Mining for Causes of Cancer: Machine Learning Experiments at Various Levels of Detail

Stefan Kramer and Bernhard Pfahringer and Christoph Helma

Abstract

This paper presents, from a methodological point of view, first results of an interdisciplinary project in scientific data mining. We analyze data about the carcinogenicity of chemicals derived from the carcinogenesis bioassay program, a long-term research study performed by the US National Institute of Environmental Health Sciences. The database contains detailed descriptions of 6823 tests performed with more than 330 compounds and animals of different species, strains and sexes. The chemical structures are described at the atom and bond level, and in terms of various relevant structural properties. The goal of this paper is to investigate the effects that various levels of detail and amounts of information have on the resulting hypotheses, both quantitatively and qualitatively. We apply relational and propositional machine learning algorithms to learning problems formulated as regression or as classification tasks. In addition, these experiments have been conducted with two learning problems which are at different levels of detail.

Quantitatively, our experiments indicate that additional information not necessarily improves accuracy. Qualitatively, a number of potential discoveries have been made by the algorithm for Relational Regression, because it is not forced to abstract from the details contained in the relations of the database.

1 Introduction

In science data analysis [3], we benefit from the luxury of precision of the data and the availability of domain knowledge, but the basic problems are the same as in other areas. Although there are not necessarily petabytes of data, the scientific problems tackled often are very hard, and the data are complex and highly structured. In case of structured data, "a flat-file form of the data is unlikely to be useful"[3]. Such data are more naturally represented in relations. This type of representation of learning problems has been studied in the field of *Inductive Logic Programming (ILP)*[9].

In this paper we present first results of an interdisciplinary project in scientific data mining. The goal of this project is to apply and develop ILP methods for learning structure-activity relationships (SARs) for carcinogenicity. SARs are models that predict the activity of chemicals in organisms from the molecular structure. Formally, the problem is to predict numbers from "relational structures" (such as labeled graphs), a problem also known as *Relational Regression* [2].

In the following, we present results from our analysis of carcinogenicity data derived from the carcinogenesis bioassay program, a long-term research study performed by the US National Institute of Environmental Health Sciences (NIEHS).

In the next section, we present previous work. Then, we describe the data used in this study. In the fourth section, we describe how we defined the learning problems and how we approached them. Subsequently, we present the results quantitatively and qualitatively. In the sixth section, we briefly sketch directions of further research.

2 Related work

Several SAR studies [5][6] using ILP methods have been published. Generally, the comparisons of ILP algorithms with other approaches (linear regression, neural networks) showed no statistically significant differences between the accuracies of the investigated methods. So the experiments demonstrated that the advantage afforded by comprehensible theories is not gained at the expense of predictive accuracy.

This work is also much in the spirit of studies comparing various methods (FOIL vs. PROGOL [14], propositional learning vs. relational learning [15]) in the domain of mutagenicity. However, the closest work in the literature is [7], which reports the application of the ILP algorithm PROGOL to one of the databases also used in this study.

3 Description of the Data

In this section we describe the datasets used in our experiments "as is", without the data engineering steps to define the learning problems. The next step dealing with the precise definition of the learning problems will be documented in the subsequent section.

Our starting point are two databases: The first one, provided by King and Srinivasan [7](abbreviated by K&S), contains information about the carcinogenicity of 330 compounds, as classified by the NIEHS. The second database, the Carcinogenic Potency Database (CPD)[4], is provided by Gold and co-workers, and contains information about bioassays including the species, the strain and the sex of the animals, and the route of administration of the compound. Figure 1 gives an overview of the relations in both databases.[1] Given these two databases, we were able to join them in order to obtain even more detailed data about carcinogenicity.

The chemicals in the K&S database are identified by the CAS registry number, which is a unique number for chemical compounds. They are described at the atom and bond level (see Figure 1). More precisely, the chemical structure of a compound is described by the atoms it contains (relation `atom`) and by a specification of the bonds between the atoms (relation `bond`). Atoms are characterized by the element, the atom type according to the molecular modelling package QUANTA, and the partial charges. The bonds are defined as

[1] This overview is slightly simplified.

relations between atoms, and also have types (according to QUANTA). Additionally, functional groups (such as benzene rings and methyl groups) and so-called "structural alerts" are represented. Structural alerts are structural properties which have been identified by domain experts as indicators of carcinogenicity. Finally, the outcome of tests for mutagenicity have been included. A chemical is said to be mutagenic if it damages DNA. This is known to be highly correlated with carcinogenicity.

The NIEHS has classified these chemicals as non-carcinogenic, equivocal and carcinogenic. The classification is based on statistically significant bioassays, additional biological data and domain knowledge of experts.

Relations in Carcinogenicity Database provided by King and Srinivasan (K&S)

```
niehs_assessment(CAS, NIEHS_Assessment). /* assessment of NIEHS as           */
                                         /* non-carcinogenic, equivocal or   */
                                         /* carcinogenic                     */

atom(CAS, AtomID, Element, AtomType,     /* description of chemicals         */
     PartialCharge).                     /* at the atom ...                  */

bond(CAS, Atom1ID, Atom2ID, BondType).   /* ... and bond level.              */

contains_benzene(CAS).                   /* Definitions of functional groups */
...                                      /* (9 relations)                    */
contains_halide10(CAS).                  /* Definitions of structural alerts */
...                                      /* (28 relations)                   */
ames(CAS).                               /* Positive test for mutagenicity   */
```

Relation in Carcinogenic Potency Database (CPD) provided by Gold et al.:

```
test(CAS, Species, Strain, Sex, Route, TumorigenicSite, PVal, TD50).
```

Figure 1: Relations in the databases to be analyzed.

Gold's CPD contains detailed descriptions of bioassays performed by the NIEHS and other organizations. Each example in the CPD consists of the compound used, the species, the strain and the sex of the animals, and the route of administration of the compound. The carcinogenicity of a substance may differ depending on the species, the strain, the sex and the route. For each bioassay, we know the statistical significance of the outcome ($PVal$) and the tumorigenic dose for 50% of the animals ($TD50$). The unit of the $TD50$ is mg/kg/day. Qualitatively, the effect of a compound is described by $PVal$: if the value indicates statistical significance, then the compound is carcinogenic. Quantitatively, the activity of a substance is described by $TD50$. E.g., if the $TD50$ is very low for the animals in a group, then the substance is highly carcinogenic.

For simplicity, we ignored another piece of information provided by the CPD: the "tumorigenic site" or target organ. This is the organ where tumors are found at the end

of a bioassay. We did not take into account this information in our first approach to the problem, since there are 76 different types of tumorigenic sites.

As is, the CPD contains no chemical descriptors. So, the effects of compounds cannot be analyzed without further information. Fortunately, the chemicals in the CPD are identified by the CAS registry number, a unique number for chemical compounds. Thus, we were able to combine the CPD with the K&S database, which is a rich source of information about chemicals. Through this combination, we have very detailed information about the bioassays as well as about the chemicals used. In Table 1, a quantitative overview of the three databases is given.

The joined DB is the basis for further investigations concerning species-specific, strain-specific, sex-specific and route-specific models for carcinogenicity. From a biological point of view, this is one of the novelties of our project.

	K&S DB	CPD	Joined DB
# examples	330	6823	6823
# tuples	21031	6823	27524
# relations	41	1	42
# features in propositional version of the data	38	6	42

Table 1: Quantitative overview of the databases.

4 Description of the Approach

In this section we describe our approach to analyzing the data. Our goal is to investigate the effects of increasing levels of detail in the data, both in the independent and the dependent variables. We investigated the following dimensions:

- Classification/regression
- Propositional learning/relational learning
- Chemicals as examples/tests as examples

Prerequisites for a quantitative comparison are that

1. the *problems* are *identical* (same examples).
2. the *same measures* are applied to the different formulations of the problem.

So a problem with a quantitative comparison stems from the different learning problems for the chemicals and for the tests. However, all combinations $\in \{classification, regression\} \times \{propositional, relational\}$ are quantitatively comparable

for both chemicals and tests, since classification accuracy can also be calculated for regression models. Besides, *qualitative* statements are possible in any case, e.g. about the features that play a role in the induced models. In the following, we will discuss the dimensions investigated in this study. A summary of the definitions of the learning problems can be found in Table 2.

Classification/Regression
Classification problems can be derived from regression problems by discretization of the dependent variable. If a regression problem is too hard, it may be easier to leave out the details and perform classification. This is what we did for tests as examples. The dependent variable here ($TD50$) is continuous, and we chose a simple discretization: if the value is bigger than the median, then the example belongs to class 1, otherwise it belongs to class 0.

The chemicals are classified as non-carcinogenic, equivocal or carcinogenic. This can be used directly for classification. However, the dependent variable is ordinal and we are not aware of learning algorithms for dependent variables that are ordinal. As a "work-around", we formulated a regression problem by mapping the NIEHS assessment to $\{-1, 0, 1\}$. Since we evaluate the results by the relative error, the scale does not play a role. We calculated the classification accuracy in the following way: if a regression rule predicts a negative value, then we predict "non-carcinogenic", else we predict "carcinogenic".

Propositional Learning/Relational Learning
To obtain a propositional version of the learning problems, we utilize the high-level chemical information from the K&S database. We define features for the existence of functional groups, structural alerts and the result of the ames test. In such a way, we obtain 38 features describing the compounds. These features can be used to propositionalize both learning problems (the chemicals and the tests as examples).

In the ILP setting, the examples are additionally defined at the atom and bond level. So obviously the description of the chemicals is more detailed than in the propositional setting. Vice versa, specifying features instead of the complete structural information can be viewed as abstraction.

Chemicals as Examples/Tests as Examples
For the chemicals, we are learning the NIEHS assessments (non-carcinogenic, equivocal or carcinogenic). The examples are exactly the same as in the database provided by King and Srinivasan. The data do not have to be transformed, as the learning problem is clear and well-defined.

For tests, the dependent variable is the tumorigenic dose ($TD50$). However, the value of $TD50$ is valid only if the substance *is* carcinogenic, i.e., if $PVal < 0.05$. This is only the case for 2897 out of the original 6823 examples. Unfortunately, there are conflicting $TD50$-values for some instances that are identical otherwise. This is due to the projection to the attributes that we selected (species, strain, sex, route), and could be resolved by including the tumorigenic site. Since we decided to ignore it here, we chose another solution

that makes sense from a biological point of view: from all conflicting instances, we defined the minimum $TD50$ as the value for $TD50$. So the dependent variable is the dose that is tumorigenic to at least one site for 50% of the animals. This transformation reduces the number of instances to 629.

Chemicals	Examples	As in original database of King and Srinivasan (330 examples)
	Dependent variable for classification	$\{$*non-carcinogenic, equivocal, carcinogenic*$\}$
	Dependent variable for regression	$\{-1, 0, 1\}$
Tests	Examples	Those with valid $TD50$ only (i.e., only those with significant outcome, $PVal < 0.05$) in case of conflicting $TD50$ values: one example with $min(TD50)$ (629 examples)
	Dependent variable for classification	$class = 1\ if\ TD50 > median(TD50)$ $class = 0\ otherwise$
	Dependent variable for regression	$TD50$

Table 2: Definitions of the learning problems in our study.

Algorithms Used
In Table 3, we present the algorithms used for our comparative study. For propositional classification, we use C4.5 [13], the well-known decision-tree package, and T2 [1], which induces 2-level decision trees. FOIL [11] and PROGOL[10][2] are state-of-the-art ILP algorithms. M5 [12] learns regression trees with linear regression models in the leaves. SRT [8] learns relational regression trees.

5 Experimental Results

First we discuss the quantitative results of the experiments (see Table 4 and Table 5).

For the *chemicals*, we did not observe big differences in accuracy except for Relational Regression, i.e., when all the available information is provided. SRT achieves (with statistical significance) the best accuracy for the chemicals. Comparing classification and regression, we observed a small improvement of the regression results over the classification results. However, the biggest difference was between using all the information, and using only part of the information.

[2]The experiment with PROGOL has been described in [7].

Examples	Formalism	Learning task	Algorithm(s)
Chemicals	Propositional	Classification	C4.5, T2
Chemicals	Propositional	Regression ($\{-1, 0, 1\}$)	M5
Chemicals	Relational	Classification	FOIL, PROGOL
Chemicals	Relational	Regression ($\{-1, 0, 1\}$)	SRT
Tests	Propositional	Classification	C4.5, T2
Tests	Propositional	Regression	M5
Tests	Relational	Classification	FOIL
Tests	Relational	Regression	SRT

Table 3: Algorithms applied to the two learning problems and different formulations of them.

Quantitatively, the results for the *tests* are in total contrast to the results for the chemicals. Here, the details provided do not seem to pay off: propositional classification algorithms are quantitatively superior to the rest. This may be due to the huge differences in the $TD50$, which may cause problems for regression algorithms. These differences are no longer visible if we perform classification instead. The bad performance of FOIL may be due to the multiple classifications which are counted as misclassifications.

Next we present the major discoveries and findings from our experiments. One of the authors is an expert in toxicology, and interprets the theories induced by the learning algorithms.

The rules found by C4.5 and FOIL are relatively lengthy, and do not provide many new insights. The rules reflect mostly what we specified as indicators of carcinogenicity, namely the ames tests and structural alerts. (Note that they also could have used the functional groups.) Some of the theories are quite accurate, but they are no real discoveries. An extreme example is the theory found by the T2 algorithm, which is quite accurate, but trivial, since it contains the ames test, and tests for structural alerts in the second level.

The theories found by C4.5 and FOIL are relatively easy to interpret for an expert, because the conditions in the theories relate to the structure of a compound. So an expert can easily draw some structures which are subsumed by a given rule. Besides, none of the rules found are *in contradiction* to "toxicological common-sense".

In contrast to C4.5 and FOIL, SRT often uses partial charges of atoms in its theories to discriminate the examples. In fact, the partial charges are the only information given about the chemical reactivity. It is a well-known fact that the lower the partial charge of an atom in a compound, the higher the probability of carcinogenicity. The drawback with tests for partial charges is that it is not possible to visualize the cases subsumed by the rule. However, it appears possible that the effect of a chemical mainly depends on partial charges, and not on the chemical structure.

The findings by C4.5 and FOIL on the one hand and by SRT on the other hand are in conflict. This raises the question whether the effect of chemical structure is stronger than the effect of partial charges, or the other way round. This interesting issue will be a point

of departure for further investigations.

The rules found by FOIL, C4.5 and SRT reveal that certain functional groups (methyl groups, benzene rings, rings of size 6) are, depending on the context, in some cases activating and in others deactivating. This pattern was found to make sense from a toxicological point of view.

Most of the qualitative insights were gained from the application of SRT. Several types of atoms have been found to be deactivating: atoms of type 8 according to QUANTA (e.g. "atoms with 2 double bonds on a 4 membered ring", and atoms of type 14 (e.g. "atoms with double bonds on a 4 membered ring with 3 double bonds"). Finally, if a sulfur atom is part of a compound, the compound is more likely to be not carcinogenic. This is amazing, as the existence of a sulfur atom in a compound is a very simple property. To the best of our knowledge, however, there is no toxicological evidence that contradicts this finding.

These observations can be made both in the application to chemicals and in the application to tests. Although these results might be real discoveries, additional analyses by independent domain experts are required to confirm them.

Approach	Algorithm	Classification Accuracy	Relative Error
Default		55.00%	–
Ames Test		63.00%	–
Propositional Classification	C4.5 prune	58.79%	–
	C4.5 rules	60.76%	–
	T2	65.00%	–
Propositional Regression	M5	69.93%	98.28%
Relational Classification	FOIL	25.15%	–
	PROGOL	63.00%	–
Relational Regression	SRT	72.46%	13.66%

Table 4: Quantitative results for chemicals obtained by 5-fold cross-validation.

Generally, SRT uses the same properties of compounds in both applications. Applied to tests, SRT additionally uses species, sex or route near the *leaves* of the trees, which results in huge differences in the predicted values for the resulting splits. This way we recognized that mice might have a much higher $TD50$ than rats. Closer examination revealed that out of all tests with $PVal < 0.05$ there are 1677 with rats, and only 1220 with mice. (In the overall database they are equally distributed.) From 189 cases which differ only in the species, there are 135 cases where the $TD50$ is higher for mice. On the average, the ratio $TD50_{mouse}/TD50_{rat}$ is 1.599. This confirms previous findings by Gold and co-workers that rats react more sensitively to carcinogens than mice.

Summing up the qualitative results, we observe that propositional learning algorithms made use of the knowledge about chemicals in the form of key attributes, but this only

Approach	Algorithm	Classification Accuracy	Relative Error
Default		50.00%	–
Propositional Classification	C4.5 prune	67.56%	–
	C4.5 rules	65.43%	–
	T2	59.86%	–
Propositional Regression	M5	56.67%	100.23%
Relational Classification	FOIL	31.39%	–
Relational Regression	SRT	56.19%	76.61%

Table 5: Quantitative results for tests obtained by 5-fold cross-validation.

yielded a high accuracy: mostly they were not capable of finding anything new. Most of the potential discoveries were obtained by an algorithm for Relational Regression, which utilizes all the available information.

6 Further Work and Conclusion

These are the next steps we are going to take:

- We will include the tumorigenic site in the description of the bioassays. Since there will be fewer conflicting $TD50$ values, more examples can be used for learning.
- We will design a machine learning algorithm for ordinal dependent variables. Like M5 or SRT, the algorithm will take into account the distance between actual and predicted values.
- Gold's CPD not only contains information from the NIEHS, but also from the literature. One of the next steps will be to determine the chemical structures of the compounds used in bioassays conducted by organizations other than the NIEHS.

In summary, we investigated the effects that various levels of detail and amounts of information have on the resulting hypotheses, both quantitatively and qualitatively. We applied relational and propositional machine learning algorithms to learning problems formulated as regression or as classification tasks. In addition, these experiments have been conducted with two learning problems which are at different levels of detail: first with chemicals as examples, second with tests as examples. Quantitatively, our experiments indicate that additional information not necessarily improves accuracy. Qualitatively, a number of potential discoveries have been made by the algorithm for Relational Regression, because it is not forced to abstract from the details contained in the relations of the database.

Acknowledgements

This research is partly sponsored by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung (FWF)* under grant number P10489-MAT. Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian Federal Ministry of Science and Transport. We would like to thank R. King, A. Srinivasan and L. Gold for providing the carcinogenicity data, and Gerhard Widmer for valuable discussions.

References

[1] P. Auer, W. Maass, and R. Holte. Theory and applications of agnostic pac-learning with small decision trees. In A. Prieditis and S. Russell, editors, *Proceedings of the 12th International Conference on Machine Learning (ML95)*. Morgan Kaufmann, 1995.

[2] S. Džeroski. *Numerical Constraints and Learnability in Inductive Logic Programming*. PhD thesis, University of Ljubljana, Ljubljana, Slovenija, 1995.

[3] U. Fayyad, D. Haussler, and P. Stolorz. KDD for science data analysis: Issues and examples. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 50–56, Menlo Park, CA, 1996. AAAI Press.

[4] L.S. Gold. Sixth plot of the carcinogenicity potency in the general literature 1989 to 1990 and by the national toxicology program 1990 to 1993. *Environmental Health Perspectives*, 103 (Suppl7):1–122, 1995.

[5] J.D. Hirst, R.D. King, and M.J.E. Sternberg. Quantitative structure-activity relationships by neural networks and inductive logic programming. the inhibition of dihydrofolate reductase by pyrimidines. *Journal of Computer-Aided Molecular Design*, 8:405–420, 1994.

[6] J.D. Hirst, R.D. King, and M.J.E. Sternberg. Quantitative structure-activity relationships by neural networks and inductive logic programming: The inhibition of dihydrofolate reductase by triazines. *Journal of Computer-Aided Molecular Design*, 8:421–432, 1994.

[7] R.D. King and A. Srinivasan. Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environmental Health Perspectives*, 1997.

[8] S. Kramer. Structural regression trees. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, 1996.

[9] S. Muggleton, editor. *Inductive Logic Programming*. Academic Press, London, U.K., 1992.

[10] S. Muggleton. Inverse Entailment and Progol. *New Generation Computing*, 13:245–286, 1995.

[11] J.R. Quinlan. Learning logical definitions from relations. *Machine Learning*, 5:239–266, 1990.

[12] J.R. Quinlan. Learning with continuous classes. In Sterling Adams, editor, *Proceedings AI'92*, pages 343–348, Singapore, 1992. World Scientific.

[13] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[14] A. Srinivasan, S. Muggleton, and R.D. King. Comparing the use of background knowledge by Inductive Logic Programming systems. In *Proceedings of the 5th International Workshop on Inductive Logic Programming (ILP-95)*, pages 199–230. Katholieke Universiteit Leuven, 1995.

[15] A. Srinivasan, S. Muggleton, R.D. King, and M. Sternberg. Theories for mutagenicity: a study of first-order and feature based induction. *Artificial Intelligence*, 85(1-2):277–299, 1996.