First order models for the Predictive Toxicology Challenge 2001

Hendrik Blockeel, Kurt Driessens, Nico Jacobs,
Raymond Kosala, Stefan Raeymaekers, Jan Ramon,
Jan Struyf, Wim Van Laer, Sofie Verbaeten

Katholieke Universiteit Leuven, Department of Computer Science
Celestijnenlaan 200A, B-3001 Leuven, Belgium

August 24, 2001

Abstract

This paper discusses the "Leuven" submission[1] to the Predictive Toxicology Challenge 2001. A brief account of some preparatory work is given, followed by a more detailed description of the approach that in the end led to the submitted model, and of the model itself. Both the approach and the model are evaluated from a data mining point of view, and a number of conclusions are drawn.

1 Introduction

The Predictive Toxicology Challenge 2000–2001 concerns the use of knowledge discovery methods in the field of carcinogenesis prediction. More specifically, the task is to build models that (a) from the description of a compound used for biochemical assays on rodents (divided into male and female rats and mice), with high confidence predict the outcome of the assay; and (b) are sufficiently interpretable for humans to be able to learn from them.

Many tasks in the biochemical domain concern structural knowledge discovery, in which the available data are of a complex nature (typically involving structural descriptions, such as the structure of a molecule). Many classical data mining algorithms do not work with such structural descriptions but work in a so-called attribute-value format, where each data element is described using the same fixed set of attributes and the domain of each attribute contains atomic values only (i.e., numbers or symbols, not sets or graphs).

[1] More specifically, the submission by the Machine Learning group of the Laboratory for Declarative Languages and Artificial Intelligence of the Computer Science Department of the Katholieke Universiteit Leuven.

In such contexts one can distinguish two kinds of approaches: one is to transform the description into attribute-value format (so-called propositionalisation approaches, see e.g. [7]), after which standard data mining techniques can be employed; the other is to use a knowledge discovery approach that handles structural data directly (such as inductive logic programming [9]). One could say that the former approach distinguishes two phases (first construction of attributes from structural data, then construction of models based on these attributes; these separate phases were explicitly present in the schedule of the challenge) while the latter approach mixes these phases.

This paper discusses the submission of the DTAI lab (Declarative Languages and Artificial Intelligence) of the Katholieke Universiteit Leuven, and in doing so attempts to draw some conclusions that could be useful for the knowledge discovery community in general. While both approaches mentioned above were explored to some extent, the focus is on the second approach (inductive logic programming).

What this paper does *not* provide, is an evaluation of the submission from a biochemical point of view. While such an evaluation is of course important, it can be done only by domain experts, which none of the authors are.

Beyond this point we assume familiarity with the Predictive Toxicology Challenge (see `www.informatik.uni-freiburg.de/~ml/ptc/` for full details) and with some data mining and machine learning terminology ([13] is an excellent introduction).

The rest of the paper is organised as follows. In Section 2 we give an overview of preliminary experiments that were conducted before selecting a particular approach. In Section 3 we discuss in some more detail the first order models that we induced and from which our actually submitted models were derived; the latter are discussed in Section 4. Section 5 concludes the paper.

2 Preparatory Experiments

2.1 Tools and Data Sets

The following data mining tools and techniques were used in these experiments:

- Weka [13], a general-purpose data mining tool from which we used : **J48**, a decision tree induction [5, 11] system based on Quinlan's C4.5 [12]; and several **feature selection** algorithms [13]
- ACE [3], an inductive logic programming [9] tool, in which we used : **Tilde**, an algorithm that induces first order decision trees [2], and is based on C4.5 [12]; and **Warmr**, an algorithm for first order pattern discovery [6] based on Apriori [1]

The tools were used in combination with various data sets available on the PTC website (`www.informatik.uni-freiburg.de/~ml/ptc/`).

2.2 Building Predictors

Our attempts at building predictive models can be grouped along a number of dimensions, which we now briefly discuss.

2.2.1 Structural Versus Propositional Methods

From the data mining point of view, one can distinguish techniques that attempt to construct relevant features themselves, and techniques that assume all relevant features of the data to be present already.

In our case, the first approach comprised the construction of structural patterns that hopefully would correlate with some assay result of the molecule. This mostly focused on patterns described as a set of functional groups that are at certain distances of each other or are bound to each other through a certain chain of atoms; and discovery of simpler substructures such as chains of 2 or 3 atoms, or simply the occurrence of a certain kind of atom in the molecule. These experiments mainly involved the use of the ILP-systems Warmr and Tilde.

The second approach in our case boiled down to building predictors based on the propositional feature sets made available on the PTC website. These experiments mainly involved the Weka tool (J48, feature selection). Warmr was also used on these features in an attempt to find frequently occurring combinations of features. Correlations between propositional features were computed in an attempt to reduce the number of features we should look at, but little reduction was obtained in this way.

2.2.2 All-at-once Versus One-at-a-time

Given that there are four different target variables (classification of assays on male rats, female rats, male mice and female mice), the probably most straightforward approach is to build a model for each target variable separately (i.e., four models in total). We call this the *one-at-a-time* approach.

It is possible to build a single model that works for all classes at once, however (*all-at-once* models). We distinguish two approaches here:

- *Predict a vector of four variables.* Here, the target variable is in fact a vector with four components; each component correspond to one of the original target variables. The model is built so as to maximise performance on all four components at once.
- Predict a single variable, using the sex and type of animal as extra attributes in the data. In other words, instead of giving as input a description of the molecule and getting as output predictions for four assays, the input is now the molecule plus the type of assay, and the output is the prediction for that assay.

The second approach can easily be achieved by deriving from each example with four target values, four examples (one for each assay) and then just running any data mining system. Note that missing class information is very naturally

handled in this way: when some of the target values are missing for an example it will just give rise to fewer derived examples. An interesting feature of this approach is that the induced model decides for itself in what cases it should distinguish between male and female animals, or between mice and rats; also, the influence of the animal's sex and type on the assay result is made explicit in this way.

2.2.3 Classification Versus Regression

The results of assays does not always seem to be clear-cut; a distinction is made between "clear evidence", "some evidence", etc. A straightforward way to model this is by assigning a number to the result, indicating to what the degree the result is thrustworthy. The problem is then turned into a regression (numerical prediction) problem. A disadvantage of this approach is that our assignment of numbers to symbolic results is ad hoc. Nevertheless, both classification and regression were tried in our preliminary experiments.

2.3 Conclusions from Preliminary Experiments

Of all the different directions mentioned above, none gave really satisfying results.

One of the main problems with the propositional features, as available on the PTC web page, is that there are thousands of them, while there are only a few hundred data elements; there is a considerable risk of overfitting in such a situation, and with several experiments it was unclear to us to what extent overfitting was happening. Moreover, in those cases where a reasonably trustworthy model was built, its predictive accuracy was usually disappointing. It seems very difficult to obtain a predictive accuracy above the default accuracy (i.e., the accuracy obtained by always predicting the majority class).

The latter remark may seem of little relevance, given that it was known in advance that the evaluation would be based on ROC analysis [10], not predictive accuracy. Still, predictive accuracy is usually the first measure one sees when building a model with, e.g., J48 or Tilde. Our perception is that this often has a discouraging effect on people: when exploring a certain direction of experiments, there is a strong tendency to abandon this direction when models derived in this exhibit consistently bad performance. In hindsight, it would probably have been a good idea to consistently use ROC diagrams to evaluate models during the experimentation.

3 Some First Order Models

3.1 Regression and Classification Trees

As none of the approaches followed for the preliminary experiments clearly outperformed the other, we eventually focused on prediction using first order decision trees, the technique we are most familiar with.

Our submissions are based on models built using ACE/Tilde, both in regression and in classification mode. The data set consisted of only the structural descriptions of the molecules as built by the Leuven group (see PTC webpage `www.informatik.uni-freiburg.de/~ml/ptc/`).

More specifically, we adopted two all-at-once approaches to build the final models:

- Tilde in regression mode, predicting vectors in 4-D where results were encoded (ad hoc) as 1 (CE), 0.5 (SE), 0.75 (P); 0 (E, EE, IS); -0.5 (N) and -1.0 (NE). The background knowledge consisted of atoms and bonds only, no functional groups.

- Tilde in classification mode. Results were encoded into two classes as follows: the CE and P classes were merged into "positive", the N and NE into "negative"; other examples were ignored. Here the background knowledge included structural features (functional groups and their counts), as well as the sex and type of the animal.

These have been chosen more or less randomly from the many different approaches explored during the preliminary experimentation phase.

The model built using the regression approach is shown in Figure 1. The classification model contains 18 nodes, which we found rather complex to show here. An interesting observation is that the sex and type attributes were not used in this tree, i.e., they were considered of little interest by the learner.

3.2 ROC Analysis of Models

The results were analysed using ROC diagrams produced on the basis of a tenfold crossvalidation (equivocal examples were filtered out from the test set). Note that regression models can be turned into classifiers by adding a threshold to them (prediction above threshold = positive, below = negative); by varying this threshold different points on a ROC diagram are obtained, so that the whole regression model is represented as a curve in the ROC diagram. While a classification tree in itself is a single point, one can order its predictions according to the confidence with which they are made (depending on the size and purity of the leaf that makes the prediction); by imposing a confidence threshold on positive predictions a curve can be constructed in exactly the same way as for regression models.

Figure 2 shows the ROC curves obtained using the regression approach. A first remarkable fact is that these curves are not convex; the reason for this is unclear. Nevertheless, by constructing the convex hulls around these curves we get a better impression of the usefulness of the models. Besides the ROC curves, the $y = x$ line is drawn, which corresponds to random prediction, as well as a second line that corresponds to classifiers that achieve the default predictive accuracy ($y = \frac{p_-}{p_+}x$ with p_+ and p_- the current proportion of positives/negatives in the data). This line is relatively steep due to the fact that there are more negatives than positives in the data; it confirms that our models have serious

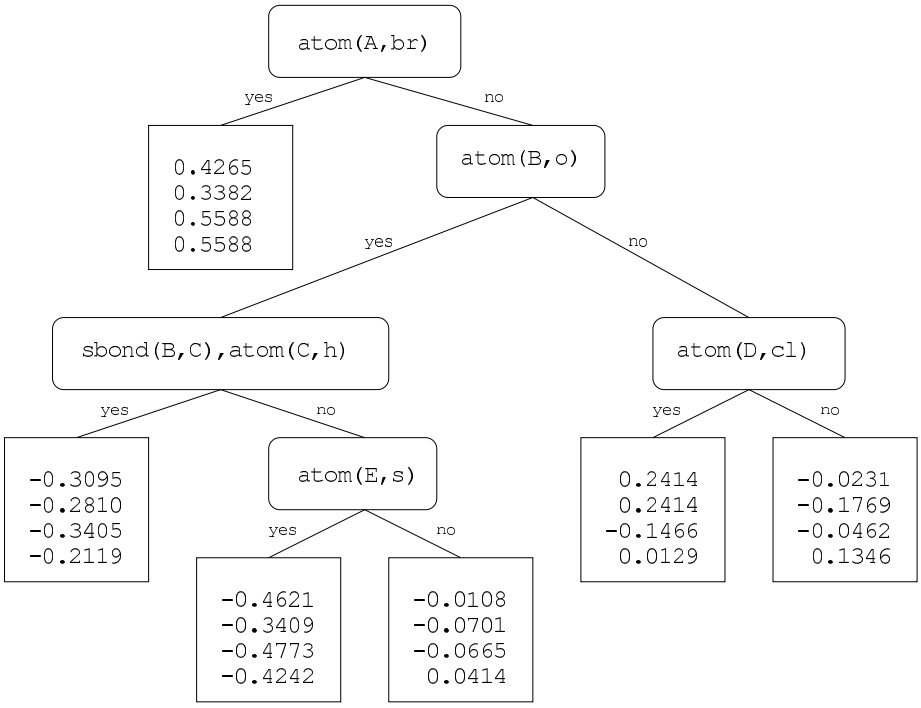

Figure 1: 4-D regression model. The numbers indicate to what extent an assay result is positive, respectively for female mice, male mice, female rats, and male rats. Capitals A through E are variables referring to atoms in the molecule; for instance, *atom(A,br)* indicates there exists an atom A of element bromine (elements symbols are lower-case here). *sbond* indicates a bond between two atoms.

difficulty achieving an accuracy above default, and shows that those classifiers that do achieve such an accuracy tend to focus on the lower left corner of the ROC diagram.

With the classification approach, no valleys in the ROC curves occurred: the actual curves were close to the convex hull. We therefore immediately present the convex hulls of the approach and compare them with those obtained using the regression approach: see Figure 3.

These diagrams suggest that our regression approach works better at the upper end of the diagram (low false positive costs). In the extreme lower left corner both approaches seem approximately equivalent, but the classification approach quickly gains an advantage over the regression approach.

These results beg the question whether the differences between both approaches are related to the regression / classification issue or to the different sets of background knowledge. We therefore ran the regression experiment on the richer background knowledge from which the classifier was derived. The

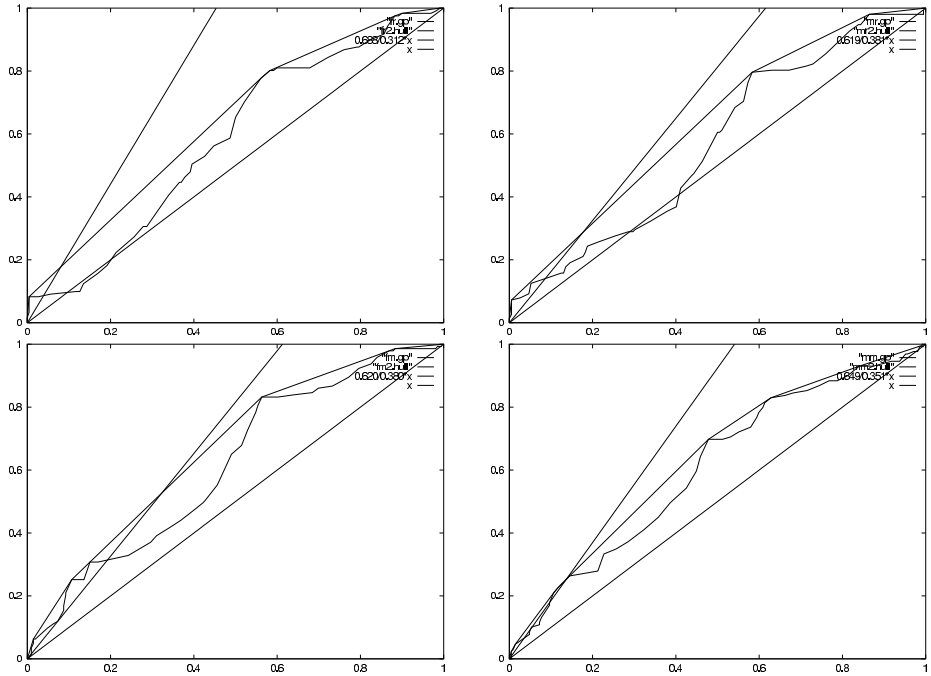

Figure 2: ROC curves for the first order 4-D regression model. Left: female; right: male; upper: rats; lower: mice. Besides ROC curves and their convex hull, lines corresponding to "random prediction" and "obtaining default accuracy" are shown.

ROC curves obtained here were very close to the classification curves in Figure 3 (but slightly below them) which suggests that the difference is mainly due to the change in background knowledge.

4 Our Submissions

4.1 The Submitted Models

Given that participants can submit up to three models, we have decided to submit models that focus on different areas of the ROC diagram.

- Model 1: the regression model with a low threshold vector; this model is suitable in environments with very low cost of false positives.
- Model 2: looking at the ROC diagrams, there are two "peaks" that are good candidates. Since it is difficult to make a choice here, we have looked at the total area under the curve. For rats the winner then seems to be the classification model, whereas for mice it is the regression model. We have

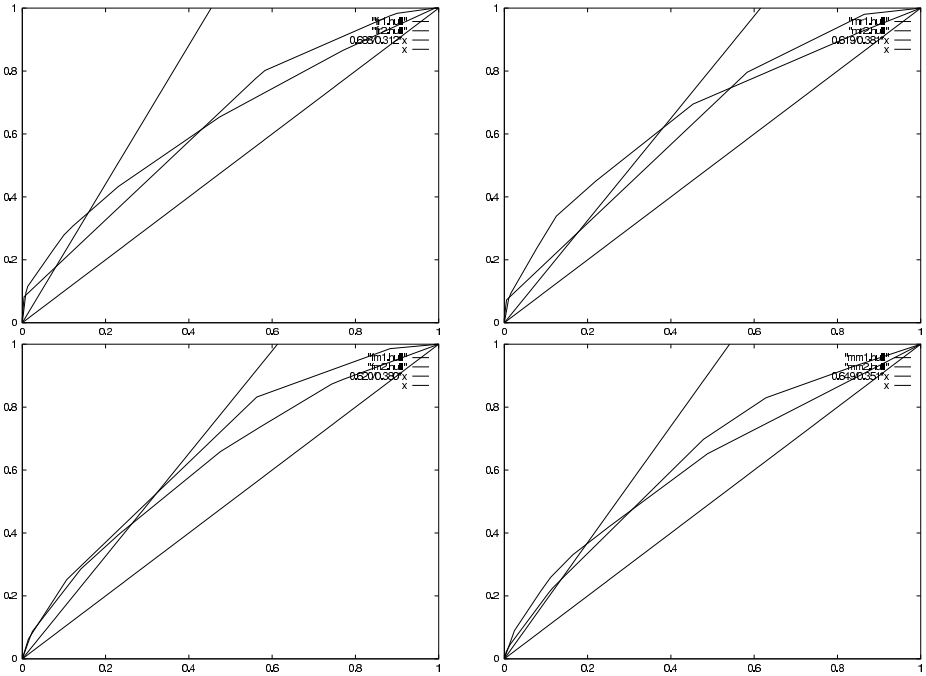

Figure 3: ROC curves for the first order classification models. Left: female; right: male; upper: rats; lower: mice.

decided to create a hybrid model: for rats, use the classification model; for mice, use the regression model.

- Model 3: the regression model with a high threshold vector; this model is suitable in environments with very high cost of false positives.

Except for Model 2 for rats, these models are extremely simple; an English translation of them is shown in Figure 4. The estimated performance of these models as measured using a tenfold crossvalidation is listed in the form of contingency tables in Table 1.

The reason for choosing "extreme" cases for Models 1 and 3 is that such models could be used for screening molecules. Assuming that our estimates are reliable, a "negative" prediction by 1 or a "positive" prediction by 3 is very likely to be correct[2], hence assays for these molecules could be given lower priority. Our results suggest that in this way, some 5% of truly positive compounds and some 10% of truly negative compounds would be selected.

[2]The exact percentage of correct predictions depends of course on the actual class distribution.

Model 1: (most "paranoid" model – low cost of false positives)
Consider a compound *inactive* if it contains oxygen and sulphur and no bromine and no hydrogen bound to oxygen, otherwise *active*.

Model 2: (neutral model for mice)
Consider a compound *active* if at least one of the following conditions is true:
- contains bromine
- does not contain oxygen
- contains oxygen but no sulphur and no hydrogen bound to oxygen
Otherwise consider it *inactive*.

Model 3: (most "permissive" model – high cost of false positives)
Consider a compound *active* if it contains bromine, otherwise *inactive*.

Figure 4: Explanation of the simpler submitted models.

4.2 Predictions on the Test Set

When making predictions on the test set, we noticed that the number of "positive" predictions on the test set was much lower than expected, and this could not be explained by a mere shift in the class distribution. For instance, our Model 2 for mice scored approximately FP=0.6 and TP=0.8, so for any mix of positive and negative molecules we would expect a proportion of positive predictions between these two numbers. In fact, only about 40% of the test instances were predicted positive. This indicates a shift in the distribution of some properties used to make the predictions, and we expect this to further deteriorate our test set predictions.

Note that this is in general a problem for those machine learning techniques that build models by selecting the most relevant features. Models that use all features at once, such as lazy learners, probably have an advantage in this respect: they cannot be misled into focusing on features that afterwards turn out to be irrelevant for the test set (unless, of course, they rely heavily on feature weighting or selection themselves).

It should be noted that test examples, even if unclassified, still contain information that may be useful for learning (a point which has recently been made in other contexts as well [8]).

5 Conclusions

As mentioned before, the authors of this paper are unable to draw conclusions with respect to the biochemical relevance of the results. From the data mining point of view, a number of lessons have been learnt that seem worth mentioning:

- While many data mining tools report the predictive accuracy of a model, one should be careful not to be (mis)guided too much by this measure.

Model 1

(R)	pos	neg	
ce,p,se	10	111	121
n,ne	1	229	230
	11	340	351

(R)	pos	neg	
ce,p,se	10	111	152
n,ne	1	229	192
	11	340	344

(R)	pos	neg	
ce,p,se	9	134	143
n,ne	3	203	206
	12	337	349

(R)	pos	neg	
ce,p,se	6	123	129
n,ne	3	204	207
	9	327	336

	TP	FP
FR	0.0826	0.0043
MR	0.0724	0.0052
FM	0.0629	0.0146
MM	0.0465	0.0145

Model 2

(CL)	pos	neg	
ce,p	32	72	104
n,ne	28	202	230
	60	274	334

(CL)	pos	neg	
ce,p	40	78	118
n,ne	24	168	192
	64	246	310

(R)	pos	neg	
ce,p,se	36	107	143
n,ne	22	184	206
	58	291	349

(R)	pos	neg	
ce,p,se	34	95	129
n,ne	30	177	207
	64	272	336

	TP	FP
FR	0.3077	0.1217
MR	0.3400	0.1250
FM	0.2517	0.1068
MM	0.2636	0.1449

Model 3

(R)	pos	neg	
ce,p,se	118	3	121
n,ne	203	27	230
	321	30	351

(R)	pos	neg	
ce,p,se	149	3	152
n,ne	166	26	192
	315	29	344

(R)	pos	neg	
ce,p,se	141	2	143
n,ne	182	24	206
	323	26	349

(R)	pos	neg	
ce,p,se	120	9	129
n,ne	174	33	207
	294	42	336

	TP	FP
FR	0.9752	0.8826
MR	0.9803	0.8646
FM	0.9860	0.8835
MM	0.9302	0.8406

Table 1: Contingency tables and true/false positive ratios for the submitted models, estimated using tenfold crossvalidation. From above to below: female rats, male rats, female mice, male mice; R=regression model, CL = classification model.

- When the distribution of attribute values may shift over time, it is a bad idea to use models that select a minimal set of most relevant features to base their prediction on. Models that take all features into account (e.g., Naive Bayes), or approaches where predictions are made without explicitly building a model (the so-called transductive setting) have advantages in this respect. Note that this is to some extent incompatible with the wish for simple, interpretable theories.

- Regression models do not yield a single point but a curve in the ROC diagram, which can be turned into a set of classifiers. Also from a single classifier, multiple classifiers can be generated that are tuned towards different misclassification cost ratios; this is what was done to produce Models 1 and 3. This methodology was not exploited to its full potential here because only three classifiers could be submitted for each type of assay, but results of other experiments [4] suggests that improvements to the ROC convex hull might be obtainable in this way.

Acknowledgements

HB is a post-doctoral fellow, JS and SV research assistants, of the Fund for Scientific Research of Flanders (FWO-Vlaanderen). RK and WVL are supported by project G.0246.99 ("Query languages for data mining") of the Fund for Scientific Research of Flanders. JR is supported by the Flemish Institute for the Advancement of Scientific and Technological Research in Industry (IWT).

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. The MIT Press, 1996.

[2] H. Blockeel and L. De Raedt. Top-down induction of first order logical decision trees. *Artificial Intelligence*, 101(1-2):285–297, June 1998.

[3] H. Blockeel, B. Demoen, L. Dehaspe, G. Janssens, J. Ramon, and H. Vandecasteele. Executing query packs in ILP. In J. Cussens and A. Frisch, editors, *Proceedings of the 10th International Conference in Inductive Logic Programming*, volume 1866 of *Lecture Notes in Artificial Intelligence*, pages 60–77, London, UK, July 2000. Springer.

[4] H. Blockeel and J. Struyf. Frankenstein classifiers : Some experiments on the Sisyphus data set. In *Proceedings of IDDM-01 - Workshop on Integration of Data Mining, Decision Support, and Meta-Learning*, Freiburg, Germany, 2001.

[5] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[6] L. Dehaspe and H. Toivonen. Discovery of frequent datalog patterns. *Data Mining and Knowledge Discovery*, 3(1):7–36, 1999.

[7] S. Kramer, B. Pfahringer, and C. Helma. Stochastic propositionalisation of non-determinate background knowledge. In *Proceedings of the Eighth International Conference on Inductive Logic Programming*, volume 1446 of *Lecture Notes in Artificial Intelligence*, pages 80–94. Springer-Verlag, 1998.

[8] T.M. Mitchell. The role of unlabeled data in supervised learning. In *Proceedings of the Sixth International Colloquium on Cognitiive Science*, San Sebastian, Spain, 1999.

[9] S. Muggleton and L. De Raedt. Inductive logic programming : Theory and methods. *Journal of Logic Programming*, 19,20:629–679, 1994.

[10] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48. AAAI Press, 1998.

[11] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[12] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann series in machine learning. Morgan Kaufmann, 1993.

[13] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.