Representational/Efficiency Issues in Toxicological Knowledge Discovery

Bernhard Pfahringer
Austrian Research Institute for AI
Schottengasse 3, A-1010 Vienna
bernhard@ai.univie.ac.at

Eva Gottmann
Institute for Environmental Hygiene
Kinderspitalgasse 15, A-1095 Vienna
a9011094@unet.univie.ac.at

Stefan Kramer
Austrian Research Institute for AI
Schottengasse 3, A-1010 Vienna
stefan@ai.univie.ac.at

Christoph Helma
Institute for Tumor Biology – Cancer Research
Borschkegasse 8a, A-1090 Vienna
Christoph.Helma@univie.ac.at

Abstract

We tackle the problem of automatic detection of Structure-Activity relationships by means of Inductive Logic Programming (ILP). We describe an algorithm for constructing features from background knowledge (stochastic propositionalization – SP) and an abstract level representation for chemical compounds.

Introduction

The current standard representation of chemical compounds used in ILP comprises relations describing single atoms and single bonds for capturing the two-dimensional molecular structure. Additionally, ad-hoc definitions of functional groups as well as predefined structural alerts are usually supplied. Apart from the obvious inefficiencies due to missed indexing possibilities we have identified two major problems with this kind of representation:

- Dynamically constructing arbitrary (connected) fragments of compounds and checking all compounds for the presence of these fragments quickly exceeds computational resources even for medium-sized fragments. Cyclic structures pose a particularly severe problem in this context. Sub-graph isomorphism detection is NP-hard in general, and due to the fine-grained atom/bond representation only rather small fragments can be investigated effectively. Additionally, the total number of distinct reasonable-sized fragments grows beyond any limit quickly.
- Even though the ad-hoc definitions of both functional groups and structural alerts seem useful in guiding search, they might as well drive heuristic search into local minima quickly. Furthermore, since such ad-hoc sets of definitions come without any guarantees on completeness, it is not obvious that all definitions necessary for solving a respective task have been supplied.

Our attempts at solving these problems are three-fold:

- We have already devised a stochastic schema for constructing small sets of fragments of high expected utility given a prespecified task, e.g. the prediction of mutagenicity or carcinogenicity. Such fragments are therefore task-specific and one should expect to induce different sets of fragments for different tasks. On the other hand, for related tasks a transfer of relevant features should be possible. The next section will describe this approach in more detail.
- We have defined and currently are implementing an abstract high-level acyclic representation for compounds based on functional groups. Most of the chemical information relevant to our tasks should still be represented at the abstract level. The abstract level will allow search to consider more complex substructures.
- This abstract-level representation will enable systematic search – based on association rule ideas – to construct fragments that can span larger parts of a compound than fragments formulated in an atom/bond representation ever could. In addition, the stochastic procedure (SP) might benefit from the high-level representation as well.

In the following two sections will describe these topics in more detail.

Stochastic Propositionalization

Learning algorithms require an adequate representation of the data to perform well in practice. Usually such a representation is either engineered manually by domain experts or derived automatically by means of so-called constructive induction. Inductive Logic Programming (ILP) algorithms put a somewhat less burden on the data engineering effort as they allow for a structured, relational representation of background knowledge. In chemical domains, a common representational device for graph-like structures is so-called non-determinate relations. Manually engineered features in such domains typically test for or count occurrences of specific substructures having specific properties. However, representations containing non-
determinate relations pose a serious efficiency problem for most standard ILP algorithms (Quinlan 90; Muggleton 95; Lavrac & Džeroski 94). Therefore, we have devised a stochastic algorithm to automatically derive features from non-determinate background knowledge. The algorithm conducts a top-down search for first-order clauses, where each clause represents a binary feature. These features are used instead of the non-determinate relations by the subsequent learning algorithms.

Feature construction (including SP) strives for features that help to create useful partitions of the examples. Additionally, in practice we often know some expert-provided features (such as the so-called “Hansch attributes” or “structural alerts” in toxicology), and we are interested in finding features that complement them. Since we have to select from a large number of features, we have to specify preferences. We used the following constraints that should be fulfilled by features constructed from background knowledge. The constructed features should be

- (C1) not too specific and not too general.
- (C2) not too complex. (So they have at least a potential for being comprehensible.)
- (C3) different from one another.
- (C4) different from existing, expert-provided features.

SP is an evolutionary algorithm that grows sets of useful features by means of a refinement operator defined by schemata (Silverstein & Pazzani 91; Blockeel & DeRaedt 98). The above constraints are implemented by min- and max-cover parameters for single clauses (C1), an MDLP-based evaluation formula (C2), and enforcement of different extensions to already present clauses (C3 and C4). Sets of clauses are evaluated together to ensure proper coverage of the training examples. This is also the motivation for the universal suffrage selection algorithm presented in (Giordana et al. 94.).

SP has been evaluated empirically in three chemical domains: the prediction of biodegradation (Džeroski and Kompare 95), mutagenity (Srinivasan et al. 95), and carcinogenicity (King and Srinivasan 97). For all these domains, SP was able to construct features that allow for the formulation of small and accurate decision trees. Features were constructed automatically, and yet are of a quality comparable to expert-provided features.

Examples of constructed features are listed in the following. The first feature was constructed in the mutagenicity prediction domain:

\[
\text{new}_f7(A) : \leftarrow \text{atm}(A, _{-}, _{-}, 27, _{-}), \text{sym}_{-\text{bond}}(A, B, _{-}, 21, _{-}), \text{atm}(A, C, _{-}, 29, _{-}).
\]

This definition captures atoms of type 29 occurring connected to an atom of type 27, where types 27 and 29 designate different kinds of aromatic carbon atoms.

A sample of example features from a successful generation of features in the carcinogenicity domain looks like this:

\[
\text{new}_f8(A) : \leftarrow \text{atm}(A, _{-}, br, _{-}, _{-}).
\]

\[
\text{new}_f9(A) : \leftarrow \text{atm}(A, B, _{-}, 21, _{-}), \text{sym}_{-\text{bond}}(A, B, C, _{-}), \text{atm}(A, C, _{-}, 3, _{-}).
\]

\(\text{new}_f7\) means a carbon atom in a benzene ring connected to another aromatic or hetero-aromatic ring and the presence of nitrogen in the compound. This feature covers (in the training set) mainly benzidine compounds which are well known for their carcinogenic activity.

\(\text{new}_f8\) covers bromine-containing compounds, mainly carcinogenic brominated aliphatic hydrocarbons (in the training set).

\(\text{new}_f9\) points to a carbon atom in an aromatic five-membered ring with an hydrogen attached to it. It is interesting that the presence of hydrogen excludes substituted and conjugated carbons.

Generally, as shown in (Kramer et al. 98), SP appears to construct features that allow for the formulation of small and accurate decision trees. Features are constructed automatically, and yet are of a quality comparable to expert-provided features.

**Abstract representation**

We are in the process of implementing a more abstract representation capturing the two-dimensional structure of compounds in a more coarse-grained acyclic graph-structure. Such an abstract higher-level representation reduces the number of possible fragments and therefore allows for deeper searches. The inherent risk of losing essential information in the abstraction step is balanced by grounding the design of the building blocks for the abstract level on functional groups wherever possible. Thus, most of the relevant chemical information should still be represented at the abstract level.

Basically, we are abstracting all parts of a compound as described below. Everything not covered (e.g. an Mg atom, or Pb, etc.) will simply be lifted “as is” into the abstract representation as to make it complete. Currently we distinguish the following sub-structures (SMILES strings (Weininger 88) are used throughout to clarify the used nomenclature):

- Alkyl chains: compute the length and the number of multiple bonds present in the chain.
- Rings: compute the size, determine aromaticity, heterogeneity, and for systems of multiple rings whether they are condensed or merged.
- Oxygen-specific substructures: we currently extract the following:
  - Aldehydes, *C=O
  - Ketones, *C(=O)*
  - Quinones, e.g. O=cclcc(c=O)cc1
  - Carboxylic acids, *C(=O)O
  - Esters, *C(=O)O*
  - Alcohols, *O
  - Phenols, e.g. Oclccccc1
- Methoxy groups, *C(=O)OC
- Ethers, *O"

• Nitrogen-specific substructures:
  - nitroso compounds, *N=O
  - nitro compounds, *[N+(=O)][O-]
  - nitriles, *C#N
  - azo compounds, *N=N
  - diazo compounds, *[N+]#N

• Sulphur-specific substructures:
  - thiols, *S
  - sulfenic acids, *SO
  - sulfides, *S*
  - sulfenic acids, *S(=O)O
  - sulfones, *S(=O)

• Phosphor-specific substructures:
  - phosphanes (PnH2n)
  - phosphoranes (PH5 and and its hydrocarbyl derivatives)
  - phosphine oxides, *P(=O)(*)
  - phosphine sulfides, *P(=S)(*)
  - phosphinic imides, *P(=N)(*)
  - phosphoric acids, *P(=O)O
  - phosphinous acids, *PO
  - phosphonic acids, *P(=O)(O)O
  - phosphonous acids, *P(O)O
  - phosphoramic acids, N(*)P(=O)(O)N(*), N(*)P(=O)(O)O
  - phosphor triamides, N(*)P(=O)(N(*)*)N(*)*

Due to the way the abstract representation is computed, it is significantly smaller than the base-level representation while at the same time (hopefully) still capturing most relevant chemical information.

% prolog-terms for smiles_code : Cc1cccc1
  atm(108_88_3,108_88_3_1,1,h,0,0).
  bond(108_88_3,108_88_3_1,108_88_3_1_3,1).
  atm(108_88_3,108_88_3_1,108_88_3_1_2,h,0,0).
  bond(108_88_3,108_88_3_1,108_88_3_1_1,h,0,0).
  atm(108_88_3,108_88_3_1,108_88_3_1_1_1,1).
  atm(108_88_3,108_88_3_1,108_88_3_2,c,0,0).
  atm(108_88_3,108_88_3_3_1,h,0,0).
  bond(108_88_3,108_88_3_3_1,108_88_3_3_1_1,1).
  atm(108_88_3,108_88_3_3_1,108_88_3_3_2,1).
  atm(108_88_3,108_88_3_3_2,c,0,0).
  atm(108_88_3,108_88_3_3_3_1,h,0,0).
  bond(108_88_3,108_88_3_3_3_1,108_88_3_3_3_1_1,1).
  atm(108_88_3,108_88_3_3_3_1,108_88_3_3_3_1_2,1).
  atm(108_88_3,108_88_3_4,c,0,0).
  bond(108_88_3,108_88_3_4,108_88_3_3_4_1,h,0,0).
  atm(108_88_3,108_88_3_4,108_88_3_3_4_1_1,1).
  atm(108_88_3,108_88_3_4,108_88_3_3_4_2,c,0,0).
  bond(108_88_3,108_88_3_4,108_88_3_3_5,c,0,0).
  bond(108_88_3,108_88_3_5_1,108_88_3_3_5_1,1).
  bond(108_88_3,108_88_3_5_1,108_88_3_3_5_2,c,0,0).
  bond(108_88_3,108_88_3_5_2,108_88_3_3_5_3,7).
  bond(108_88_3,108_88_3_5_2,108_88_3_3_5_4,7).
  bond(108_88_3,108_88_3_5_2,108_88_3_3_5_5,7).
  atm(108_88_3,108_88_3_3_5_5_1,h,0,0).
  bond(108_88_3,108_88_3_5_5_1,108_88_3_3_5_5_1_1,1).
  atm(108_88_3,108_88_3_3_5_5_1,108_88_3_3_5_5_1_2,c,0,0).
  bond(108_88_3,108_88_3_3_5_5_1_2,108_88_3_3_5_5_1_3,7).
  atm(108_88_3,108_88_3_3_5_5_1_3,h,0,0).
  bond(108_88_3,108_88_3_3_5_5_1_3,108_88_3_3_5_5_1_4,7).
  atm(108_88_3,108_88_3_3_5_5_1_4,c,0,0).
  atm(108_88_3,108_88_3_3_5_5_1_4,108_88_3_3_5_5_1_5,c,0,0).
  atm/5 facts describe the single atoms of compound, whereas bond/4 facts list the (possibly multiple) bindings between respective atoms. The abstract representation is much smaller:

% the abstract level representation
  abs_atm(108_88_3,0,alkyl(1,f),[108_88_3_1]).
  abs_atm(108_88_3,1,ring(6,uniform, resonant [108_88_3_2,108_88_3_3,108_88_3_4,108_88_3_5,108_88_3_6,108_88_3_7]).
  abs_bond(108_88_3,0,1,1).

Two "abstract atoms" numbered 0 and 1 and comprise the abstract compound. "Abstract atom" 0 is an alkyl chain of length 1 with no multiple bonds. "Abstract atom" 1 is an uniform (i.e. all-carbon), resonant ring of size six. Both "abstract" atoms are linked together by an "abstract bond". We also keep links back to the base-level thus identifying each base-level atom involved in any "abstract atom".

Search

The smaller and acyclic nature of the abstract representation allows for comparatively large (in terms of base-level representation) substructures to be constructed and searched for.

We are currently working on/ respectively plan to use this abstract level representation for the following:

A simple example will illustrate the achievable simplification. The base-level representation of Methyl-Benzene (CAS 108-88-3, depicted in Figure 1) is:

\[
\text{H}_3\text{C} \quad \text{Figure 1: Methyl-Benzene}
\]
 Depth-limited complete search (like CASE/MULTICASE (Klopman & Rosenkranz 94)).

- Run SP on it trying to uncover useful complex abstract features.

- Investigate whether one or more of the so-called Ashby-alerts can be rediscovered by the former two approaches.

- Utilize these features as the input for some standard clustering algorithm, e.g. (Jarvis & Patrick 73).

Complete search ala CASE usually explores connected fragments up to a size of around 10 atoms. The acyclic abstract representation should allow for about the same number of abstract atoms in an abstract fragment. Effectively, such fragments could cover literally dozens of base-level atoms, thus vastly increasing the search horizons, albeit in a limited knowledge-based manner. With these experiments we will assess the real benefit of the specific abstract level representation we have devised.

Summary

We have described an algorithm for stochastic proposition-alization (SP) and we have devised an abstract level representation for chemical compounds. Such a representation can be utilized in several distinct ways. All different ways are aimed at enabling the discovery of larger than state-of-the-art-sized fragments in compounds. Even though some success has been achieved, there is still a lot of room for improving the representation of chemical structure for Machine Learning purposes. An important aspect that we have omitted so far from our experiments is the proper treatment of three-dimensional structure. Currently, we only incorporate some global properties (e.g. volume, surface, or logP) which partially capture some of the 3D properties of a compound. Obviously, a more thorough representation of 3D information will have to be a focus of our future research.

Acknowledgements

This research is part of the project “Carcinogenicity Detection by Machine Learning” supported by the Austrian Federal Ministry of Science and Transport. Partial support is also provided by the “Jubiläumsfonds der Österreichischen Nationalbank” under grant number 6930. The Austrian Federal Ministry of Science and Transport provides general financial support for the Austrian Research Institute for Artificial Intelligence.

References


