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

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

Bernhard Pfahringer
Computer Science Dept.,
University of Waikato
Private Bag 3105
Hamilton, New Zealand
bernhard@cs.waikato.ac.nz

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

Abstract

This paper presents first results of an interdisciplinary project in scientific data mining. We analyze data about the carcinogenicity of chemicals derived from the carcinogenesis bioassay program 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 can utilize all the information contained in the relations of the database as well as in the numerical dependent variable.

Introduction[1]

In science data analysis (Fayyad, Haussler, & Stolorz 1996), we benefit from the luxury of precision of the data and the availability of domain knowledge, but often scientific data are complex and highly structured. Therefore "a flat-file form of the data is unlikely to be useful"(Fayyad, Haussler, & Stolorz 1996). Such data are more naturally represented by relations as is done in *Inductive Logic Programming (ILP)*(Muggleton 1992).

In this paper we present first results of an interdisciplinary project in scientific data mining. The goal of this project is to develop and apply 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* (Džeroski 1995). The data used were derived from the carcinogenesis bioassay program, a long-term research study performed by the US National Institute of Environmental Health Sciences (NIEHS).

Related Work

Several SAR studies (e.g., (Hirst, King, & Sternberg 1994)) using ILP methods have been published. Generally, the comparisons of ILP algorithms with other approaches (linear regression, neural networks) showed no statistically significant differences in predictive accuracies, but ILP-generated theories tend to be more comprehensible. This work is also much in the spirit of studies comparing various methods (FOIL vs. Progol (Srinivasan, Muggleton, & King 1995), propositional learning vs. relational learning (Srinivasan *et al.* 1996)) in the domain of mutagenicity. (King & Srinivasan 1997) report on the application of Progol to one of the databases also used here.

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.

Our starting point are two databases: The first one (King & Srinivasan 1997)(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)(Gold 1995) contains information about bioassays including the species, the strain and the sex of the animals, and the route of administration of the compound.

The chemicals in the K&S database are identified by the (unique) CAS registry number. The compounds are described at the atom and bond level using two relations `atom` and `bond`. Atoms are characterized by the element, the atom type according to the molecular

	K&S DB	CPD	Joined DB
# examples	330	6823	6823
# tuples	21031	6823	27524
# relations	41	1	42
# features	38	6	42

Table 1: Quantitative overview of the databases.

modelling package QUANTA, and the partial charges. The bonds are defined as relations between atoms, and also have types (according to QUANTA). Additionally, the existence of functional groups (such as benzene rings and methyl groups) and of so-called "structural alerts" is represented. Finally, the outcome of the *ames test* for mutagenicity, which is highly correlated with carcinogenicity, has been included. The NIEHS has classified these chemicals as non-carcinogenic, equivocal and carcinogenic.

Gold's CPD contains detailed descriptions of bioassays performed by the NIEHS and other organizations. 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 body weight/day. The *effect* of a compound is described by $PVal$: if the value indicates statistical significance, then the compound is carcinogenic. 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.

As is, the CPD does not contain chemical descriptors, but the chemicals used are identified by the CAS registry number. So we joined the CPD with the K&S database via the CAS registry number, and obtained a database containing information about bioassays as well as information about the chemicals used. 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. Table 1 gives a quantitative overview of the three databases.

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. The dimensions investigated are chemicals vs. tests as examples, classification vs. regression, and propositional vs. relational learning. To allow for meaningful comparisons, the examples of the learning problems have to be the same, and the same measures of accuracy have to be applied. Therefore, results for chemicals and tests are not directly comparable. 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.

Chem.	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.

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.

Chemicals as Examples/Tests as Examples

For the first series of experiments, we used the K&S database with *chemicals* as examples. In the second series of experiments, the *tests* (i.e., individual bioassays) from the joined database were used as examples. We focused on a few attributes (species, strain, sex and route of administration), and on only those tests with carcinogenic substances ($PVal < 0.05$). We defined the dependent variable as the minimum $TD50$ of all examples that are identical otherwise, i.e., the dose that in at least one case is tumorigenic.

Classification/Regression

For the *chemicals*, we were learning the NIEHS assessments (non-carcinogenic, equivocal or carcinogenic). Basically, this dependent variable is ordinal, but it can also be used for classification. Since we are not aware of *relational* learning algorithms dealing with ordinal dependent variables, we formulated a regression problem by mapping the NIEHS assessment to $\{-1, 0, 1\}$. The scale does not play a role because we evaluated the results by the *relative error* (see e.g. (Quinlan 1992)). The *classification accuracy* was calculated in the following way: if a regression rule predicts a negative value, then we predict "non-carcinogenic", else we predict "carcinogenic".

For *tests* as examples, we derived the classification problem from the regression problem by discretization of the dependent variable. We chose a simple discretization of the tumorigenic dose ($TD50$): if the value is bigger than the median, then the example be-

Examples	Form.	Task	Algorithm(s)
Chemicals	Prop.	Class.	C4.5, T2
Chemicals	Prop.	Regr.	M5
Chemicals	Rel.	Class.	FOIL, Progol[2]
Chemicals	Rel.	Regr.	SRT
Tests	Prop.	Class.	C4.5, T2
Tests	Prop.	Regr.	M5
Tests	Rel.	Class.	FOIL
Tests	Rel.	Regr.	SRT

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

longs to class 1, otherwise it belongs to class 0.

Propositional Learning/Relational Learning
To obtain a propositional version of the learning problems, we utilized the high-level chemical information from the K&S database, including functional groups, structural alerts and the result of the ames test. The ILP setting includes additional low-level structural details about atoms and bonds.

Algorithms Used
In Table 3, we present the algorithms used for our comparative study. For propositional classification, we used C4.5 (Quinlan 1993), and T2 (Auer, Maass, & Holte 1995), which induces 2-level decision trees. FOIL (Quinlan 1990) and Progol (Muggleton 1995) are state-of-the-art ILP algorithms. M5 (Quinlan 1992) learns regression trees with linear regression models in the leaves. SRT (Kramer 1996) learns relational regression trees.

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. SRT achieves (with statistical significance) the best accuracy for the chemicals. Regression seems to perform slightly better than classification. However, the biggest difference was between using all the information, and using only part of the information.

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. 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.

[2] The experiment with Progol has been described in (King & Srinivasan 1997).

Approach	Algorithm	Accuracy	Rel. E.
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%	0.98
Relational Classification	FOIL	25.15%	–
	Progol	63.00%	–
Relational Regression	SRT	72.46%	0.14

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

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 these algorithms 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 commonsense".

In contrast to C4.5 and FOIL, SRT often uses partial charges of atoms in its theories to discriminate the examples. Therefore, it appears possible that the effect of a chemical depends more on partial charges than on chemical structure.[3] This interesting issue will be a point of departure for further investigations.

The rules found by C4.5, FOIL and SRT reveal that certain functional groups (methyl groups, benzene rings, 6 membered ring) are, depending on the context, in some cases activating and in others deactivating. This pattern is in accordance with present toxicological knowledge.

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"), atoms of type 14 (e.g. "atoms with double bonds on a 4 membered ring with 3 double bonds"), and sulfur atoms.

These observations can be made both in the application to chemicals and in the application to tests.

[3] Note that in some sense partial charges are caused by the chemical structure.

Approach	Algorithm	Accuracy	Rel. E.
Default		50.00%	–
Propositional Classification	C4.5 prune	67.56%	–
	C4.5 rules	65.43%	–
	T2	59.86%	–
Propositional Regression	M5	56.67%	1.23
Relational Classification	FOIL	31.39%	–
Relational Regression	SRT	56.19%	0.77

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

Although these results might be real discoveries, additional analyses by independent domain experts are required to confirm them.

In general, 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. This way we recognized that mice might have a much higher $TD50$ than rats. 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 towards carcinogens than mice.

Qualitatively, we observed that propositional learning algorithms utilize chemical knowledge only in the form of key attributes. Most of the potential discoveries were obtained by an algorithm for Relational Regression which utilizes all the available information.

Further Work and Conclusion

One of our next steps will be to include the tumorigenic site (i.e., the target organ) in the description of the bioassays. Since there will be fewer conflicting $TD50$ values, more examples can be used for learning.

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 can utilize all the information contained in the relations of the database as well as in the numerical dependent variable.

Acknowledgements This research was sponsored by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung (FWF)* under grant number P10489-MAT and under grant number Schrödingerstipendium Nr.J01269-MAT. Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian Federal Ministry of Science and Transport. Bernhard Pfahringer acknowledges the support of the ML group of the University of Waikato. We would like to thank R. King, A. Srinivasan and L. Gold for providing the carcinogenicity data, and G. Widmer for valuable discussions.

References

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

Džeroski, S. 1995. *Numerical Constraints and Learnability in Inductive Logic Programming*. Ph.D. Dissertation, University of Ljubljana, Ljubljana, Slovenija.

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

Gold, L. 1995. 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.

Hirst, J.; King, R.; and Sternberg, M. 1994. 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.

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

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

Muggleton, S., ed. 1992. *Inductive Logic Programming*. London, U.K.: Academic Press.

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

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

Quinlan, J. 1992. Learning with continuous classes. In Adams, S., ed., *Proceedings AI'92*, 343–348. Singapore: World Scientific.

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

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

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