A_n$, where A_i are atoms. This formula should be read as "if query ?- $A_1,\ldots,A_k$ succeeds then extended query ?- $A_1,\ldots,A_n$ succeeds also". The *confidence* of query extension R can be computed as the ratio of the frequencies of queries ?- $A_1,\ldots,A_n$ and ?- $A_1,\ldots,A_k$. The *frequency* (or *support*) of query extension R is the frequency of query ?- $A_1,\ldots,A_n$.

Substructure discovery with WARMR

We now briefly describe the WARMR algorithm used in the experiment. More details can be found in (Dehaspe & Toivonen 1999). WARMR is the first general purpose ILP system to employ the efficient levelwise method known from the APRIORI algorithm (Agrawal *et al.* 1996). In (Dehaspe & Toivonen 1999) we show how WARMR can be tuned to simulate APRIORI and some other well-known algorithms for frequent pattern discovery. A stand-alone version of WARMR is freely available for academic purposes upon request.

The levelwise algorithm (Mannila & Toivonen 1997) is based on a breadth-first search in the lattice spanned by a specialization relation $\preceq$ between patterns, where $p1 \preceq p2$ denotes pattern "$p1$ is more general than pattern $p2$", or "$p2$ is more specific than pattern $p1$". The specialisation relation used in WARMR is θ-subsumption, a stronger variant of the subset relation: $p1$ θ-subsumes a $p2$ if and only if there exists a (possibly empty) binding of the variables of $p2$, such that every logical atom of the resulting query occurs in $p1$. The levelwise method looks at a level of the lattice at a time, starting from the most general patterns. The method iterates between candidate generation and candidate evaluation phases: in *candidate generation*, the lattice structure is used for pruning non-frequent patterns from the next level; in the *candidate evaluation* phase, frequencies of candidates are computed with respect to the database. Pruning is based on monotonicity of $\preceq$ with respect to frequency: if a pattern is not frequent then none of its specialisations are frequent. So while generating candidates for the next level, all the patterns that are specialisations of infrequent patterns can be pruned.

Experiments

In a first experiment, only using atom-bond information, no substructure described with less than 7 logical atoms is found to be related to carcinogenicity. This places a lower limit on the complexity of rules that are based exclusively on chemical structure.

Other experiments based on all available information revealed that the Ashby alerts were not used by any rules. We believe this reflects the difficulty humans and machine have in discovering general chemical substructures associated with carcinogenicity. However, it is possible that the intuitive alerts used by Ashby were incorrectly interpreted and encoded in PROLOG by (King & Srinivasan 1996).

Two particularly interesting rules that combine biological tests with chemical attributes were found. It is difficult to compare these directly with existing knowledge, because most work on identifying structural alerts has been based on alerts for carcinogenicity, while both rules identify alerts for non-carcinogenicity. It is reasonable to search for non-carcinogenicity alerts as there can be specific chemical mechanisms for this, e.g. cytochromes specifically neutralize harmful chemicals. The query extension

cytogen_ca(C,n), sulfide(C,S) $\rightsquigarrow$ *non_carcin(C)*
FREQ:0.06 ; CONF:0.86

for identifying non-carcinogenic compounds is interesting. The combination of conditions in the rule seems to be crucial: the cytogen and sulfide tests in isolation seem to do worse. Within query extension

atomch(C,A,X),
$X \leq$ *-0.215, salmonella(C,n)* $\rightsquigarrow$ *non_carcin(C)*
FREQ:0.27 ; CONF:0.62

the addition of the chemical test makes the biological test more accurate at the expense of less coverage. As the rule refers to charge this rule may be connected to transport across cell membranes.

It is interesting and significant that no atom-bond substructures described with less than 7 conditions were found to be related to carcinogenicity. This result is not inconsistent with the results obtained by (King & Srinivasan 1996) and (Srinivasan *et al.* 1997) using PROGOL because most of the substructures there involve partial charges, and the ones that don't do not meet the coverage requirements set in the WARMR experiments.

Although the lack of significant atom-bond substructures found in the WARMR experiments is disappointing, it is perhaps not too surprising. The causation of chemical carcinogenesis is highly complex with many separate mechanisms involved. Therefore predicting carcinogenicity differs from standard drug design problems, where there is normally only a single well defined mechanisms. We consider that it is probable that the current database is not yet large enough to provide the necessary statistical evidence required to easily identify chemical mechanisms. Biological tests avoid this problem because they detect multiple molecular mechanisms; e.g., the Ames test for mutagenesis detects many different ways chemicals can interact with DNA and cause mutations; biological tests also detect whether the compound can cross cell membranes and not be destroyed before reaching DNA. Biological tests vary in expense, speed, and accuracy. At the extreme cheap

and fast and relatively inaccurate end is the Ames test for mutagenicity, this is fast and uses bacteria (so there are no ethical issues). At the other end are long expensive trials which involve the dissection of thousands of rodents.

The ultimate goal of our work in predictive toxicology is to produce a program that can predict carcinogenicity *in humans* from just input chemical structure. Such a system would allow chemicals to be quickly and cheaply tested without harm to any animals. This goal is still far distant. Our results suggest that an intermediate goal for data mining in this predictive toxicology problem is to identify the combination of biological tests and chemical substructures that provides the most cost-effective tests for testing chemical carcinogenesis.

Conclusions

We presented a data mining problem in a biochemical database. The goal is to discover frequent substructures of chemical compounds in relation to their possible carcinogenicity. Rather than trying to predict the toxicity of unknown compounds, our purpose is to assist chemical experts in discovering chemical mechanisms of toxicology.

One result of our experiment is a repository of frequent substructures in a general DATALOG format. We believe this repository constitutes a new description of the data that is useful for chemists and data miners looking for predictive theories. We have also identified substructures, both known and new, that could be related to carcinogenicity. On the other hand, we have found that, within this biochemical database, short, accurate and highly significant rules apparently do not exist.

Acknowledgements. Luc Dehaspe is supported by ESPRIT Long Term Research Project No 20237, ILP^2. Hannu Toivonen is supported by the Academy of Finland. R.D. King was partly supported by grant BIF08765 from the BBSRC and grant GR/L262849 from the EPSRC. We are grateful to Luc De Raedt and Heikki Mannila for discussions and comments, and to Wim Van Laer and Hendrik Blockeel for their share in the implementation of WARMR.

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. I. 1996. Fast discovery of association rules. In Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining.* Menlo Park, CA: AAAI Press. 307 – 328.

Ashby, J., and Tennant, R. W. 1991. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutation Research* 257:229–306.

Bristol, D.; Wachsman, J.; and Greenwell, A. 1996. The NIEHS predictive-toxicology evaluation project. *Environmental Health Perspectives* Supplement 3:1001–1010.

De Raedt, L., and Dehaspe, L. 1997. Clausal discovery. *Machine Learning* 26:99–146.

De Raedt, L., and Džeroski, S. 1994. First order *jk*-clausal theories are PAC-learnable. *Artificial Intelligence* 70:375–392.

Dehaspe, L., and Toivonen, H. 1999. Discovery of frequent Datalog patterns. *Data Mining and Knowledge Discovery* 3(1):7 – 36.

Djoko, S.; Cook, D. J.; and Holder, L. B. 1995. Analyzing the benefits of domain knowledge in substructure discovery. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, 75 – 80.

King, R., and Srinivasan, A. 1996. Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environmental Health Perspectives* 104(5):1031–1040.

King, R.; Muggleton, S.; Srinivasan, A.; and Sternberg, M. 1996. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences* 93:438–442.

Kramer, S.; Pfahringer, B.; and Helma, C. 1997. Mining for causes of cancer: machine learning experiments at various levels of detail. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, 223 – 226.

Mannila, H., and Toivonen, H. 1997. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery* 1(3):241 – 258.

Shen, W.; Ong, K.; Mitbander, B.; and Zaniolo, C. 1996. Metaqueries for data mining. In Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining.* Menlo Park, CA: AAAI Press. 375–398.

Srinisavan, A.; King, R. D.; Muggleton, S. H.; and Sternberg, M. J. E. 1997. The predictive toxicology evaluation challenge. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97).* Morgan Kaufmann.

Srinivasan, A.; King, R.; Muggleton, S.; and Sternberg, M. 1997. Carcinogenesis predictions using ILP. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, Lecture Notes in Artificial Intelligence, 273–287. Springer-Verlag.

Wang, X.; Wang, J. T. L.; Shasha, D.; Shapiro, B.; Dikshitulu, S.; Rigoutsos, I.; and Zhang, K. 1997. Automated discovery of active motifs in three dimensional molecules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, 89 – 95.