(The Futility of) Trying to Predict Carcinogenicity of Chemical Compounds

Bernhard Pfahringer

University of Waikato, Hamilton, New Zealand,
bernhard@cs.waikato.ac.nz,
http://www.cs.waikato.ac.nz/~bernhard

Abstract. This paper describes my submission to one of the sub-problems formulated for the Predictive Toxicology Challenge 2001. The challenge is to predict the carcinogenicity of chemicals based on structural information only. I have only tackled such predictions for bioessays involving male rats. As we currently do not know the true predictions for the test-set, all we can say is that one of the models supplied by us seems to be optimal over some subrange of the ROC spectrum. The successful model uses a voting approach based on most of the sets of structural features made available by various other contestants as well as the organizers in an earlier phase of the Challenge. The WEKA Machine Learning workbench served as the core learning utility. Based on a preliminary examination of our submission we conclude that reliable prediction of carcinogenicity is still a far away goal.

1 Introduction

Environmentally-induced cancers are a serious health problem. All relevant data was collected by the US National Toxicology Program (NTP) which conducts standardized chemical bioassays. The Predictive Toxicology Challenge was initiated to motivate Machine Learning researcher into tackling this important prediction problem.

In the first stage various research groups generated sets of features that they thought would be relevant in predicting carcinogenicity. All these were made available from the PTC webpage [4]. We used six of those feature sets as well as one additional set we generated locally. The next section will discuss how single classifiers for each of those sets were first selected and then combined into a final two-level model. The section thereafter will then discuss the seven base-level models in some detail. Finally, we draw a few pessimistic conclusions.

2 Selecting and Voting Classifiers

Two very useful pieces of applied machine learning knowledge are:

- Using the right features makes all the difference.
- Ensembles of classifiers usually outperform single classifiers.

Following that advice we initially downloaded everything available, converted it all into an appropriate format for our tools[1], and did a couple of initial experiments trying to get "a feel" for the problem. These initial results were rather disappointing. All algorithms that we tried performed rather poorly. Basically two findings were made:

1. No single method improved prediction by more than 10 percentage points over the default prediction. 65% proved to be the limit.
2. Methods like logistic discriminant, linear support vector machines and Naive Bayes classification outperformed decision-tree or rule-based learning methods. Interestingly, all these methods can be thought of as alternative ways of computing a single hyperplane separating two classes. Like in text classification, these methods seem to do well in domains where there is an abundance of weak features, but no need or possiblity of forming reasonable logical combinations of features.

These rather disappointing results caused us to search for alternative feature sets, hoping that these might improve prediction. Most of these attempts were futile as well, only one attempt of capturing some three-dimensional knowledge performed on a par with the better ones of the supplied feature sets. Consequently, we decided to use the following seven feature sets to induce appropriate classifiers for each set and to simply vote these classifiers for deriving final predictions. We have excluded any feature set where the improvements over default accuracies were insignificant. The final feature sets are:

1. KULEUVEN features: a very small set of just 10 summary features. A logistic discriminant was constructed.
2. SENS features: 13 features representing linear sub-fragments that were derived in a class-sensitive manner. A logistic discriminant was constructed.
3. NTP features: a set of 24 physico-chemical descriptors of compounds. A bag of 10 logistic discriminants was constructed.
4. DRAGON features: a set of 839 features computed by the DRAGON program. A linear support vector machine was constructed.
5. FCSS codes: a set of 402 features supplied by the VINITI research group. A linear support vector machine was constructed.
6. BONDS3D features: 324 features capturing distances in 3D space between various pairs of bonds.
7. BCI fingerprints: 5212 features describing substructures. A Naive Bayes classifier was constructed.

[1] The WEKA machine learning workbench is available under the Gnu Public License and can be downloaded from `http://www.cs.waikato.ac.nz/~ml`

Voting these seven classifiers worked pretty well, adding a couple of percentage points lifting predictive accuracies to about 70% in cross-validation tests over the **training-set**.

So when receiving the final test-set for prediction we were quite surprised to find that almost no chemical was predicted to be carcinogenic. Obviously voting by simply summing the predicted class probabilities returned from each of the seven classifiers did not work as expected. Some inspection revealed that most of the time if the more likely class was "carcinogenic", its probability was only slightly larger than 0.5, so a few strong votes for "non-carcinogenic" with probabilities close to one could easily mask any indications of carcinogenicity. Consequently we turned to categorical predictions for the seven base level learners. Table 1 depicts the distribution of "carcinogenic" votes over the test-set compounds.

Table 1. Voting distribution: number of votes versus number of compounds receiving exactly that number of "carcinogenic" votes.

n_{votes}	$n_{compounds}$
0	72
1	55
2	37
3	11
4	6
5	2
6	2
7	0

Clearly, if we insist on a majority decision, only $2 + 2 + 6 = 10$ compounds would be classified as "carcinogenic". So we decided to set a threshold such that the predicted distribution would be similar to the distribution found in the training-set, which is the only reasonable reference point available given the lack of any further information. The exact same distribution could be generated by a cutoff somewhere between one and two votes, so we submitted actually three sets of predictions:

- Model M1: predicts "carcinogenic" if at least 1 of the seven classifiers says so, which is a rather cautious approach trying to minimize false-negatives as far as possible. M1 predicts "carcinogenic" for 113 of the 185 compounds in the test-set.
- Model M2: predicts "carcinogenic" if at least 2 of the seven classifiers say so. M2 predicts "carcinogenic" for only 58 compounds.
- Model M3: uses probabilities and adjusts the cutoff to closely mimic the training-set distribution. M3 predicts "carcinogenic" for 83 compounds.

Model M1 is the best of these three models according to the Challenge organizers and is also optimal compared to all submissions for some range of error cost as determined by the organizers using ROC curves (a good introduction to ROC curves is given in [5]).

To judge the individual contribution made by each classifier one can look at the number of compounds being predicted as "carcinogenic" by each classifier in total, as well as the number of compounds that are being predicted "carcinogenic" uniquely by one classifier. Table 2 summarizes these numbers.

Table 2. Classifier contributions: total and unique counts of "carcinogenic" predictions for each feature set.

Feature Set	Total	Unique
FCSS	66	24
NTP	35	7
DRAGON	28	6
BONDS3D	27	8
BCI	22	3
SENS	20	6
KULEUVEN	10	1

Clearly, the method contributing most is the classifiers built over the FCSS features. There does not seem to be much of a difference between the remaining ones. The KULEUVEN entry has to be taken with a pinch of salt as we were not able to utilize the full set of features, just an extremely limited subset. Also, we need to be careful with these kind of judgements, as we currently do not know how many of these predictions are actually correct.

3 Individual classifiers

In this section we will describe the individual classifiers produced for the seven sets of features. As all the learning methods used here basically are just estimating a separating hyperplane, we depict each model as a kind of regression equation, where we would predict "carcinogenic" if the outcome of the equation is positive, and "non-carcinogenic" if the outcome is negative. For the smaller sets of features we will give the full equations, for the larger feature sets this is infeasible and pointless anyway, so we will only try to extract the most important features.

How can one determine the importance of features in a regression equation? First of all the sign of the coefficient of each feature indicates the general tendency: positive coefficients indicate features correlates positively with "carcinogenicity", negative coefficients indicate the opposite. Of course this is only true for positive feature values, but most of our feature values are positive. As for judging the magnitude of influence, the coefficients are not sufficient unfortunately, as the ranges of values of different features may vastly differ. A useful heuristic for estimating importance is provided by the absolute value of the product of the coefficient and the mean of the respective feature. This value is the average contribution to classification made by a particular feature. We will be using this heuristic to sort small feature sets, and to extract the more important features from larger sets.

In the following subsections we describe the seven base-level models. For all features sets we have generally performed extensive experiments comparing decision-tree and rule-based methods, logistic discriminants, support vector machines, and Naive Bayes as well as bagging and boosting. Due to excessive runtimes, for larger data-sets some methods proved infeasible, e.g. computing logistic discriminants is least of the order of $O(a^2)$ where a is the number of attributes. Clearly we cannot apply logistic discriminant to datasets like the BCI fingerprint set featuring 5212 attributes per compound. The classifiers chosen and reported below were the best-performing ones in cross-validations over the training-set.

3.1 KULEUVEN features

This is the smallest set of all. Due to last-minute problems encountered in transforming test-set features into the ARFF format mandated by WEKA, we finally chose a very small set of just 10 summary features. Table 3 depicts the full equation.

Table 3. The regression equation for the KULEUVEN feature set.

Coefficient	$Coeff * Mean$	Feature
-0.0552	-1.410	N_ATOMS: number of atoms
0.0006	0.127	D_FG: functional group distance
0.4372	0.119	N_ARO2N2: two aromatic rings with N atoms each
0.4372	0.119	N_ARO2N: two aromatic rings, one N at least
0.0082	0.070	N_FG: total number of functional groups
-0.2573	-0.048	MAX_DELTA_CHARGE: maximal charge difference
0.0001	0.023	WEIGHT: molecular weight
-0.0065	-0.010	N_RING: number of rings
0.0500	0.008	MAX_DISTANCE: maximal distance between two atoms
-0.0047	-0.004	N_AROMATIC: number of aromatic rings
0.7635		Intercept

Clearly, the features N_ARO2N2 and N_ARO2N are collinear, and so one should have been dropped. For a more detailed description of the features please see the documentation supplied by the KULEUVEN group.

3.2 Sens features

These 13 features represent linear sub-fragments of compounds that were derived in a class-sensitive manner by the Freiburg group. Table 4 depicts the full equation.

Table 4. The regression equation for the Sens feature set.

Coefficient	$Coeff * Mean$	Feature
6.6183	0.8657	Br
-5.6828	-0.4461	Br-C
14.6016	0.3402	Br-C-C-Br
0.7179	0.0709	C-c:c:c:c:c:c-N
0.5704	0.0547	N-c:c-O
0.5704	0.0547	N-c:c:c:c:c:c-O
-0.1893	-0.0160	Br-C-C
0.0092	0.0015	c:c-c:c:c:c-N
0.0092	0.0015	c:c:c-c:c:c:c-N
0.0092	0.0015	c:c:c:c-c:c:c:c-N
0.0092	0.0015	c:c:c:c:c-c:c:c:c-N
0.0184	0.0015	c:c:c:c:c:c-c:c:c:c-N
0.0184	0.0015	N-c:c:c:c-c:c:c:c-N
-0.4892		Intercept

Clearly again we see quite a few collinear features which should have been dropped at closer inspection. Still, the equation seems reasonable in qualitative terms, according to my amateur knowledge of chemistry. The presence of Bromium acts as a strong indicator for carcinogenicity, as does the presence of both oxygen and nitrogen connected to an aromatic ring structure.

3.3 NTP features

This feature set consists of 24 physico-chemical descriptors of compounds supplied again by the Freiburg group. The actual classifier induced for this feature set is a bag of 10 logistic discriminants. To save space, we only depict one of the ten regression equations in Table 5. Clearly the coefficients vary between the various bags, but the ranking of the features is mostly the same, so reproducing just one equation should be sufficient.

Obviously the first seven features heavily dominate the final decision.

3.4 Dragon features

This feature set comprises 839 features as computed by the Dragon program. We just depict the top 15 features in Table 6. More information about the Dragon program and generated feature set is available from the PTC webpage [4].

Table 5. The first of ten similar regression equations for the NTP feature set.

Coefficient	$Coeff * Mean$	Feature
-859.0440	-8036.27	IONIZATION_POTENTIAL
-479.3035	4477.94	HOMO
-386.3687	1865.19	ELECTRONEGATIVITY
-186.6642	1685.59	HOMO_LUMO
-2.4965	-915.47	TOTAL_ACCESS
2.5028	722.77	NON_POLAR_ACCESS
2.5204	196.38	POLAR_ACCESS
-0.4108	-9.06	POLARIZA
0.0023	-5.54	TOTAL_ENERGY
-0.0002	2.93	ELECTRONIC_ENERGY
0.0124	2.83	MOLECULAR_WEIGHT
0.0358	2.81	PERC_NONPOLAR
7.1313	-2.22	LUMO
0.0005	1.90	STABIL
-0.0430	-1.14	DIPOLE
0.5135	1.08	LOGP
0.3726	0.98	POINT_CHG_DIPOLE
0.0523	0.47	LARGEST_INNERATOMIC_DISTANCE
-0.0089	0.20	STRAIN
-0.0008	0.17	DELTAHF
0.1665	0.14	HYBRID_DIPOLE
0.0025	-0.13	HEAT_OF_FORMATION
1.9928	0.13	CHARGE
-19.4291	-0.11	RADICAL
-3.1908		Intercept

3.5 FCSS codes

This feature set comprises 402 features expressed using the FCSS language submitted by the VINITI research group from Russia. Again, just the top 15 features are depicted in Table 7.

For an interpretation of the attributes selected please refer to the document describing FCSS available from the PTC webpage [4].

3.6 Bonds3D features

This set of 324 features is a naive attempt to capture some kind of 3D information about compounds. Basically distances in 3D space between various pairs of bonds were computed. For each pair of types of bonds we count how many such pairs are present and what their minimal and their maximal 3D distance is in every compound. If some pair is not present, we use zero for both the minimal and maximal distance, which is a reasonable null value in regression equations. Actually this set of features is a subset comprising only those pairs, where one

Table 6. The top 15 features of the DRAGON feature set.

Coefficient	$Coeff * Mean$	Feature
-0.4318	-12829.8	A106: TPCM total multiple path count
0.2608	3832.5	A142: SRW10 self-returning walk count order 10
0.1609	895.6	A112: GMTIV Gutman MTI valence vertex degrees
0.2288	539.4	A140: SRW08 self-returning walk count order 8
0.0972	352.7	A111: SMTIV Schultz MTI valence vertex degrees
-0.1178	-285.5	A364: W3D 3D-Wiener index
0.0513	225.0	A78: IDMT total info-content distance magnitude
-0.4319	-142.0	A108: PCD diff of multiple path counts to path counts
0.0380	87.1	A120: SMTI Schultz Molecular topological index
0.1452	83.9	A141: SRW09 self-returning walk count order 9
0.0326	72.5	A100: GMTI Gutman Molecular topological index
0.1591	71.1	A138: SRW06 self-returning walk count order 6
-0.1139	-62.3	A105: TPC total path count
-0.1083	-44.1	A366: DDI distance-distance index
-0.0927	-38.2	A382: Mor01u 3D-MORSE signal 01
2.027163		Intercept

bond is either a single bond between two carbon atoms, or a single bond between a carbon atom and a hydrogen atom. There is no particular justification for using this subset except for the fact that its cross-validation performance was superior to all other subsets tested. Again, we just depict the top 15 features in Table 8.

An attribute “n_C1C-H1N” represents the total count of pairs of bonds of type “C-C” and of type “N-H” that are present in a compound, where as an attribute like “max_C1C-C1H” measures the maximal 3D distance for any pair of bonds of type “C-C” and type “C-H”.

The numbers and especially the signs in Table 8 seem rather counter-intuitive, e.g. the presence of “Cl” seems to decrease the likelihood of carcinogenicity.

3.7 BCI fingerprints

This is the most extensive of all feature sets comprising 5212 BCI fingerprints. BCI fingerprints have been supplied by George Cowan of Pfizer Global Research and Development. As the top rank is dominated by negative coefficients here, we have decided to have both the top 10 negatives as well as the top 10 positive terms depicted in Table 9.

We have simply extracted the definition for each BCI fingerprint from the supplied dictionary, separating multiple entries with commas. Please consult the BCI description for more information on the meaning of these descriptors (again available from the PTC webpage [4]).

Table 7. The top 15 features of the FCSS feature set.

Coefficient	$Coeff * Mean$	Feature
-0.952	-0.571	AC-6-06
1.133	0.238	AC-200331
-1.188	-0.099	AC-201131
-1.188	-0.075	AC-1300241
0.516	0.068	AC-1200331
0.483	0.057	AC-1201411
1.175	0.053	AC-500051
-1.323	-0.044	AC-1200241
-0.563	-0.039	AC-66-10
0.446	0.037	AC-1301331
-0.302	-0.035	AC-1201131
-0.288	-0.035	AC-3100331
0.600	0.034	AC-2400331
0.589	0.034	AC-500331
-1.369	-0.033	AC-264021
0.18823		Intercept

4 Conclusions and further directions

Given all the information and figures above reporting rather meager performance gains overall it is obvious that we currently cannot predict carcinogenicity of new compounds reliably. Investigating alternative approaches to strict yes/no decisions like predicting rankings for sets of compounds or even trying to predict $LD50$ dosages directly, might be promising.

Additionally, we need to be careful when interpreting coefficients and signs thereof in a regression context. Signs can be wrong for various reasons that are discussed in [3]. An attempt to counter these reasons is described in [1]. But even variables with large standardized coefficients are not necessarily the most important ones, as work on socalled random forests [2] has shown.

But the major current shortcoming – we suppose – is simply a lack of data given the diversity of compounds encountered. A few hundred data points just does not seem to suffice. We are confident that methods similar to those described here should deliver good prediction rates when supplied with larger datasets describing at least a few thousand compounds. The methods described above would scale to datasets of such sizes. Unfortunately, such data currently is proprietary knowledge of chemical companies only.

References

1. Pazzani, M. J., Bay, S. D.: The Independent Sign Bias: Gaining Insight from Multiple Linear Regression, in Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society, 1999, `http://www.ics.uci.edu/ sbay/#publications`.

Table 8. The top 15 features of the BONDS3D feature set.

Coefficient	$Coeff * Mean$	Feature
-0.473	-39.06	n_C1H-C1H
0.710	24.60	n_C1H-C2C
-0.247	-23.81	n_C1C-C1H
0.702	17.88	n_C1C-C2C
-0.906	-16.82	n_C1H-C1N
0.528	11.36	n_C1H-C1O
0.553	6.70	n_C1C-C1O
-0.904	-5.58	max_C1C-C1H
0.955	4.79	n_C1C-H1N
-0.441	-4.59	n_C1C-C1N
-0.597	-4.15	max_C1H-C1H
-0.809	-4.12	max_C1C-C1C
0.657	3.03	n_C1C-C2O
-0.505	-2.77	n_C1CL-C1H
-0.445	-2.49	n_C1C-C1CL
0.531305		Intercept

2. Breiman L.: Random forests, random features, Technical Report 567, University of Berkeley, 1999, `http://www.stat.berkeley.edu/users/breiman/`.
3. Mullet G.: Why Regression Coefficients have the Wrong Sign, in Journal of Quality Technology, 8(3), 1976.
4. Helma C., King R.D., Kramer S., Srinivasan A.: The Predictive Toxicology Challenge for 2000-2001, `http://www.informatik.uni-freiburg.de/~ml/ptc/`.
5. Provost F., Fawcett T.: Robust Classification for Imprecise Environments, in Machine Learning Journal, 42(3), 2001.

Table 9. The top 10 negative and positive features of the BCI fingerprints feature set.

Coefficient	$Coeff * Mean$	Feature
-1.194	-1.145	AttrCor-1244: AAAAacAA
-0.591	-0.519	AttrCor-13: AAAAcsAA
-0.588	-0.441	AttrCor-158: AS4Aaa4Aaa4Aaa4A, ASC aaC aaC aaC
-0.967	-0.420	AttrCor-122: ASAAcsAAcsAAcsAA
-0.553	-0.404	AttrCor-190: AA4Aaa4Aaa4A, AAC aaC aaC
-0.497	-0.349	AttrCor-1209: ASAAarAAarAAarAAarAAacAA
-0.502	-0.341	AttrCor-2984: AS4Aaa4Aaa4Aaa4Aaa4Aaa4A
-0.504	-0.339	AttrCor-317: RCAAarAAarAAarAAarAAarAAar
-0.469	-0.330	AttrCor-2672: ASAAarAAarAAarAAacAA
-0.476	-0.322	AttrCor-295: ASC aaC aaC aaC aaC aaC
0.600	0.133	AttrCor-487: AP4A2 2 6 4A2 2, APC 2 2 6 C 2 2
0.531	0.120	AttrCor-3096: APAA2 3 7 AA2 2
0.388	0.118	AttrCor-4355: AS5Acs4Arn4Arn4Arn4A, ASN csC rnC rnC rnC
0.498	0.112	AttrCor-786: AP4A2 3 7 4A2 2, APC 2 3 7 C 2 2
0.464	0.111	AttrCor-1565: APAA2 2 6 AA2 2
1.231	0.110	AttrCor-1814: acC arC arC arC arC arC acN
0.360	0.109	AttrCor-381: AS5Acs4Arn4Arn4Arn4Arn4A, ASN csC rnC rnC rnC rnC, AS5Acs4Arn4Arn4Arn4Arn4Arn4A, ASN csC rnC rnC rnC rnC rnC
0.336	0.104	AttrCor-1733: AS5Acs4Arn4Arn4A, ASN csC rnC rnC
1.121	0.103	AttrCor-4272: AS6Aac4Aar4Aar4Aar4Aar4Aar4Aac5A
0.333	0.102	AttrCor-288: AA4Arn4Arn4Acs5A, AAC rnC rnC csN
-0.241		Intercept