Generating Comprehensible QSAR Models

Cecilia Sönströd[1], Ulf Johansson[1] and Ulf Norinder[2]

Abstract—**This paper presents work in progress from the INFUSIS project and contains initial experimentation, using publicly available medicinal chemistry datasets, on obtaining comprehensible QSAR models. Three techniques are evaluated on both predictive performance, measured as accuracy, and comprehensibility, measured as model size. The chosen techniques are J48 decision trees and JRip and Chipper decision lists. The results show that J48 obtains superior accuracy and that Chipper performs best of the two decision list algorithms on accuracy. Furthermore, it is seen that, regarding accuracy, all techniques benefit from feature reduction, which almost always results in increased accuracy. Regarding comprehensibility, JRip obtains the smallest models, followed by Chipper, with J48 producing the largest models. For model size, feature reduction is not seen to be universally beneficial; only J48 produces smaller models for the reduced datasets, while both decision list algorithms actually produce larger models on average. The overall conclusion is that, for these datasets, there exists a definite tradeoff between accuracy and comprehensibility that needs to be investigated further.**

Index Terms—**Classification, Comprehensibility, Concept Description, Data Mining, QSAR.**

I. Introduction

In this paper, several techniques producing transparent models are evaluated with respect to their ability of producing comprehensible Quantitative Structure-Activity Relationship (QSAR) models. 8 publicly available datasets and 6 different attribute sets are used. The overall purpose is to evaluate the different techniques regarding both predictive performance and comprehensibility.

II. Background

When performing classification, it is sometimes desirable to obtain models that enable human comprehension, both to increase confidence in the model and to provide understanding and explanation from the model. Indeed, the CRISP-DM [1] data mining framework specifically identifies a data mining problem type called *concept description*, where the aim is to provide insights into interesting relationships in the underlying domain. The obvious way to achieve comprehensibility is to use transparent models, i.e. models that enable tracing of predictions, typically by following the conditions that a classified instance has to fulfill in order to obtain its class label.

However, transparency alone does not guarantee comprehensible models, since some transparent models can be very complex and contain literally hundreds of conditions. Comprehensibility is of course not an entirely objective property, but depends, among other things, on the domain and the person interpreting the model. However, some properties of a transparent model that should be deemed desirable for comprehensibility can be identified. Among the most natural are:

- ability to survey the whole model
- clear representation of the most important relationships
- that relationships shown are general

Most often, model size is used to evaluate comprehensibility and models are required to have high accuracy on unseen data to guarantee that relationships are general.

It is known that techniques producing transparent models in general have lower predictive performance than techniques that produce opaque models, such as neural networks or support vector machines. To obtain comprehensibility, accuracy is often sacrificed by the use of simpler models; a tradeoff termed the *accuracy vs. comprehensibility* tradeoff.

A. The Chipper Algorithm

Within the field of machine learning, there are many techniques producing transparent models, most notably using the decision tree and ordered rule set (decision list) representations. However, very few techniques are specifically aimed at comprehensible models and even fewer contain explicit means for controlling the accuracy vs. comprehensibility tradeoff. In [2], we introduced the decision list algorithm Chipper, specifically aimed at 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

[1] CSL@BS Research Group, School of Business and Informatics, University of Borås, Sweden. Email: {cecilia.sonstrod, ulf.johansson}@hb.se

[2] AstraZeneca R&D Södertälje, Sweden. Email: ulf.norinder@astrazeneca.com

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.

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. 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. Chipper has been evaluated in a number of studies both on benchmark and medicinal chemistry datasets, see e.g. [3] and [4], and has been shown to perform well regarding comprehensibility, but sometimes at the expense of low accuracy.

The main purpose of the present study is to use a large number of medicinal chemistry datasets, with different characteristics, to evaluate the predictive performance and comprehensibility of Chipper models. This evaluation and comparisons against models produced by other techniques will provide further insights into the strengths and weaknesses of the Chipper algorithm.

III. Method

The following techniques were evaluated in this study:

- RIPPER [5], implemented as JRip in Weka [6]
- C4.5 [7] trees, implemented as J48 in Weka
- Chipper, also implemented in Weka

The motivations for choosing RIPPER and C4.5 is that they represent the state-of-the-art for decision lists and decision trees, respectively, and that these representations are the two most natural for transparent classification models. For all three techniques, the default parameter settings were used; meaning among other things that both J48 and Jrip use pruning, and that Chipper used an *ignore* value of 5% and a *stop* value of 95%.

A. Datasets

In this study, eight different medicinal chemistry datasets from the study of Bruce et al. are used [8]. These datasets were originally used by Sutherland et al. [9]. The datasets are related to 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 (THER), 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.

The continuous numerical values for activity (pIC50 for the first five datasets 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). Since each dataset showed a uniform distribution of activity values, the transformation used the median activity value as a threshold between the two classes to create a 50/50 split of active/inactive observations, creating two-class datasets with a balanced class distribution.

It should be noted that Bruce et al. used two separate feature sets, *2.5D descriptors* (2.5D) and *linear fragment descriptors* (Frags.) to characterize the chemical structures in the data sets. Both these attribute sets are used in this study, together with a further four attribute sets obtained from AstraZeneca, described below.

oeSelma, the AstraZeneca in-house oeSelma program generates 2D descriptors related to size, ring structure, flexibility, hydrogen bonds, polarity, electronic environment and lipophilicity. The attribute sets used in this study contains 93 such descriptors.

AstraZeneca Descriptor Set (AZ Desc), an in-house modification of the oeSelma descriptors, consisting of 196 physical-chemical descriptors.

sign12, signature fingerprints that characterize a molecule by a set of canonical subgraphs where the vertices denote the atoms in the molecule and the edges correspond to the bonds between those atoms. Each subgraph is rooted on a different vertex (atom) with a predefined level of branching referred to as the height of the fingerprint (heights 1 and 2 used here) [10].

ecfi descriptors are the Scitegic Extended Connectivity Fingerprints. The fingerprint assigned to an atom is based on the number of connections, the element type, the charge and the mass of the atom in question [11]. The ecfi attribute set in this study contains 1024 molecular fingerprints.

In all six feature sets, initial investigation of the datasets showed strong correlation between several input attributes, indicating that feature reduction could be beneficial for both predictive performance and comprehensibility. The standard procedure for feature selection in Weka (called CfsSubSetEval) was used to perform the attribute reduction. Simply put, CfsSubSetEval, when using standard settings, performs a best first search, keeping only attributes relatively strongly correlated with the target variable. In addition, attributes strongly correlated with other attributes, already in the reduced feature set, are discarded. The feature selection resulted in a significant reduction of the number of features for all datasets. Tables I and II below summarize the characteristics of the resulting datasets.

TABLE I
CHARACTERISTICS OF DATA SETS USED – PHYSICAL-CHEMICAL DESC.

Dataset	**Inst.**	**Attribute set**					
		2.5D	2.5D Red	oeSelma	oeSelma Red	AZ Desc.	AZ Desc. Red
ACE	114	56	12	93	14	196	17
AchE	111	63	9	93	11	196	16
BZR	163	75	16	93	15	196	23
COX2	322	74	12	93	8	196	14
DHFR	397	70	10	93	16	196	17
GPB	66	70	2	93	8	196	6
THER	76	64	10	93	11	196	17
THR	88	66	7	93	7	196	11

TABLE II
CHARACTERISTICS OF DATA SETS USED – FINGERPRINT DESC.

Dataset	**Instances**	**Attribute set**					
		Frags	Frags Red	ecfi	ecfi Red	sign12	sign12 Red
ACE	114	1024	11	1024	29	332	13
AchE	111	774	16	1024	32	211	14
BZR	163	832	20	1024	45	450	26
COX2	322	660	24	1024	45	573	30
DHFR	397	951	14	1024	33	487	17
GPB	66	692	14	1024	32	239	21
THER	76	575	21	1024	23	251	15
THR	88	527	7	1024	22	220	8

A. *Experiments*

All experiments were performed in Weka, using all datasets obtained by using 6 different attribute sets, with and without reduction, for the 8 datasets available, i.e., 6*2*8=96 datasets in total. The use of both unreduced and reduced data was motivated by the need to investigate the effects of feature reduction on both accuracy and comprehensibility.

Two main experiments were conducted. Experiment 1 concerned the predictive performance, measured as accuracy, of each technique. Since all datasets contain relatively few instances, this experiment employed 10x10-fold cross-validation (10x10CV), with identical folds for all techniques, to ensure reliable results. In Experiment 2, comprehensibility was evaluated using the simple criterion of model size, measured as number of atomic conditions in the model. This measure was calculated on the model output by Weka, which is built using the whole dataset.

IV. RESULTS

In Table III below, the summarized results for Experiment 1 are presented. The *Avg. Acc.* columns present the average accuracy for each technique over all 8 datasets for a certain attribute set.

TABLE III
ACCURACY RESULTS AND RANKS – EXPERIMENT 1 – 10X10CV

Attribute set	**Technique**					
	JRip		**J48**		**Chipper**	
	Avg. Acc.	Avg. Rank	Avg. Acc.	Avg. Rank	Avg. Acc.	Avg. Rank
2.5D	0.686	2.63	0.720	1.25	0.693	2.13
2.5D Red	0.722	1.63	0.732	1.63	0.683	2.63
oeSelma	0.747	2.00	0.749	2.25	0.753	1.75
oeSelma Red	0.751	2.25	0.758	1.75	0.741	2.00
AZ Desc.	0.748	2.00	0.748	1.88	0.747	2.13
AZ Desc. Red	0.765	1.88	0.771	2.00	0.754	2.13
Frags	0.689	2.50	0.739	1.13	0.700	2.38
Frags Red	0.742	2.50	0.755	1.50	0.761	2.00
ecfi	0.700	2.25	0.716	1.63	0.722	2.00
ecfi Red	0.743	2.63	0.768	1.75	0.795	1.63
sign12	0.701	2.38	0.742	1.38	0.708	2.25
sign12 Red	0.729	2.63	0.754	1.63	0.760	1.75
Average Unred.	0.712	2.29	0.736	1.58	0.721	2.10
Average Red.	0.742	2.25	0.756	1.71	0.749	2.02
Average All	0.727	2.27	0.746	1.65	0.735	2.06

Several quite clear results emerge from this experiment. First, there is a clear ordering of the three techniques, with J48 consistently achieving best, with Chipper in second place and JRip being the worst overall. This is seen both in average accuracies and in ranks. This result is a bit surprising, since earlier studies have had JRip outperform Chipper on accuracy.

The picture becomes even clearer when considering the number of times each technique obtains the best result, see Table IV below. There are a number of occasions where two techniques share the best result, and hence the total number of wins for an attribute set is sometimes 9, even though there are only 8 datasets for each attribute set.

TABLE IV
WINNING TECHNIQUES – EXPERIMENT 1 – 10X10CV

Attribute set	**Technique**		
	JRip Wins	**J48** Wins	**Chipper** Wins
2.5D	0	6	2
2.5D Red	5	4	0
oeSelma	1	3	4
oeSelma Red	0	5	3
AZ Desc.	2	3	3
AZ Desc. Red	2	3	3
Frags	0	7	1
Frags Red	1	4	3
ecfi	0	5	3
ecfi Red	1	3	4
sign12	0	6	2
sign12 Red	0	4	4
Total wins Unred.	3	30	15
Total wins Red.	9	23	17
Total wins All	12	53	32

A rather typical result set, using the unreduced 2.5D attribute set, is shown in Table V below. Typically, there are quite small differences in accuracy. However, when looking at all available results, it is seen that each of techniques have a few really bad results, and also a few outstanding ones. In the result set below, e.g., Chipper obtains a really high accuracy on the BZR dataset, whilst performing quite badly on the DHFR dataset. J48 is, however, the most robust of the techniques, obtaining the worst accuracy only on 19 (out of a possible 96) datasets.

TABLE V

DETAILED ACCURACY FOR 2.5D DATASETS – EXPERIMENT 1 – 10X10CV

Dataset	Technique					
	JRip		J48		Chipper	
	Avg. Acc.	Rank	Avg. Acc.	Rank	Avg. Acc.	Rank
ACE	0.839	3	0.856	1	0.841	2
AchE	0.633	3	0.651	2	0.660	1
BZR	0.696	3	0.710	2	0.759	1
COX2	0.700	2	0.744	1	0.665	3
DHFR	0.773	2	0.778	1	0.744	3
GPB	0.604	3	0.703	1	0.655	2
THER	0.639	2	0.675	1	0.618	3
THR	0.604	3	0.641	1	0.605	2
Average acc.	0.686		0.720		0.693	
Average rank		2.63		1.25		2.13
Total wins		0		6		2

The second main result from Experiment 1 is that, on these datasets, attribute reduction is almost always beneficial for predictive performance measured as accuracy. This is in accordance with statements from domain experts that these datasets are quite easy and also contain highly correlated attributes. Again, this is seen clearly when tabulating how many times out of the possible 8 that accuracy is increased for the reduced datasets, see Table VI below.

TABLE VI

FEATURE REDUCTION IMPROVEMENTS – EXPERIMENT 1 – 10X10CV

Attribute set	Technique		
	JRip #improve	J48 #improve	Chipper #improve
2.5D	5	6	6
oeSelma	4	6	2
AZ Desc.	6	6	6
Frags	7	6	8
ecfi	8	8	7
sign12	7	5	8
Total	37	37	37

Concerning Experiment 2, the main results regarding model size are presented in Table VII below. The column *Avg. Size* contains the average model size, measured as the total number of conditions in the model, over the 8 datasets for each attribute set, and the column *Avg. Rank* contains ranks averaged in the same way.

TABLE VII

MODEL SIZE AND RANKS – EXPERIMENT 2

Attribute set	Technique					
	JRip		J48		Chipper	
	Avg. Size	Avg. Rank	Avg. Size	Avg. Rank	Avg. Size	Avg. Rank
2.5D	5.13	1.38	12.63	2.75	5.00	1.38
2.5D Red	4.75	1.38	11.13	2.25	6.13	1.88
oeSelma	3.38	1.13	15.88	2.88	5.38	1.88
oeSelma Red	4.75	1.25	10.75	2.50	6.63	2.13
AZ Desc.	5.25	1.25	12.75	2.75	4.75	1.88
AZ Desc. Red	4.38	1.25	10.00	2.75	5.75	1.88
Frags	4.25	1.13	9.25	2.13	14.13	2.63
Frags Red	5.13	1.63	4.50	1.38	10.25	2.75
Ecfi	4.50	1.13	15.13	2.50	7.50	2.13
ecfi Red	6.88	1.25	8.88	1.63	15.63	2.75
sign12	4.00	1.00	12.88	2.50	8.88	2.25
sign12 Red	6.00	1.50	7.13	1.50	11.63	2.75
Average Unred.	4.42	1.17	13.08	2.58	7.60	2.02
Average Red.	5.35	1.36	8.82	2.01	9.47	2.35
Average All	4.89	1.27	10.95	2.30	8.54	2.18

As can be seen from the table, results regarding comprehensibility (size) are very clear regarding how the different techniques perform. This time, however, the ordering from Experiment 1 is reversed, with JRip excelling in producing small models, and J48 producing quite large models.

Another striking result from Experiment 2 is the effect of feature reduction, which is largely detrimental on model size for both decision list techniques, but beneficial for J48. This effect is most pronounced for the fingerprint attribute sets; indeed JRip produces larger models for all reduced fingerprint datasets, while the decrease in size for J48 is quite dramatic with more or less half the average model size for these datasets.

Rank averages indicate that JRip wins a majority of the datasets, and this is confirmed by looking at the number of wins for each technique over all datasets, summarized in Table VIII below.

TABLE VIII

WINNING TECHNIQUES – MODEL SIZE - EXPERIMENT 2

Attribute set	Technique		
	JRip Wins	J48 Wins	Chipper Wins
2.5D	5	1	5
2.5D Red	6	2	2
oeSelma	7	0	2
oeSelma Red	6	1	1
AZ Desc.	6	1	2
AZ Desc. Red	6	0	2
Frags	7	1	1
Frags Red	4	5	0
ecfi	7	1	1
ecfi Red	7	3	1
sign12	8	0	2
sign12 Red	4	5	0
Total wins Unred.	40	4	13
Total wins Red.	29	11	6
Total wins All	69	15	19

This table reinforces the picture that J48 is the technique that benefits from feature reduction, obtaining most of its winning results on the reduced datasets, at the expense of both decision list algorithms.

A typical set of results regarding comprehensibility is the unreduced oeSelma attribute set, shown in Table IX below.

TABLE IX
DETAILED MODEL SIZE FOR OESELMA DATASET – EXPERIMENT 2

Dataset	Technique					
	JRip		J48		Chipper	
	Size	Rank	Size	Rank	Size	Rank
ACE	1	1	5	3	2	2
AchE	1	1	5	2	6	3
BZR	3	1	22	3	6	2
COX2	10	2	35	3	8	1
DHFR	6	1	35	3	6	1
GPB	1	1	6	3	5	2
THER	1	1	6	3	5	2
THR	4	1	13	3	5	2
Average size	3.38		15.88		5.38	
Average rank		1.13		2.88		1.88
Total wins		7		0		2

Here, JRip produces several models containing just a single test, while Chipper has a relatively stable model size and J48 sometimes produces very large models that are quite incomprehensible. Viewing these comprehensibility results in relation to accuracy, it is noteworthy that for this attribute set, JRip obtains the best accuracy on only one dataset, with J48 winning four and Chipper the remaining three. This shows that the extremely small JRip models are, in general, unable to adequately capture the relationships in these datasets.

V. CONCLUSION

The results clearly indicate that on these medicinal chemistry datasets, there is a definite tradeoff between accuracy and comprehensibility. J48 is the technique that overall produces that most accurate models, but these models are also the largest. There are some instances when models produced by J48 are simply too large to be deemed comprehensible at all, even if individual predictions can, in principle, be traced. JRip obtains the smallest average model size, but also the worst average accuracy.

Chipper obtains a compromise between accuracy and comprehensibility. Regarding accuracy, Chipper clearly outperforms JRip, the other decision list algorithm, despite not using any pruning or optimization. Also, it is noteworthy that the two decision list algorithms produce smaller models, indicating that this representation is suitable for concept description. Finally, feature reduction is seen to have a beneficial effect on accuracy for all techniques. However, feature reduction is detrimental to model size for both decision list algorithms, but results in smaller models for J48 trees.

VI. DISCUSSION AND FUTURE WORK

In this study, all techniques were run with their default settings. Past studies, see e.g. [3], have indicated that Chipper obtains relatively stable results regarding accuracy over different parameter settings, so it is natural to evaluate how much Chipper model size can be reduced by using a higher *ignore* value (permitting slightly less accurate rules) and a lower *stop* value (permitting earlier formation of the default rule), without significant loss of accuracy. Initial experiments in this direction show quite promising results, with Chipper settings of *ignore* at 10% and *stop* at 80% seeming capable of bringing model size down significantly without loss of accuracy. Similarly, the possibility of increasing JRip predictive performance and decreasing J48 model size without loss of accuracy needs to be investigated.

Regarding comprehensibility, model size is a rather blunt measure, especially if the aim of concept description is taken into consideration. It is questionable whether a model containing a single test succeeds in bringing insights into the underlying domain. One can argue that a slightly larger model, from which several important relationships are discernable, would be more desirable. This is especially true if the model clearly shows which relationships are most important and/or general.

In [12], it was proposed that comprehensibility be broken down into the three aspects of brevity, relevance and interpretability. For brevity and relevance, numeric measures have been proposed, in [12] and [3] respectively, and it seems natural to evaluate the models from the present study on these measures.

REFERENCES

[1] The CRISP-DM Consortium, CRISP-DM 1.0, www.crisp-dm.org, 2000.

[2] U. Johansson, C. Sönströd, T. Löfström and H. Boström, "Chipper – A Novel Algorithm for Concept Description", *10th Scandinavian Conference on Artificial Intelligence*, IOS Press, pp. 133-140, Stockholm, Sweden, 2008.

[3] C. Sönströd, U. Johansson and T. Löfström, "Evaluating Algorithms for Concept Description", *The 2009 International Conference on Data Mining (DMIN09)*, Las Vegas, NV, 2009.

[4] C. Sönströd, U. Johansson, U. Norinder, and H. Boström, "Comprehensible Models for Predicting Molecular Interaction with Heart-Regulating Genes" , *7th International Conference on Machine Learning and Applications (ICMLA '08)*, Orlando, FL, IEEE press, pp. 559 – 564, 2008.

[5] W. Cohen, "Fast Effective Rule Induction", *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*, p. 115-123. Tahoe City, CA, 1995.

[6] I. H. Witten and E. Frank, *Data Mining – Machine Learning Tools and Techniques*, 2nd ed, Morgan Kaufman, 2005.

[7] J. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993.

[8] C. L. Bruce, J. L. Melville, S. D. Pickett, and J. D. Hirst, "Contemporary QSAR Classifiers Compared", *J. Chem. Inf. Model.*, 47:219-227, 2007.

[9] J. J. Sutherland, L.A. O'Brien, and D.F. Weaver, "A Comparison of Methods for Modeling Quantitative Structure–Activity Relationships", *J. Med. Chem.*, 47: 5541–5554, 2004.

[10] C. J. Churchwell, M. D. Rintoul, S. Martin, D. P. Visco Jr., A. Kotu, R-S. Larson, L. O. Sillerud, D. C. Brown and J.-L. Faulon, "The signature molecular descriptor: 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides", *J. Mol. Graphics Model.*, 22:4, p. 263-273

[11] Scitegic Pipeline Pilot, 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123-1365, U.S.A. Available from SciTegic Inc. at http://www.scitegic.com

[12] C. Sönströd, U. Johansson and R. König, "Towards a Unified View on Concept Description", *The 2007 International Conference on Data Mining (DMIN07)*, Las Vegas, NV, 2007.