Pin-Pointing Concept Descriptions

Cecilia Sönströd, Ulf Johansson
School of Business and Informatics
University of Borås, Sweden
cecilia.sonstrod, ulf.johansson@hb.se

Henrik Boström
Department of Computer and Systems Sciences
Stockholm University, Sweden
henrik.bostrom@dsv.su.se

Ulf Norinder
AstraZeneca R&D
Södertälje, Sweden
ulf.norinder@astrazeneca.com

***Abstract*—In this study, the task of obtaining accurate and comprehensible concept descriptions of a specific set of production instances has been investigated. The suggested method, inspired by rule extraction and transductive learning, uses a highly accurate opaque model, called an oracle, to coach construction of transparent decision list models. The decision list algorithms evaluated are JRip and four different variants of Chipper, a technique specifically developed for concept description. Using 40 real-world data sets from the drug discovery domain, the results show that employing an oracle coach to label the production data resulted in significantly more accurate and smaller models for almost all techniques. Furthermore, augmenting normal training data with production data labeled by the oracle also led to significant increases in predictive performance, but with a slight increase in model size. Of the techniques evaluated, normal Chipper optimizing FOIL's information gain and allowing conjunctive rules was clearly the best. The overall conclusion is that oracle coaching works very well for concept description.**

***Index Terms*—Data mining, Concept description, Decision lists**

I. Introduction

One of the most intensive areas of research within the pharmaceutical industry today is to collect and analyze data on absorption, distribution, metabolism, excretion and toxicity (ADMET) [1]. The overall purpose is to learn how various compounds interact with the human body in order to guide drug development projects in the search for promising compounds. Specifically, compounds unsuitable as drug candidates, e.g., due to toxicity, should be detected as early as possible.

Currently, a commonly adopted approach is to leverage large libraries of chemicals (acquired or synthesized to meet stringent quality criteria) and use high-throughput screening (HTS) to test for biological activity. Promising compounds found in this way become the focus for continued research, which typically leads to further synthesis and screening. Synthesis and screening processes are, however, often time consuming and costly, making it desirable to estimate the biological activity, as well as ADMET properties, before synthesis. When computer software is used for this initial modeling, the procedure is referred to as *in silico* modeling [1]. If successful, in silico modeling saves much time and investments by excluding non-promising compounds, thus allowing earlier focus on drug candidates with high potential.

Obviously, in silico modeling can be performed by using powerful machine learning techniques, such as artificial neural networks (ANNs) or support vector machines (SVMs), which produce opaque models with high predictive performance. However, domain experts (e.g. computational chemists) also have a need for comprehensible models, to help gain insights into which attributes of compounds that are of importance for possessing certain biological properties. An obvious criterion for a model to be comprehensible is that it is transparent, but one could also argue that it should be relatively small, include the most important relationships and describe these succinctly. These comprehensible models should, of course, also be as accurate as possible. It is a well-known fact that there exists a trade-off between comprehensibility and accuracy, in that techniques that produce transparent models generally obtain worse predictive performance than, e.g., ANNs or SVMs.

A common situation in early phases of drug development is that a large number of compounds with known values, obtained by HTS, for a certain type of biological activity are available, but that sets of newly acquired or synthesized compounds should be evaluated by using in silico modeling. Naturally, the targeted biological activity for these new compounds needs to be estimated in an accurate and comprehensible way. In data mining terms, this can be identified as an instance of the problem type *concept description*, as described by the CRISP-DM data mining framework [2]. Simply put, the overall purpose of concept description is to gain insights. So, rather than focusing only on producing models with high predictive accuracy, obtaining adequate descriptions of the most important relationships in the data is essential. Another important feature of concept description that is mentioned in [2] is that models need not capture the whole dataset, i.e., partial models are perfectly acceptable.

The main idea in this paper is to utilize existing libraries of compounds as training data to build high-performance opaque models, and then use these models to coach the building of comprehensible models to obtain concept descriptions for new sets of data. This idea of using coaching to obtain high accuracy on specific (production) instances is obviously similar to transductive learning, but we explicitly focus on situations where the final model must be transparent, leading to a process where a stronger model coaches a weaker.

More generally, this study targets the specific situation where production input vectors are already determined and available when building the concept description model. In real-world applications, this is actually a very common situation, meaning that the model is explicitly built for the task and production set at hand. In this exact situation, the unlabeled production instances, i.e., the very same instances that later

978-1-4244-6587-3/10/$25.00 É2010 IEEE

will be used for actual predictions and descriptions, could also be used for building the model. The purpose of this paper is thus to study how coaching techniques can be used for this kind of concept description modeling, using the drug discovery domain described above as a motivating example.

II. Background and related work

When supervised learning is used to build a predictive classification model, a function mapping data instances to the target class variable is learned from a set of labeled instances; i.e., the class value is known for each training instance. Semi-supervised learning, on the other hand, uses both labeled and unlabeled data for the model construction. In some situations, typically when labeled data is hard or costly to obtain, but unlabeled data is relatively cheap and easily accessible, semi-supervised models will outperform models built using labeled data only. In such situations, the number of labeled instances used is normally much smaller than the number of unlabeled instances. Naturally, several fundamental methods based on semi-supervised learning exist; for a good survey see [3]. Several standard approaches first build a model using only labeled data, and then use this model to label unlabeled instances, thus creating more training instances. This process may be repeated over several iterations, and there are many variations, but the final classifier is normally trained using a majority of initially unlabeled instances.

Transductive learning is, in a strict sense, the opposite of inductive learning; i.e., the inference is directly from training instances to production instances. Or, put in another way, transductive learning omits the model building, thus solving a less general problem than standard predictive classification. Using this definition, however, even transductive support vector machines become inductive learners since they partition the entire input space. With this in mind, the term transductive learning is often used to characterize an algorithm utilizing both labeled and unlabeled data to obtain high accuracy on *specific* production instances; see e.g. [4].

The main inspiration for the method suggested in this paper comes, however, from the field of *rule extraction*. Rule extraction, which is the process of generating a transparent model based on a corresponding opaque model, has been used mainly to understand and analyze ANN models; for a good survey see [5].

Black-box rule extraction algorithms approach rule extraction as a learning task, where the target concept is the function originally learned by the opaque model. A training instance is, consequently, the original input vector and the corresponding prediction from the opaque model. Most black-box rule extraction algorithms maximize *fidelity* i.e., the number of identical classifications.

Most importantly, for this application, the opaque model is also a very accurate model of the function between input and output, so it could be used to label novel instances with unknown target values, as they become available. Naturally, these newly labeled instances could then be used as learning examples.

We have previously shown that the procedure of letting a highly-accurate model (called an oracle) act as a coach can be successfully applied to both rule extraction [6] and standard tree or rule induction [7].

Within the field of machine learning, there are many techniques producing transparent models, typically either decision trees or ordered rule sets, also called *decision lists*. Unfortunately, very few techniques specifically aim for comprehensible models, and even fewer contain explicit means for controlling the accuracy vs. comprehensibility trade-off. In [8], we introduced the decision list algorithm Chipper, which is tailor-made for concept description. The basic idea in Chipper is to, in every step, search for the rule that classifies the maximum number of instances using a split on one attribute. For continuous attributes, this means a single comparison using a relational operator. For nominal attributes, this is translated to a set of instances having identical values for that attribute.

Two main parameters, called *ignore* and *stop*, are used to control the rule generation process. The *ignore* parameter specifies the misclassification rate (as percentage of remaining instances) that is acceptable for each rule and can have different values for each output class. The motivation for the *ignore* parameter is that it can be used to view the data set at different levels of detail, with higher values prioritizing the really broad discriminating features of data items and with low values trying to capture more specific rules. The *stop* parameter specifies the proportion of all instances that should be covered by rules before formulating a default rule and terminating. The motivation for this parameter is that it can be used to find only the most general relationships in the data, instead of trying to find rules to cover particular instances in the training data. This parameter is also motivated by the observation in CRISP-DM that concept description models may well be partial. In effect, these two parameters control the level of granularity of the decision list.

III. Method

The first subsection introduces the drug discovery data sets used. The second subsection describes the different techniques evaluated, and the third subsection, finally, gives the experimental details.

A. Datasets

The data sets used are from the medicinal chemistry domain and consist of 8 different data sets, from the study of Bruce et al. [9], originally used by Sutherland [10]. In the study by Bruce et al., the two attribute sets *2.5D* and *Frags.* were used; here a further three attribute sets are available, thus bringing the total number of data sets to 40 (8 data sets, with 5 different attribute sets). All the data sets used in this study will be made publicly available on the INFUSIS homepage[1]. Of the five different attribute sets, two describe physical-chemical properties (e.g. number of atoms or types of bonds) of the compounds and the other three molecular fingerprints. The characteristics for each combination of data set and attribute

[1] www.his.se/infusis

set are summarized in Table I below, where *Inst.* means number of instances. As can be seen in the table, the data sets have relatively few instances, but a substantial number of attributes, especially for the fingerprint attribute sets.

TABLE I
DATA SET CHARACTERISTICS - NUMBER OF ATTRIBUTES

		Physical-chemical		Fingerprint		
Name	Inst.	2.5D	AZ Desc.	Frags.	ecfi1024	sign12
ACE	114	56	196	1024	1024	332
AchE	111	63	196	774	1024	211
BZR	163	75	196	832	1024	450
COX2	332	74	196	660	1024	573
DHFR	397	70	196	951	1024	487
GPB	66	70	196	692	1024	239
THER	76	64	196	575	1024	251
THR	88	66	196	527	1024	220

The motivation for the use of these data sets is that they represent the data mining situation described above, where concept descriptions of production data are desired, since relationships found are of interest to domain experts and can also be used to guide further search for promising molecules.

All data sets concern biological activity for inhibitor compounds. The continuous numerical values for activity (pIC50 for the first five data sets and pKi for the last three) in the study by Sutherland et al. were transformed by Bruce et al. into two categorical classes (active and inactive), using the median activity value as a threshold between the two classes to create a 50/50 split of active/inactive observations, since each data set showed a uniform distribution of activity values. The data sets concern the following biological targets:

- A set of 114 angiotensin converting enzyme (ACE) inhibitors. Activities of these compounds spread over a wide range, with pIC50 values ranging from 2.1 to 9.9.
- A set of 111 acetylcholinesterase (AchE) inhibitors, with pIC50 values ranging from 4.3 to 9.5.
- A set of 163 ligands for the benzodiazepine receptor (BZR), with activity (pIC50) values ranging from 5.5 to 8.9.
- A set of 322 cyclooxygenase-2 (COX2) inhibitors, having pIC50 values that range from 4.0 to 9.0.
- A set of 397 dihydrofolate reductase inhibitors (DHFR), with pIC50 values for rat liver enzyme ranging from 3.3 to 9.8.
- A set of 66 inhibitors of glycogen phosphorylase b (GPB), with pKi values ranging from 1.3 to 6.8.
- A set of 76 thermolysin inhibitors (THERM), having pKi values ranging from 0.5 to 10.2.
- A set of 88 thrombin inhibitors (THR) with pKi values ranging from 4.4 to 8.5.

B. Techniques

As mentioned in the introduction, the purpose of this study was to evaluate whether the use of a high-accuracy opaque model (serving as a coaching oracle) may be beneficial for creating concept descriptions using decision list algorithms. More specifically, rule sets induced directly from training data only were compared to rule sets built using different combinations of training data and oracle data, i.e., the production instances with corresponding ensemble predictions as target values. For simplicity, and to allow easy replication of the experiments, the Weka data mining workbench [11] was used for all experiments. In this study, large Random Forest [12] ensemble models were used as oracles. For producing the decision lists, JRip and Chipper were used. The motivation for including JRip is that it represents the state-of-the-art rule inducer RIPPER [13], providing a benchmark to compare Chipper's concept description performance against. The JRip parameter settings were left at their default values, but with Chipper, four different parameter settings were tried, varying granularity by using different *ignore* and *stop* parameter values and also varying the rule selection criteria between accuracy and optimizing FOIL's information gain. In addition, two different representation languages were also tried. The difference was whether conjunctions were allowed or not when building the rules. The four Chipper settings used were:

- *Chip1-A*: Here Chipper was set to produce very detailed decision lists, by using a 1% *ignore* and a 99% *stop* value, and also using rule selection based on accuracy.
- *Chip1-F5*: The same *ignore* and *stop* values as above, but rule selection based on FOIL's information gain and allowing up to 5 conjunctions in each rule.
- *Chip2-A*: Here Chipper was used with its normal prediction settings, using a 5% *ignore* and a 95% *stop* value, and using rule selection based on accuracy.
- *Chip2-F5*: The same *ignore* and *stop* values as above, but rule selection based on FOIL's information gain and allowing up to 5 conjunctions in each rule.

C. Experiments

For the experimentation, 4-fold cross-validation was used. The reason for not using the more standard value of ten folds was the fact that the use of only four folds results in what we believe to be a more representative proportion between training and production data. On each fold, the Random Forest ensemble was first trained, using training data only. This ensemble (the oracle) was then applied to the production instances, producing production predictions. This resulted in two different data sets:

- The *training* data: this is the original training data set, i.e., original input vectors with corresponding correct target values.
- The *oracle* data: this is the production instances with corresponding ensemble predictions as target values.

In the experimentation, training and oracle date were evaluated as training data for the decision list algorithms, both separately and together. In practice, this means that JRip and Chipper optimized different combinations of training accuracy and production fidelity towards the oracle. More specifically, each experiment had the following different setup:

- **Experiment 1** - Induction (I): Standard induction using

original training data only. This maximizes training accuracy.

- **Experiment 2** - Explanation (X): Uses only oracle data, i.e., maximizes fidelity towards the oracle on production data.
- **Experiment 3** - Indanation[2] (IX): Uses training data and oracle data, i.e., will maximize training accuracy and oracle fidelity on the production data.

Table II below summarizes the different setups.

TABLE II
SETUPS

	Data		Maximizes	
Setup	Training	Oracle	Training Accuracy	Production Fidelity
I	x		x	
X		x		x
IX	x	x	x	x

Evaluation was performed using the three measures accuracy, area under the ROC curve (AUC) and size. The size measure is the number of conditions in a rule set. In Figures 1 and 2 below, sample JRip and Chipper rule set are shown.

(HBAsum $\geq$ 6.505) and (MaxNegChargeGH $\geq$ -0.9987) and (MMSPEC_VDW_EP_N_AREA $\geq$ 0.392) $\rightarrow$ activity = 0 (72.0/3.0)

(ClogP $\leq$ 3.618) and (VDW_HB_D_AREA $\leq$ 24.26) $\rightarrow$ activity = 0 (28.0/4.0)

(MM_RNCS $\leq$ 2.826) and (M3M $\leq$ 4.299) and (HOMO $\leq$ -0.5016) $\rightarrow$ activity = 0 (17.0/1.0)

(VDW HB A AREA $\geq$ 51.45) and (MM SAS EP N AREA $\geq$ 259) and (MM VDW EP P VAR $\leq$ 40.75) $\rightarrow$ activity = 0 (9.0/0.0)

(SAS POL AREA $\leq$ 97.62) and (AverPosCharge GM $\geq$ 0.0707) and (Polarizability $\leq$ 38.42) $\rightarrow$ activity = 0 (16.0/3.0)

(MM HACA $\geq$ 0.5937) $\rightarrow$ activity = 0 (8.0/1.0)

$\rightarrow$ activity = 1 (172.0/23.0)

Number of Rules : 7

Fig. 1. Sample JRip rule

IF Chi6p $\leq$ 1.119 THEN 0 [56/5]
IF Chi5p $\geq$ 2.322 THEN 1 [50/2]
DEFAULT: 0 [8/4]

Number of Rules : 3

Fig. 2. Sample Chipper rule

IV. RESULTS

In Table III below, the accuracy results from Experiment 1, i.e., normal rule induction, are shown aggregated over the five different attribute sets. Since there are some fundamental differences in characteristics between the chemical-physical (*2.5D* and *AZ Descriptors*) and fingerprint (*Frags.*, *sign12* and *ecfi*) data sets, results are also aggregated over these two groups. The chemical-physical group thus consists of 16 data sets and the fingerprint group of 24 data sets. The total mean ranks are, of course, over all 40 data sets.

TABLE III
EXPERIMENT 1 - AGGREGATED ACCURACY RESULTS

	JRip	Chip1		Chip2	
		A	F5	A	F5
2.5D	.678	.681	.678	.679	.681
AZ Desc.	.732	.714	.712	.735	.720
Mean Acc Chem-Phys	**.705**	**.697**	**.695**	**.707**	**.700**
Mean Rank Chem-Phys	**2.94**	**2.81**	**3.00**	**2.94**	**3.13**
Frags.	.680	.712	.708	.707	.734
sign12	.695	.705	.705	.705	.716
ecfi1024	.689	.704	.705	.720	.712
Mean Acc Fingerpr.	**.688**	**.707**	**.706**	**.711**	**.721**
Mean Rank Fingerpr.	**3.75**	**2.83**	**2.88**	**3.08**	**2.29**
Total Mean	.695	.703	.702	.709	.713
Total Mean Rank	3.43	2.83	2.93	3.03	2.63

As can be seen in the table, when performing normal rule induction there are only small differences in accuracies between the techniques on the chemical-physical data sets and this is reflected by the mean ranks. On the fingerprint data sets, there is a slight advantage in average accuracies for Chip2, but this is not reflected in the mean ranks. To establish whether there are any statistically significant differences between the techniques for the fingerprint data sets, we follow the procedure recommended by Demšar [14] for comparing several classifiers over a number of data sets, i.e., a Friedman test [15], followed by a Nemenyi post-hoc test [16]. The result of these tests should however be treated with some care in this study as it is not obvious that sets of compounds represented by different feature sets can be considered to be independently selected datasets. Hence, this, and subsequent statistical tests employed, should be seen as approximate tests. With five classifiers and 24 data sets, the critical distance (for $\alpha = 0.05$) is 1.25, so based on these tests, the only statistically significant difference is that Chip2-F5 obtained significantly higher accuracy than Jrip on the fingerprint data sets.

To illustrate the origin of the aggregated results, Table IV below shows the detailed accuracy results for the *AZ Descriptor* attribute set. There are, of course, another four sets of accuracy results, one for each attribute set.

TABLE IV
EXPERIMENT 1 - ACCURACY FOR *AZ Descriptors* ATTRIBUTE SET

	JRip	Chip1		Chip2	
		A	F5	A	F5
ACE	.846	.790	.791	.835	.824
AchE	.626	.637	.635	.634	.614
BZR	.698	.666	.655	.724	.698
COX2	.681	.700	.674	.669	.653
DHFR	.760	.744	.743	.712	.733
GPB	.593	.624	.617	.647	.657
THR	.637	.642	.665	.647	.663
THERM	.585	.646	.646	.590	.602
Mean	**.678**	**.681**	**.678**	**.679**	**.681**
Mean Rank	**3.13**	**2.50**	**2.88**	**3.13**	**3.25**

[2]This name, combining the terms induction and explanation, is of course made-up

Here it is clearly seen that the small differences in average differences do not mean that all techniques perform similarly on all data sets, indeed there are quite large differences at times. However, as shown by the mean ranks, no technique consistently performs better than the others.

As can be seen in Table V below, the picture for AUC is similar to the accuracy results, but with a more pronounced advantage for Chip2.

TABLE V
RESULTS EXPERIMENT 1 - AUC

	JRip	Chip1		Chip2	
		A	F5	A	F5
2.5D	.685	.675	.679	.688	.699
AZ Desc.	.740	.710	.713	.743	.736
Mean AUC Chem-Phys	**.713**	**.693**	**.696**	**.715**	**.718**
Mean Rank Chem-Phys	**2.63**	**2.94**	**2.88**	**2.31**	**2.00**
Frags.	.689	.721	.719	.738	.740
sign12	.701	.710	.709	.722	.744
ecfi1024	.698	.704	.705	.740	.736
Mean AUC Fingerpr.	**.696**	**.712**	**.711**	**.737**	**.740**
Mean Rank Fingerpr.	**3.83**	**3.08**	**3.17**	**1.88**	**1.63**
Total Mean	.703	.704	.705	.728	.731
Total Mean Rank	3.35	3.03	3.05	2.05	1.78

Notably, for AUC there is a difference between Chip2 and the other techniques also on the chemical-physical data sets, even if the differences are not statistically significant. Looking at the mean ranks for the fingerprint data set, however, there are significant differences between both Chip2 variants and JRip and Chip1-F5. When considering the total ranks, the critical distance (for $\alpha = 0.05$) with 40 data sets is 0.96, so both Chip2 variants performed significantly better than the three other techniques.

Turning to size, Table VI below shows the aggregated size results for Experiment 1.

TABLE VI
RESULTS EXPERIMENT 1 - SIZE

	JRip	Chip1		Chip2	
		A	F5	A	F5
2.5D	8.4	14.4	17.1	5.9	14.4
AZ Desc.	7.7	12.6	15.0	5.2	12.0
Mean Size Chem-Phys	**8.1**	**13.5**	**16.0**	**5.6**	**13.2**
Mean Rank Chem-Phys	**2.06**	**3.69**	**4.44**	**1.00**	**3.69**
Frags.	7.0	35.2	38.0	12.6	22.9
sign12	8.2	32.3	35.5	10.0	20.5
ecfi1024	8.5	21.0	24.1	8.2	17.4
Mean Size Fingerpr.	**7.9**	**29.5**	**32.5**	**10.3**	**20.3**
Mean Rank Fingerpr.	**1.29**	**4.42**	**4.46**	**1.71**	**3.00**
Total Mean	8.0	23.1	25.9	8.4	17.4
Total Mean Rank	1.60	4.13	4.45	1.43	3.28

Here, both JRip and the Chip2-A consistently produced quite compact rule sets. There is also a notable difference between Chip2-A and Chip2-F5, in that the latter has about twice the average size, which is mainly due to the use of conjunctions. Chip1, with its high demand on rule accuracy and small amount of instances in the default rule, often produced very large models.

Summarizing Experiment 1, then, Chip2-F5 obtained the best predictive performance, measured both as accuracy and as AUC, but this was achieved at the expense of comprehensibility, with models about twice the size of Chip2-A. JRip produced small models, but performed worst on both accuracy and AUC. Overall, a good compromise between predictive performance and compact models was achieved by Chip2-A. Looking at the results on predictive performance (both measured as accuracy and AUC) and size, it is clear that Chip1 is over-training, producing very detailed models at the expense of generalization ability. When the task is prediction, this is of course not a desirable property, but if the aim, as in Experiment 2 and 3, is to describe a relatively small set of production instances, it might actually be beneficial.

Turning to Experiment 2, where the transparent models were built using only oracle data, Table VII shows the aggregated accuracy results.

TABLE VII
RESULTS EXPERIMENT 2 - X - ACCURACY

	JRip	Chip1		Chip2	
		A	F5	A	F5
2.5D	.724	.741	.741	.738	.762
AZ Desc.	.734	.750	.750	.751	.767
Mean Acc Chem-Phys	**.729**	**.746**	**.746**	**.745**	**.764**
Mean Rank Chem-Phys	**3.81**	**2.69**	**2.81**	**3.06**	**1.50**
Frags.	.671	.736	.736	.736	.729
sign12	.691	.719	.719	.734	.754
ecfi1024	.690	.753	.753	.728	.750
Mean Acc Fingerpr.	**.684**	**.736**	**.736**	**.733**	**.744**
Mean Rank Fingerpr.	**4.58**	**2.00**	**1.96**	**2.96**	**2.04**
Total Mean	.702	.740	.740	.738	.752
Total Mean Rank	4.28	2.28	2.30	3.00	1.83

Here, the differences between the techniques are much more marked than in Experiment 1, showing that Chipper managed to use the oracle data in a more effective way than JRip. This is especially true for the fingerprint data sets, where JRip actually obtains lower average accuracy when using oracle data compared to using training data only. It is also notable that Chip1 obtains the same averages for all attribute sets, regardless of rule selection criterion and representation, suggesting that the same rules are often chosen. Closer inspection of the results, on data set level, shows that while results are very similar, they are not identical for most data sets. For the chemical-physical data sets, a Friedman test followed by a Nemenyi post-hoc test, with the critical distance for $\alpha = 0.05$ being 1.52 for 16 data sets and five classifiers, shows that the only significant differences are that Chip2-F5 performed significantly better than JRip and Chip2-A. For the fingerprint data sets, where the critical distance is 1.25, all Chipper techniques performed significantly better than JRip, and all these differences hold for the total ranks. In addition, Chip2-F5 performed significantly better than Chip2-A overall.

In Table VIII, the aggregated AUC results for Experiment 2 are shown.

TABLE VIII
RESULTS EXPERIMENT 2 - X - AUC

	JRip	Chip1		Chip2	
		A	F5	A	F5
2.5D	.725	.741	.743	.746	.771
AZ Desc.	.738	.751	.751	.756	.776
Mean AUC Chem-Phys	**.731**	**.746**	**.747**	**.751**	**.774**
Mean Rank Chem-Phys	**3.25**	**2.63**	**2.56**	**2.94**	**1.38**
Frags.	.675	.738	.738	.738	.748
sign12	.693	.720	.720	.746	.766
ecfi1024	.694	.754	.754	.738	.764
Mean AUC Fingerpr.	**.687**	**.737**	**.737**	**.740**	**.759**
Mean Rank Fingerpr.	**4.58**	**2.29**	**2.29**	**2.38**	**1.21**
Total Mean	.705	.741	.741	.745	.765
Total Mean Rank	4.05	2.43	2.40	2.60	1.28

For the chemical-physical data sets, the significant differences are the same as for accuracy, but for the fingerprint data sets, Chip2-F5 obtained significantly better AUC than all other techniques. It also still holds that all Chipper techniques are significantly better than JRip on the fingerprint data sets. Over all data sets, Chip2-F5 performed significantly better than all other techniques and JRip was significantly worse than all Chippers.

In Table IX, the size results from Experiment 2 are given.

TABLE IX
RESULTS EXPERIMENT 2 - X - SIZE

	JRip	Chip1		Chip2	
		A	F5	A	F5
2.5D	6.1	5.1	5.1	3.4	5.2
AZ Desc.	5.9	4.6	4.6	3.1	4.6
Mean Size Chem-Phys	**6.0**	**4.9**	**4.8**	**3.3**	**4.9**
Mean Rank Chem-Phys	**4.38**	**3.00**	**2.88**	**1.56**	**3.06**
Frags.	6.0	10.8	10.6	6.2	7.8
sign12	5.9	11.3	10.9	5.7	7.9
ecfi1024	6.4	7.3	7.1	4.8	6.4
Mean Size Fingerpr.	**6.1**	**9.8**	**9.5**	**5.6**	**7.4**
Mean Rank Fingerpr.	**2.21**	**4.58**	**3.54**	**1.96**	**2.46**
Total Mean	6.1	7.8	7.6	4.7	6.4
Total Mean Rank	3.08	3.95	3.28	1.80	2.70

The overall impression is that all techniques produce very small rule sets, which is to be expected since the task is now to build a model using oracle data only, which for some data sets contains as few as 17 instances. There are some significant differences in ranks, of course, such as Chip2-A being significantly better than JRip on the chemical-physical data sets. Interestingly, JRip performs very well regarding size on the fingerprint data sets, being significantly better (together with Chip2-A) than both Chip1 variants. This, in conjunction with the relatively poor results on accuracy and AUC, suggests too much emphasis on generalization in this situation, where the pruning and rule optimization procedures in the JRip algorithm end up pruning away useful conjuncts.

To summarize Experiment 2, then, Chip2-F5 once again obtained the highest accuracy and AUC, and also produced reasonably small models. A comparison of the results between Chip1 and Chip2 shows that using parameter settings favoring very detailed rule sets is not the best option even for the task of describing only a small set of production instances; rather the normal Chipper prediction settings should be used. It is, however, beneficial for accuracy and AUC to use the FOIL rule selection and allow conjunctions in rules. Regarding JRip, it is quite clear that this task does not suit the technique at all. Even though it produces rule sets of competitive size, performance on accuracy and AUC are significantly worse than most Chipper variants.

Turning to Experiment 3, where the transparent models are built using both training and oracle data, Table X shows the aggregated accuracy results.

TABLE X
RESULTS EXPERIMENT 3 - IX - ACCURACY

	JRip	Chip1		Chip2	
		A	F5	A	F5
2.5D	.737	.745	.742	.735	.755
AZ Desc.	.762	.762	.748	.760	.763
Mean Acc Chem-Phys	**.749**	**.753**	**.745**	**.747**	**.759**
Mean Rank Chem-Phys	**3.13**	**2.81**	**3.31**	**3.00**	**2.31**
Frags.	.719	.735	.734	.734	.742
sign12	.735	.726	.721	.732	.754
ecfi1024	.735	.750	.707	.739	.759
Mean Acc Fingerpr.	**.730**	**.737**	**.721**	**.735**	**.752**
Mean Rank Fingerpr.	**3.38**	**2.63**	**3.50**	**3.13**	**1.88**
Total Mean	.738	.744	.730	.740	.755
Total Mean Rank	3.28	2.70	3.43	3.08	2.05

Here, the picture is again back to that of Experiment 1, i.e. quite small differences between the techniques, especially for the chemical-physical data sets. For the fingerprint data sets, Chip2-F5 is again the clear winner and is significantly better than all other techniques except Chip2-A. The same significant differences also hold when the comparison is made over all data sets.

In Table XI below, the corresponding AUC results are shown.

TABLE XI
RESULTS EXPERIMENT 3 - IX - AUC

	JRip	Chip1		Chip2	
		A	F5	A	F5
2.5D	.745	.735	.741	.741	.770
AZ Desc.	.766	.741	.750	.774	.775
Mean AUC Chem-Phys	**.756**	**.738**	**.746**	**.758**	**.773**
Mean Rank Chem-Phys	**2.69**	**3.75**	**3.00**	**2.38**	**1.56**
Frags.	.726	.736	.741	.765	.758
sign12	.741	.720	.721	.759	.771
ecfi1024	.743	.745	.705	.751	.764
Mean AUC Fingerpr.	**.737**	**.734**	**.723**	**.758**	**.764**
Mean Rank Fingerpr.	**3.29**	**3.29**	**3.54**	**1.75**	**1.75**
Total Mean	.744	.736	.732	.758	.768
Total Mean Rank	3.05	3.48	3.33	2.00	1.68

Again, Chip2-F5 clearly performed best, and the interesting significant differences are that both Chip2 variants were significantly better than all other techniques.

Table XII shows the size results from Experiment 3.

TABLE XII
RESULTS EXPERIMENT 3 - IX - SIZE

	JRip	Chip1		Chip2	
		A	F5	A	F5
2.5D	9.7	16.1	20.6	5.8	15.6
AZ Desc.	8.8	14.3	18.0	5.1	13.4
Mean Size Chem-Phys	**9.2**	**15.2**	**19.3**	**5.4**	**14.5**
Mean Rank Chem-Phys	**2.00**	**3.75**	**4.88**	**1.00**	**3.81**
Frags.	8.9	37.3	41.6	11.1	21.3
sign12	10.6	37.5	43.4	9.7	21.5
ecfi1024	10.6	23.3	30.5	8.0	18.3
Mean Size Fingerpr.	**10.1**	**32.7**	**38.5**	**9.6**	**20.3**
Mean Rank Fingerpr.	**1.58**	**4.33**	**4.67**	**1.42**	**3.00**
Total Mean	9.7	25.7	30.8	7.9	18.0
Total Mean Rank	1.75	4.10	4.75	1.25	3.10

The picture is again similar to Experiment 1, with JRip and Chip2-A consistently producing small rule sets; indeed these techniques are significantly better than the three other techniques both on the fingerprint data sets and over all data sets. The differences in rule set size are also about the same as in Experiment 1.

To summarize Experiment 3, where both training and oracle data were used, Chip2-F5 obtained the best predictive performance, measured both as accuracy and AUC, and JRip and Chip2-A produced the smallest models. Again, there are clear benefits from using Chipper with its normal prediction settings, and there is the possibility to gain some predictive power, at the expense of comprehensibility, by using rule selection based on FOIL and allowing conjunctive rules. JRip again produced small models, but had worse predictive performance than Chip2.

The results section will conclude with some comparisons across the three experiments, with the aim of illustrating the effect of using oracle data and to compare the different techniques.

To get a clearer picture of the effect of using oracle data, Table XIII below shows significant gains in accuracy for the X and IX setups. A + indicates a significant improvement, using a sign test, compared to normal rule induction; i.e. the detailed accuracy results from Experiment 1. For the 16 chemical-physical data sets, 13 wins are needed for a statistically significant difference (for $\alpha = 0.05$) and for the 24 fingerprints data sets, 18 wins are necessary to establish a significant difference.

TABLE XIII
EFFECTS OF ORACLE DATA - ACCURACY

		Jrip	Chip1		Chip2	
			A	F5	A	F5
X	Chem-Phys.	+	+	+		+
	Fingerprint		+	+	+	+
IX	Chem-Phys.	+	+	+	+	+
	Fingerprint	+	+		+	+

The overall impression from this is that all Chipper techniques utilized oracle data very well. The increases in accuracy were generally bigger when using oracle data to supplement training data, confirming IX to be the best setup. JRip performed rather badly on oracle data only and was furthermore not able do exploit the additional oracle data in setup IX to increase its predictive performance as much as the two best Chippers.

Looking at the size results across the three experiments, there is a clear ordering with the smallest models being obtained when using setup X, i.e., oracle data only, followed by setup I, i.e., normal rule induction, and the largest models when using setup IX, i.e., both training and oracle data. This ordering corresponds to the number of instances used for model building in the different setups. The result is to be expected since decision list algorithms operate by sequential covering and more instances to cover means that more rules need to be formulated before the default rule. That the differences in size are much more accentuated for Chip1 is also a consequence of how the algorithm works and the parameter settings requiring very accurate rules and few instances covered by the default rule.

For a comparison of the techniques, Table XIV below shows the averaged accuracy results and mean ranks over all three experiments.

TABLE XIV
COMPARISON OF TECHNIQUES - ACCURACY

		Jrip	Chip1		Chip2	
			A	F5	A	F5
I	Mean Acc.	.695	.703	.702	.709	.713
	Mean Rank	3.43	2.83	2.93	3.03	2.63
X	Mean Acc.	.702	.740	.740	.738	.752
	Mean Rank	4.28	2.28	2.30	3.00	1.83
IX	Mean Acc.	.738	.744	.730	.740	.755
	Mean Rank	3.28	2.70	3.43	3.08	2.05
Total	Mean Acc.	.712	.729	.724	.729	.740
	Mean Rank	3.66	2.60	2.88	3.03	2.17

This confirms the picture from the experiments that Chip2-F5 performed best overall on accuracy and that the other Chipper techniques also outperformed JRip. Regarding AUC, the overall result was that Chipper performed relatively better on AUC than JRip, and that this held regardless of model size. As noted in the summaries of each experiment above, Chip2 quite often obtained significantly better AUC than all other techniques.

Finally, looking at model size for the different techniques, JRip and Chip2-A consistently obtained the smallest models, often significantly better than the other techniques. Of the other techniques, Chip2-F5 was almost always better than Chip1.

V. CONCLUSION

In this paper, different ways of utilizing oracle data for a concept description task in the drug discovery domain have been evaluated using the standard decision list technique JRip and the Chipper technique, specifically aimed at concept description, with four different parameter settings. The evaluation was carried out using both predictive performance, measured

as accuracy and AUC, and comprehensibility, measured as rule set size.

When the task was normal rule induction, all techniques performed similarly on accuracy, but with a small advantage for Chipper over JRip, with the best Chipper variant being normal prediction settings augmented with rule selection based on optimizing FOIL's information gain and with conjunctive rules allowed. When measuring AUC, both this Chipper and the normal Chipper performed significantly better than all other techniques.

The use of oracle data only for building models resulted in clear gains in accuracy and AUC, coupled with smaller models for all Chipper variants, showing that Chipper is very well suited to producing concept descriptions for a set of production instances. JRip had some problems with this task, with putting too much emphasis on generalization being the most probable explanation.

The best way of using oracle data was together with normal training data; indeed all techniques consistently performed much better on both accuracy and AUC with this setup, compared to normal rule induction. There were, however, some increases in average model size when more data was used to build the models.

When looking at how the different techniques performed, the clear winner regarding predictive performance was normal Chipper with rule selection based on FOIL's information gain and conjunctions. This technique, however, produced somewhat larger models than JRip and normal Chipper, so if really compact concept descriptions are needed, normal Chipper is a good alternative. Indeed, the Chipper algorithm performed remarkably well on all tasks in this study, managing to produce accurate and compact concept descriptions both with and without oracle data.

The overall conclusion is that oracle coaching works very well for concept description. Further, it was seen that augmenting normal training data with oracle data will lead to better predictive performance for all techniques evaluated.

ACKNOWLEDGMENT

This work was supported by the INFUSIS project (www.his.se/infusis) at the University of Skövde, Sweden, in partnership with the Swedish Knowledge Foundation under grant 2008/0502.

REFERENCES

[1] H. van de Waterbeemd and E. Gifford, "Admet in silico modelling: towards prediction paradise?" *Nat Rev Drug Discov*, vol. 2, no. 3, pp. 192–204, March 2003.

[2] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0 step-by-step data mining guide," The CRISP-DM consortium, Tech. Rep., 2000.

[3] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.

[4] T. Joachims, "Transductive inference for text classification using support vector machines." Morgan Kaufmann, 1999, pp. 200–209.

[5] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowl.-Based Syst.*, vol. 8, no. 6, pp. 373–389, 1995.

[6] U. Johansson and L. Niklasson, "Evolving decision trees using oracle guides," in *CIDM*. IEEE, 2009, pp. 238–244.

[7] U. Johansson, C. Sönströd, and T. Löfström, "Oracle coached decision trees and lists," in *IDA '10: Proceedings of the 9th International Symposium on Intelligent Data Analysis*, 2010.

[8] U. Johansson, C. Sönströd, T. Löfström, and H. Boström, "Chipper – a novel algorithm for concept description," in *Proceeding of the 2008 conference on Tenth Scandinavian Conference on Artificial Intelligence*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2008, pp. 133–140.

[9] C. L. Bruce, J. L. Melville, S. D. Pickett, and J. D. Hirst, "Contemporary qsar classifiers compared," *J. Chem. Inf. Model.*, vol. 47, no. 1, pp. 219–227, January 2007.

[10] J. J. Sutherland, L. A. O'Brien, and D. F. Weaver, "A comparison of methods for modeling quantitative structure-activity relationships," *J. Med. Chem.*, vol. 47, no. 22, pp. 5541–5554, October 2004.

[11] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, June 2005.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, October 2001.

[13] W. W. Cohen, "Fast effective rule induction," in *In Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995, pp. 115–123.

[14] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[15] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of American Statistical Association*, vol. 32, pp. 675–701, 1937.

[16] P. B. Nemenyi, *Distribution-free multiple comparisons. PhD-thesis*. Princeton University, 1963.