Discovery of Toxicological Patterns with Lazy Learning

Eva Armengol and Enric Plaza

IIIA - Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research
Campus UAB, 08193 Bellaterra, Catalonia (Spain)
{eva,enric}@iiia.csic.es

Abstract. In this paper we propose the use of a lazy learning technique called LID for discovering patterns in the Toxicology dataset. LID classifies examples and builds an explanation of that classification. We analyzed the Toxicology dataset using a two-step proces: first we use LID for classifying all the cases in the dataset. Then we select a subset of explanations and use them as patterns that capture structural regularities (*patterns*) among carcinogenic chemical compounds.

1 Introduction

Computer-based Toxicology (sometimes called *Toxicoinformatics*) uses automatic tools for analysing the toxicity of a molecule based on its molecular structure. Often these tools are used by pharmaceutical industries for designing drugs with desired properties. In particular, one of these properties can be the toxicity of molecules. The use of automatic tools also requires the definition of some kind of representation of the chemical compounds. Representations widely used by commercial software are SAR (*Structure-Activity Relationship*) and QSAR (*Quantitative Structure Activity Relationship*). Both approaches use induction to detect generalizations from a set of compounds having the property of interest. Most of authors in the Predictive Toxicology Challenge [1] use relational representations based on QSAR descriptors.

In [2] we proposed an alternative representation of chemical compounds based on the ontology used by the chemists. Currently, this representation only takes into account the physical structure of the molecule. Our point is that good results could be obtained without individually describing each atom of the molecule since most of them have well-known properties. Nevertheless, we could easily add information to this representation.

In this paper we use LID, a lazy learning technique useful for solving the classification task. In addition to classify a compound, LID gives an explanation of that classification. We take benefit of this explanation for discover patterns in the Toxicology dataset.

This paper is organized as follows. Section 2 briefly describes LID. In section 3 we explain the description of the chemical compounds and how LID can support the global study of a dataset. Section 4 we discuss the LID results.

V. Palade, R.J. Howlett, and L.C. Jain (Eds.): KES 2003, LNAI 2774, pp. 919–926, 2003.

2 Lazy Induction of Descriptions

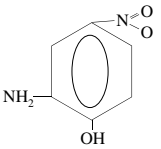
The goal of LID (*Lazy Induction of Descriptions*) is to determine the class of a problem, i.e. LID is able to learn the classification of new examples using previously defined examples. LID determines which are the more relevant features of the problem and to search in the case base for cases sharing these relevant features. The problem is classified when LID finds a set of relevant features shared by a subset of cases belonging to the same solution class. We call the structure formed by these features *similitude term*. LID uses the feature term formalism for representing cases. In section 3.1 feature terms are explained with an example. For a more formal explanation of feature terms see [2].

LID inputs are: a case base B, a similitude term D initialized to the most general feature term (i.e. the most general description), a problem p, the set S_D (*discriminatory set associated to* D) that contains all the cases that satisfy the structure described by D. Initially $S_D = B$ since D is satisfied by all the cases in B. The first step of LID is to check whether all the cases in S_D belong to the same solution class. If this stopping condition is not satisfied, LID selects one of the features of the problem p and adds it to the current similitude term D in order to construct a new similitude term D^1 that specializes D. Next, LID is recursively called using the similitude term D^1 and the discriminatory set S_{D^1} containing only those cases in S_D satisfied by D^1. This process continues until either the similitude term D^n is specific enough to satisfy the stopping condition or there are no more possible features to add to the D^n. The LID algorithm and some examples of application can be found in [3].

The result of LID is a solution class C_i and a similitude term D^n. The similitude term D^n can be seen as an explanation of why p belongs to C_i. Notice that the stopping condition means that D^n is able to discriminate some cases belonging to C_i. D^n is a *partial* description of C_i. D^n is partial because, in general, it does not satisfy all the cases belonging to C_i but only a subset of them (those sharing the features of D^n with p). The similitude term D^n depends on the new problem, therefore several partial descriptions can be built for one solution class.

3 Discovery of Patterns in the Toxicology Dataset

The Toxicology dataset contains descriptions of around 500 chemical compounds that may be carcinogenic to two animal species: rats and mice. The carcinogenic activity of the compounds has proved to be different in both species and also both sex of the same species. Therefore there are, in fact, four datasets. The chemical compounds of the dataset can be classified into eight solution classes according to the laboratory experiments: *positive*, *clear evidence*, *some evidence*, *equivocal*, *equivocal evidence*, *inadequate study*, *negative* and *negative evidence*. Nevertheless, most of the authors working on this dataset consider the classes *positive*, *clear evidence* and *some evidence* as the class "positive"; the classes *negative* and *negative evidence* as the class "negative"; and the compounds belonging to the other classes are removed.



$$TR\text{-}339 = \left[\begin{array}{l} \underline{\mathsf{compound}} \\ \mathsf{main\text{-}group} \doteq benzene \\ \mathsf{p\text{-}radicals} \doteq \left\{ \begin{array}{l} \left[\begin{array}{l} \underline{position\text{-}radical} \\ \mathsf{position} \doteq one \\ \mathsf{radicals} \doteq \left[\begin{array}{l} \underline{compound} \\ \mathsf{main\text{-}group} \doteq alcohol \end{array}\right] \end{array}\right] \\ \left[\begin{array}{l} \underline{position\text{-}radical} \\ \mathsf{position} \doteq two \\ \mathsf{radicals} \doteq \left[\begin{array}{l} \underline{compound} \\ \mathsf{main\text{-}group} \doteq amine \end{array}\right] \end{array}\right] \\ \left[\begin{array}{l} \underline{position\text{-}radical} \\ \mathsf{position} \doteq four \\ \mathsf{radicals} \doteq \left[\begin{array}{l} \underline{compound} \\ \mathsf{main\text{-}group} \doteq nitro\text{-}derivate \end{array}\right] \end{array}\right] \end{array} \right. \end{array}\right]$$

Fig. 1. Representation of the compound with identifier TR-339, *2-amino-4-nitrophenol*, using feature terms

In the next section we explain how the chemical compounds are represented using feature terms. Then we describe the application of LID to discover patterns.

3.1 Representation of the Chemical Compounds

The basis of the representation we propose is the *chemical ontology* used by chemist experts, and that is implicit in the chemical nomenclature of the compounds. Our point is that the chemical nomenclature is a systematic way of describing a molecule and that the name of a molecule provides to a chemist all the necessary information about the structure of a molecule. For instance, *benzene* is an aromatic ring composed by six carbon atoms with some well-known properties, therefore it is not necessary to describe the individual atoms in the benzene when we have the *benzene* concept in our domain ontology.

In our representation a chemical compound is represented by a feature term (see Figure 1) of sort *compound*. This sort is described by two features: main-group and p-radicals. The value of the feature p-radicals is a set whose elements are of sort *position-radical*. The sort *position-radical* is described using two features: radicals and position. The value of the feature radicals is also of sort *compound*. This is because both, main group and radicals, are the same kind of molecules, i.e. benzene may be the main group in one compound and a radical in some other compounds. The feature position of the sort *position-radical* indicates where the radical is bound to the main group.

For example, the compound in Figure 1 with identifier TR-339, is the *2-amino-4-nitrophenol*. This compound has a phenol as main group. Phenol is a molecule composed of one benzene with an alcohol radical in position one. Thus, the compound TR-339 has a benzene as main group and a set of three radicals: a radical with an *alcohol* as main group in position one; a radical with an *amine* as main group in position two; and a radical with a *nitro-derivate* in

position four. Notice that this information has been directly extracted from the chemical name of the compound following the nomenclature rules.

3.2 Analysis of Patterns for the Toxicology Dataset

In this section we explain how the similitude terms can be used to discover patterns in the Toxicology dataset. Our experiment two steps: 1) use LID with the leave-one-out method in order to generate similitude terms for classifying the cases; and 2) select a subset of these similitude terms.

The first step of the problem was to solve each problem of the case base using LID. At the end of this step there is a set of similitude terms that have been used for classifying some cases. Some of these similitude terms are totally discriminatory, since they satisfy cases belonging to only one solution class, whereas others are not. For this reason, during the second step we selected only those similitude terms that either are totally discriminatory or they satisfy a majority of cases belonging to one class. In particular, we selected those similitude terms whose associated discriminatory set contains more than the 3/4 of cases belonging to the same solution class. From now on, we call to these similitude terms *patterns* and M the set of these patterns.

In the following we explain some of the patterns found by LID that classify a molecule as positive. For some of these patterns we have encountered some evidence of toxicity in the literature but there are some other patterns whose positive toxicity has not been clearly reported.

Molecules with a chlorine radical. There is a subset of compounds that LID has classified as positive or negative using as explanation that they have a radical *chlorine* (pattern m_1). When we use m_1 to retrieve the compounds of the whole case base that have a radical chlorine we obtain the results shown in Table 1. The column labeled as *#cases* shows the number of cases in each dataset that satisfy m_1. The column $T+$ shows the number of these cases that have positive activity and $T-$ shows the number of cases that have negative activity. In particular, from the 47 compounds having a radical chlorine in the MR dataset, 15 of them are positive and 32 are negative. Notice that for female rats (FR) it seems to be clear that these molecules are not carcinogenic, nevertheless this is not so clear for the other datasets. In the literature we have found some compounds with chlorine (such as the vinyl chloride or the chloroform) that are positive. Brautbar (in www.expertnetwork.com/med2.htm) describes some experiments proving that chlorinated hydrocarbons are carcinogenic.

Molecules with an anthracene as main group. LID has classified some molecules as having positive carcinogenic activity giving as explanation that they have an anthracene as main group (pattern m_2). An analysis of the Toxicology dataset reveals that all the compounds having anthracene as main group have been considered positive in rats (see Table 1). We have not found laboratory experiments confirming this result. In fact, the anthracene cannot be considered as toxic itself, but it is a molecule having a high tendency to make associations with other molecules and these associations could easily be carcinogenic.

Table 1. Tables reporting, for each dataset, the total number of compounds satisfying a given pattern (*#cases*), cases with positive carcinogenicity ($T+$) and cases with negative carcinogenicity ($T-$)

	Pattern m_1			Pattern m_2			Pattern m_3		
Dataset	*#cases*	T +	T -	*#cases*	T +	T -	*#cases*	T +	T -
MR	47	15	32	5	5	0	5	5	0
FR	46	9	37	4	4	0	3	3	0
MM	48	25	23	4	3	1	6	4	2
FM	47	25	22	6	3	3	6	5	1

Molecules with an epoxide radical. Molecules having an epoxide (pattern m_3) are classified by LID as having positive activity. There are five compounds containing an epoxide whose effects have been studied in rats and all them have been considered as positive cacinogenicity (see Table 1). Several laboratory experiments done by Melnik (see members.nyas.org/events/conference/conf_02_0429.html) proved the carcinogenicity of the epoxides and their precursors. Also, there are studies (for instance those described by Glukster in www.fccc.edu/research/reports/report98/glukster.html) showing that an epoxide produces the positive activation of polycyclic aromatic hydrocarbons.

Molecules with a bromine radical. LID has classified as having positive carcinogenic activity some molecules with a radical bromine (pattern m_4). In the database, there are 10 compounds with a radical bromine, and three compounds having a radical with a radical bromine (pattern m_5). Table 2 shows that most of them have a confirmed positive activity. We found in the HERP index (potency.berkeley.edu/herp.html) some experiments proving that compounds with bromine (e.g. ethylene dibromide) have positive activity. In [4] some models are introduced to predict the carcinogenicity of chemical compounds. These models use domain knowledge in the form of rules to increase the predictive accuracy. One of the rules introduced by the experts is to consider as positive those compounds with bromine.

Table 2. Results of the retrieval with the patterns m_4, m_5, m_6 and m_7

	Pattern m_4			Pattern m_5			Pattern m_6			Pattern m_7		
Dataset	*#cases*	T +	T -	*#cases*	T +	T -	*#cases*	T +	T -	*#cases*	T +	T -
MR	3	3	0	10	8	2	8	8	0	5	5	0
FR	3	3	0	9	7	2	7	6	1	4	4	0
MM	2	2	0	9	5	4	6	4	2	4	3	1
FM	3	2	1	11	7	4	8	5	3	6	3	3

Table 3. Results of the retrieval with the patterns m_8, m_9, and m_{10}

	Pattern m_8			Pattern m_9			Pattern m_{10}		
Dataset	*#cases*	T +	T -	*#cases*	T +	T -	*#cases*	T +	T -
MR	3	3	0	0	0	0	5	2	3
FR	3	3	0	2	2	0	5	2	3
MM	3	2	1	2	1	1	4	3	1
FM	3	2	1	3	0	3	4	3	1

Molecules formed by linear chains of at least 9 carbons. LID has classified as having positive activity molecules having one radical in the position 9 (pattern m_6). Also, LID has classified as positive molecules having a radical in the position 10 (pattern m_7). These patterns means, in fact, that these molecules are chains (hydrocarbons) of at least either 9 or 10 carbons. An analysis of the dataset shows that most of molecules that are hydrocarbons having 9 or 10 carbons have been considered as positive in the laboratory experiments, especially for rats (see Table 2). Laboratory experiments done by Belpoggi (explained in members.nyas.org/events/ conference/conf_02_0429.html) also proved that the gasoline (a hydrocarbon with 8 carbons) is carcinogenic for rodents and also there are hydrocarbons such as 2,2,4-trimethyl pentane or the 1-chloro-2-propanol that are carcinogenic. Notice that the last compound satisfies also the first pattern above since they have a radical chlorine.

Molecules related to butane. LID has classified as positive some molecules having acyclic unsaturated butane as the main group (pattern m_8) and also some molecules having the butane as radical (pattern m_9). As shown in Table 3, the toxicity when butane is the main group seems to be clear in rats. In the literature we found that compounds such as the butylated hydroxyanisole and the 1,3-butadiene have a positive activity.

Molecules with an ether as main group. The results of this pattern (m_{10}) are not clear for rats (see Table 3). Nevertheless, most of molecules with an ether as main group are considered carcinogen in mice. Experiments reported by Belpoggi seem to confirm the carcinogenicity of some molecules with ether such as methyl-tertiary-butyl ether (MTBE), ethyl-tertiary-butyl ether (ETBE), tertiary-amyl-methyl ether (TAME), di-isopropyl ether (DIPE). Notice that some of these compounds also satisfy the previous patterns related to butane.

4 Discussion

Some of the patterns build by LID are satisfied by few molecules, so it is not possible to determine its validity when there is no experimental evidence of carcinogenicity. Nevertheless the LID patterns can suggest positive or negative tendency of the compounds and, in that way, they can support the selection of appropriate laboratory experiments to determinate the toxicity of a compound.

Because LID has been able to find patterns that are already known, we can expect that it is a good tool to be used as first step for testing carcinogenicity in unknown compounds. Moreover, LID has found positive activity patterns that are followed by only a few compounds of the dataset but that they are known as carcinogenic in the literature. For instance, Antosiewicz et al. [5] made some experiments proving that the hydrazine is carcinogenic, and in the dataset there are only two compounds with hydrazine and both are positive in rats.

LID results could be improved in three ways. The first one is to use other databases containing chemical compounds since we consider that the NTP case-base is not representative enough. This could produce more accurate patterns since LID could work with more examples. Secondly, we could use combinations of patterns to predict the carcinogenic activity of the compounds in a way similar to the described by Okada in [6].

Thirdly, we could introduce the *multiexamples* concept [7]. The idea is that the representation of an example is not unique. In the chemical domain, some molecules could be represented in several equivalent ways depending on the group that we consider as the main. For instance, DDT can be formulated either as *1,1'-(2,2,2-trichloroethylidene)bis(4-chloro)-benzene* meaning that the main group is a benzene or also as the *1,1,1-trichloro-2,2-bis(p-chlorophenyl)-ethane* meaning that the main group is an ethane. Using feature terms, both representations are different and they produce different patterns in LID.

Acknowledgements

This work has been supported by the MCYT-FEDER Project SAMAP (TIC 2002-04146-C05-01). The authors thank Dr. Lluis Bonamusa for his assistance in developing the representation of chemical molecules.

References

[1] Helma, C., King, R., Kramer, S., Srinivasan, A.: The predictive toxicology challenge 2000-2001. In: ECML/PKDD 2001. Freiburg. (2001) 919

[2] Armengol, E., Plaza, E.: Relational case-based reasoning for carcinogenic activity prediction. AIReview. Special Issue on Life Sciences. (in press) (2003) 919, 920

[3] Armengol, E., Plaza, E.: Lazy induction of descriptions for relational case-based learning. In: Machine Learning: ECML-2002. Number 2167 in Lecture Notes in Artificial Intelligence, Springer-Verlag (2001) 13-24 920

[4] Blockeel, H., Driessens, K., Jacobs, N., Kosala, R., Raeymaekers, S., Ramon, J., Struyf, J., Laer, W. V., Verbaeten, S.: First order models for the predictive toxicology challenge 2001. In: Proceedings of the Predictive Toxicology Challenge Workshop, Freiburg, Germany, 2001. (2001) 923

[5] Antosiewicz, J., Matuszkiewicz, A., Olek, R. A., Kaczor, J. J., Zilkowski, W., Wakabayashi, T., Popinigis, J.: Content and redistribution of vitamin E in tissues of wistar rats under oxidative stress induced by hydrazine. Arch. Environ. Contam. Toxicol. 42 (2002) 363-368 925

[6] Okada, T.: Discovery of structure activity relationships using the cascade model: the mutagenicity of aromatic nitro-compounds. Journal of Computer Aided Chemistry 2 (2001) 79-86 925

[7] Dietterich, T. G., Lathrop, R. H., Lozano-Perez, T.: Solving the multiple instance problem wich axis-parallel rectangles. Artificial Intelligence 89 (1997) 31-71 925