Accuracy vs. Comprehensibility in Data Mining Models

Ulf Johansson
Department of
Business and Informatics
University of Borås
Sweden
ulf.johansson@hb.se

Lars Niklasson
Department of
Computer Science
University of Skövde
Sweden
lars.niklasson@ida.his.se

Rikard König
Department of
Business and Informatics
University of Borås
Sweden
rikard.konig@hb.se

Abstract - *This paper addresses the important issue of the tradeoff between accuracy and comprehensibility in data mining. The paper presents results which show that it is, to some extent, possible to bridge this gap. A method for rule extraction from opaque models (Genetic Rule EXtraction – G-REX) is used to show the effects on accuracy when forcing the creation of comprehensible representations. In addition the technique of combining different classifiers to an ensemble is demonstrated on some well-known data sets. The results show that ensembles generally have very high accuracy, thus making them a good first choice when performing predictive data mining.*

Keywords: data mining, ensembles, rule extraction.

1 Introduction

One application area of Information Fusion (hereafter IF) is decision support. Often information and data from different sources are needed to make informed decisions. The most challenging task is to fuse information from various external sources with internal sources. In this processes several problems must be handled, e.g. definition of suitable formats, validation of the quality of the sources, definition of source precedence, fusion models, etc. This paper deals with problems were only internal sources are fused, which means that the problem of IF is somewhat simpler. Nevertheless, the problem is far from trivial since the fusion model is not known. The task is to use stored information and generate a model that fits the data; a process often termed data mining.

Of key interest here is the tradeoff between the choice of a powerful model with high performance on novel data (i.e. a model with high accuracy) and a transparent model (i.e. a model with high comprehensibility). For some tasks high performance is sufficient but in some tasks a transparent model is necessary. The latter is often true in decision support tasks where reasons for the decision must be clearly identifiable.

The problem is not only to decide between accuracy and comprehensibility, it is also to decide on which data mining technique to use. There are a number of techniques available but there is no 'standard' technique. The selection of a data mining technique for a specific problem is therefore a daunting task. Not only is the choice extremely important for the overall performance but it must often also be made early in the process. In practice the choosing of technique is often a procedure of trial and error.

With this in mind most integrated all-purpose data mining software systems try to ease the choice of data mining technique by providing guides, wizards and even automated choices of techniques. However, an approach often utilized is to create many models, using different techniques, and apply this ensemble of models to create an overall model. The motivation for using ensembles in general is obvious; they are more robust, i.e. their applicability span over a larger set of problems.

At the same time an ensemble is by definition a set of models, making it very hard to express the relationships found in original variables. The option to use an ensemble thus is a choice of a black box model prioritizing accuracy. Consequently, if comprehensibility is needed, the data miner must find and use a simpler and less powerful model; typically a decision tree.

Experience from the field of Expert System has shown that an explanation capability is a vital function provided by symbolic AI systems. In particular the ability to generate even limited explanations is absolutely crucial for the user acceptance of such systems [1]. Since the purpose of most data mining systems is to support decision making the need for explanation facilities in these systems is apparent.

In [2], Andrews, Diederich and Tickle highlight the lack of an explanation facility when using opaque models (here neural networks) and argue for rule extraction; i.e. to create more transparent representation from trained neural networks:

> *It is becoming increasingly apparent that the absence of an explanation capability in ANN systems limits the realizations of the full potential of such systems, and it is this precise deficiency that the rule extraction process seeks to reduce.* (p. 374)

Ultimately there should be a standard technique, always producing both accurate and comprehensible models. A step in this direction would be a technique reducing the tradeoff between accuracy and comprehensibility by allowing the data miner to use powerful techniques like ensembles and neural networks while still providing an explanation facility. It is the intention of this paper to take

a step in this direction by exploring the effects on accuracy when forcing a data mining technique to generate small (and therefore more comprehensible) decision trees.

2 Background

Modern computer technology enables storing of huge amounts of data at moderate cost. While most data is not stored with predictive modeling or analysis in mind the collected data unquestionably represents potentially valuable information.

The activity to transform the collected data into actionable information is termed data mining, and is increasingly becoming recognized as an important activity within both the private and public sectors. Although there exists several definitions of data mining they are quite similar. In [3], the following definition is suggested:

> *Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules.* (p. 5)

Typically data mining either builds a predictive model from stored data and then uses this model on novel data or performs a descriptive partitioning of the data set.

Although data mining is used by many companies and there exists several integrated "of the shelf" data mining software tools there is no standard system or dominating method. As a matter of fact researchers (both from the academia and the business) constantly come up with improved or even innovative techniques.

At the same time there is a strong agreement among researchers and executives alike about the criteria that all data mining techniques must meet. Most importantly the techniques must have high performance. This criterion is, for predictive modeling, typically translated into that the technique should produce models that are likely to generalize well, thus showing good accuracy when applied to novel data. At the same time the comprehensibility of the model is very important since the results should ultimately be interpreted by a human. The need for comprehensibility is particularly important when someone needs to be accountable for the decision, e.g. in medical decisions.

The method CRISP-DM[1] points at the advantage of having "a verbal description of the generated model (e.g. via rules)", thus acknowledging the importance of comprehensibility. Only with this description is it possible to "assess the rules; are they logical, are they feasible, are there too many or too few, do they offend common sense?"

Traditionally most research papers focus on high performance, although the comprehensibility criterion is highly stressed by the business.

The comprehensibility issue is tightly connected to the choice of data mining technique. Some techniques like decision trees and linear regression are regarded as transparent[2]; i.e. allowing human inspection and understanding. Other techniques, most notable neural networks, are said to be opaque and must be used as black-boxes. The descriptions above are however too simple. The comprehensibility is at least, also dependent on the size of the model. As an example; it must be questionable whether an extremely bushy decision tree could be regarded as comprehensible.

The tradeoff between high performance techniques and techniques producing comprehensible models is very interesting. This becomes even more motivating since the techniques that arguably show the highest performance in general are neural networks and ensemble methods like boosted decision trees. Neither of these techniques produces comprehensible models in general. From this it seems inevitable that a choice has to be made between performance and comprehensibility. With this in mind several researchers have tried to bridge the gap by introducing techniques for converting opaque models to transparent models, without sacrificing accuracy. Most significant are the many attempts to extract rules from trained neural networks. Although the algorithms proposed often show good performance in the case studies reported, there is still no rule extraction method recognized as superior to all others. More specifically no existing method fulfils all the demands on a reliable rule extraction algorithm. Key demands (see e.g. [4]) are accuracy (the extracted model must perform almost as well as the original on unseen data), comprehensibility (the extracted model should be "easy" to interpret by a human) and fidelity (the extracted model should perform similar to the original model).

A clear indication of the status of the different rule extraction algorithms is the fact that none of the major data mining software tools includes a rule extraction facility.

The description above leads to the following two key observations:

- The task of transferring high-accuracy opaque models to transparent models while retaining the level of accuracy is important. A rule extraction method could, as part of the data mining process, reduce the need for a tradeoff between accuracy and comprehensibility.

[1] CRoss-Industry Standard Process for Data Mining was an ESPRIT project that started in the mid-1990's. The purpose of the project was to propose a non-proprietary industry standard process model for data mining. For details see www.crisp-dm.org.

[2] It should be noted that strictly speaking, the terms transparency and comprehensibility are not synonymous. A neural network could, for instance, be regarded as transparent; i.e. the topology, activation functions and the weight matrix can easily be turned into a functional description. However, following the standard terminology, the distinction is not vital in this paper either. The main point is that transparency without comprehensibility is of limited value.

- Although there are well-established demands to evaluate rule extraction algorithms against no specific rule extraction algorithm can be regarded as the "standard" method.

3 Previous work

We have previously [5] suggested a novel method for rule extraction from neural networks called G-REX. G-REX is a black-box method where the extraction strategy is based on genetic programming (GP). The overall purpose of a black-box method for rule extraction is to use the predictions of the neural net as the target variable. An extracted rule thus describes (in a more transparent representation) not the topology of the neural network but the relationship between the input variables and output variables found by the neural net. This raises the obvious question "why not apply GP to the data set directly?" However, as reported by Dorado et al. [6], and observed in previous applications of G-REX, the accuracy of the rule extraction is slightly higher than directly applied GP. An explanation could be that a neural net with high generalization capability in a sense is a better (more general) representation of the relation than the data set itself.

In G-REX GP is used for the search process. More specifically a pool of candidate rules (which could be Boolean rules, decision trees, m-of-n rules etc.) is continuously evaluated against an evaluation ('fitness') function. The best rules are kept and combined using genetic operators to raise the fitness over time. After many iterations (generations) the most fit program is chosen as the extracted rule.

G-REX uses the "ramped half and half" strategy [7] when creating the original population. Crossover and mutation are performed in a standard fashion. Rules are chosen for reproduction using roulette wheel selection; i.e. the probability for a rule being selected for reproduction is proportional to its fitness.

The exact parameters like crossover and mutation rates, number of individuals in the population and number of generations can easily be varied depending on the problem and should probably be found from initial experimentation. Normally a larger problem (more input variables and classes) would benefit mostly from larger populations.

When using G-REX on a specific problem fitness function, function set and terminal set must be chosen. The function and terminal sets determines the representation languages while the fitness function captures what should be optimized in the extracted representation. Normally the fitness function would include components measuring fidelity (i.e. a measure of how similar a particular set of rules is to the network) and comprehensibility with the possible addition of an accuracy part. The fidelity component could for instance increase the fitness value by one for each training sample classified in the same way as the neural net. The comprehensibility component is a penalty term; typically proportional to the length of the program. An accuracy component requires a validation set; i.e. a part of the data set not used for training of the neural net. If present this component would usually increase the fitness for each validation sample classified correctly.

In the original study [5] G-REX was demonstrated on some well-known classification problems. The problems were chosen to force G-REX to create different rule representations; i.e. Boolean rules and decision trees. The extracted models were compared with the standard tool See 5 [8] and shown to have, in general, higher accuracy on novel data.

G-REX has in later studies [9] been tested on a number of classification problems, but also been extended in several ways; see [10]. More specifically the extensions to the original G-REX are:

- G-REX can also be used on regression problems producing regression trees.
- G-REX is not limited to rule extraction from neural networks but has also been used to extract rules from boosted decision trees.
- G-REX has used several different representation languages, most recently *fuzzy rules*.

4 Method

The purpose of this study is to evaluate the performance of a general-purpose ensemble followed by G-REX rule extraction. More specifically an ensemble consisting of five data mining techniques; a C&R[1] tree, a boosted C&R tree, a CHAID tree [11–12], a Multi-Layer Perceptron neural network (MLP) [13] and a Radial Basis Function neural network (RBF) [14], is evaluated on five publicly available classification problems.

The data mining software used is "Statistica Data Miner"[2], so for the exact implementations of the techniques in the ensemble consult the program documentation. It should be noted that the technique termed C&R tree, according to the documentation, is a "comprehensive implementation of the methods described as CART©" [15].

4.1 Problems Used

Five different data sets, all publicly available from the UCI machine learning repository[3] were used. Below is a short description of each data set.

BUPA liver disorders (BLD)

This data set was donated by R. S. Forsyth. The problem is to predict whether or not a male patient has a liver disorder based on blood tests and alcohol consumption.

[1] Classification and Regression tree.
[2] www.statsoft.com
[3] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/

PIMA Indians Diabetes (PID)
The diagnostic, binary-valued variable investigated is whether or not the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

Image segmentation (SEG)
This data set was used in the StatLog project. The aim is to predict the central pixel given the multi-spectral values of the other pixels in a 3x3 neighborhood.

StatLog vehicle silhouette (VEH)
This data set originated from the Turing Institute, Glasgow, Scotland. The problem is to classify a given silhouette as one of four types of vehicle, using a set of features extracted from the silhouette. The four vehicles are double decker bus, Chevrolet van, Saab 9000 and Opel Manta 400.

Waveform (WAV)
This is an artificial three-class problem based on three waveforms. Each class consists of a random convex combination of two waveforms sampled at the integers with noise added. A description for generating the data is given in [15].

Table 1 shows the characteristics for these data sets.

Table 1: Characteristics of the data sets.

Data set	Classes	Attributes	Instances
BLD	2	6	345
PID	2	7	768
SEG	7	19	2310
VEH	4	18	846
WAV	3	21	3600

In [16], Lim, Loh and Shih, evaluate altogether 33 classification algorithms on 22 data sets, including the five data sets used here. Among other results the best and the worst accuracy on each data set were reported together with the accuracy of the naïve prediction; i.e. the most frequent class. For further comparison these values are given in Table 2.

Table 2: Results from the Lim, Loh and Shih study. Percent correct on test set

Data set	Best	Worst	Naive
BLD	72%	57%	56%
PID	78%	69%	67%
SEG	98%	48%	14%
VEH	85%	51%	26%
WAV	85%	52%	33%

4.2 Experiments
Each data set was randomly split in a training set and a test set. For every data set exactly 75% of the instances were used for training. This is slightly different from the study by Lim, Loh and Shih, who for most problems applied ten-fold cross-validation; i.e. using 90% of the data set for training in each phase.

No changes were made to the default settings for each individual technique of the ensemble. Obviously the results from the individual techniques could be aggregated in different ways. In this study bagging was used; i.e. each technique voted for a specific class and the ensemble reported the class with most votes. Two different approaches were tried; either all techniques were allowed to vote or just the three techniques with the highest accuracy on the training set.

Here G-REX is used on the results from the ensemble allowing three techniques to vote. G-REX was set up slightly differently depending on the number of classes. It evolved a Boolean rule for the binary problems and extracted decision trees for the others. The fitness function chosen was based on fidelity (i.e. each sample classified the same way as the ensemble increases fitness) and included a penalty term enforcing shorter programs.

Different settings were tried for G-REX. Especially the penalty term for longer rules was varied. Since the main purpose of G-REX is to enhance comprehensibility the aim was to find rather short rules with acceptable accuracy. In addition G-REX was also tested with less severe penalties for longer rules, obviously resulting in longer, but potentially more accurate, rules.

The two most interesting questions in the study are the accuracy of the ensemble and the relative performance of G-REX. Here the performance must recognize both accuracy and comprehensibility. With this in mind it is natural to compare G-REX, CHAID and C&RT; i.e. the techniques producing transparent models.

5 Results
Table 3 shows the results for the individual techniques (given as percent correct on test sets) from the experiments.

Table 3: Results for the techniques.

Data set	C&RT	C&RT Boost	CHAID	MLP	RBF
BLD	70	66	59	66	63
PID	72	77	76	77	77
SEG	95	90	60	91	90
VEH	75	73	48	80	72
WAV	74	73	61	85	85
MEAN	**77.2**	**75.8**	**60.8**	**79.8**	**77.4**

As can be seen in Table 3 the MLP has the highest accuracy. A classification and regression tree offers a potentially more comprehensible solution but entails a loss of accuracy, compared to the MLP. However neither of the techniques exhibits accuracy similar to the best results in the experiments of Lim, Loh and Shih.

Table 4 shows the results for the ensemble approaches, and G-REX applied to the best three ensembles. The results for G-REX are from models prioritizing

comprehensibility; i.e. using a penalty function enforcing shorter rules.

Table 4: Results for ensembles and G-REX.

Data set	Ensemble best three	Ensemble all	G-REX
BLD	70	70	70
PID	78	79	78
SEG	93	94	76
VEH	84	82	70
WAV	85	84	77
MEAN	**82.0**	**81.8**	**74.2**

As can be noted in Table 4, the accuracy for the ensemble is now similar to the best results of Lim Loh and Shih. The result for G-REX is clearly a loss in accuracy for the gain of a more comprehensible model (here compact decision trees).

Below is a short discussion about the results for each problem.

BUPA liver disorders
Most techniques produce similar results on this data set. The ensembles, C&RT and G-REX all have an accuracy of 70%. The Boolean rule extracted by G-REX is fairly small; see Fig. 1.

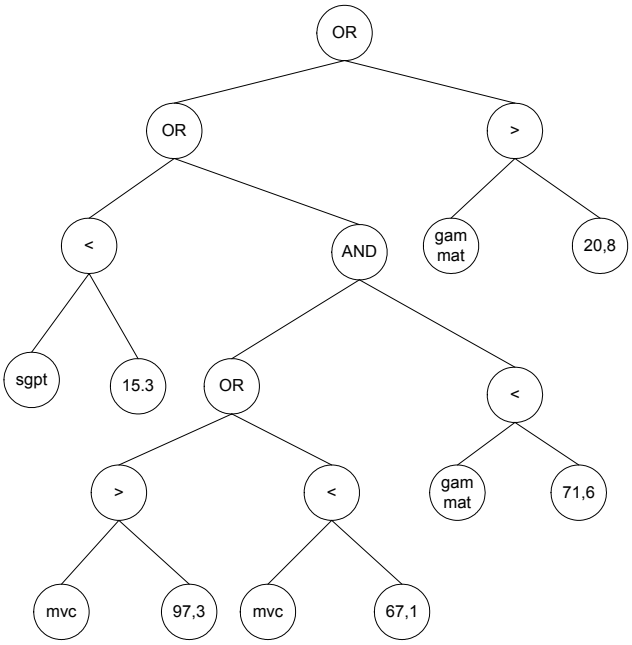

Fig. 1: Rule for BLD extracted by G-REX.

For this problem, however, the decision tree found by C&RT is even smaller; see Fig. 2. It is very interesting to see the striking similarity; both trees use almost the same tests near the root.

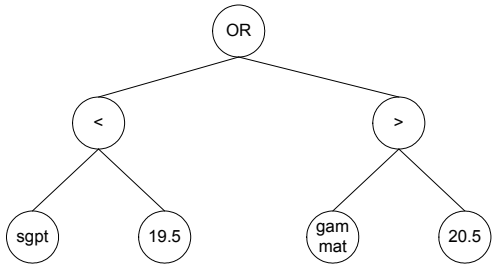

Fig. 2: Rule for BLD found by C&RT.

PIMA Indians Diabetes
Most techniques show good accuracy with the ensembles on top. G-REX comes up with an almost trivial rule; see Fig. 3. At the same time the decision tree found by C&RT contains 74 internal nodes and has worse accuracy, compared to G-REX.

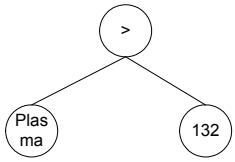

Fig. 3: Rule for PID extracted by G-REX.

Image segmentation
For this data set most techniques have very good accuracy, over 90%. Despite this the relationship is far from uncomplicated; the C&RT-tree has 15 internal nodes. G-REX has trouble finding an accurate tree for this problem. When forced to look for shorter rules the result is a rule with nine internal nodes and accuracy no better than 76%.

StatLog vehicle silhouette
On this problem the ensembles are much more accurate than most of the individual techniques. G-REX fails to find a really short rule with good accuracy. The reported rule contains 19 internal nodes. For comparison the C&RT rule has 96 internal nodes. When allowed to search for longer rules G-REX does find a more accurate rule (77% accuracy) but then the size is comparable to the C&RT-tree.

Waveform
The neural networks excel on this data set and the ensembles again show good accuracy. G-REX does find a rather short rule with a higher accuracy compared to C&RT; see Fig. 4. C&RT produces a tree with intimidating 334 interior nodes.

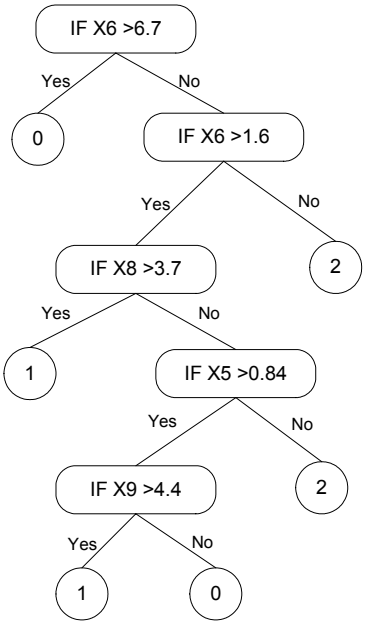

Fig. 4: Rule for WAV extracted by G-REX.

6 Conclusions

The most interesting result of this study is perhaps that the ensembles overall produce very accurate predictions. In fact for all experiments, with the exception of SEG, the accuracy is close to the best accuracy reported by Lim, Loh and Shih in [16].

G-REX is generally capable of finding short and accurate rules. For some of the problems, however, the loss of accuracy would probably not be acceptable. On the other hand G-REX was in this study forced to look for really short rules. Obviously G-REX could produce more accurate rules but that would seriously hamper the comprehensibility. A comparison between G-REX and C&RT shows that G-REX accuracy on one of the problems is significantly lower than C&RT. Leaving this problem out, the accuracy is almost identical, but the number of nodes in the decision trees produced by G-REX is significantly lower compared to those produced by C&RT. We argue that this is an important aspect and that this leads to superior comprehensibility, which is a property that must have high priority within data mining. If this is not the case then there is really no motivation for rule extraction in the first place!

7 Discussion and future work

This paper has not supplied a solution to the problem of the tradeoff between accuracy and comprehensibility, but we have taken a small step to bridge this gap. The G-REX algorithm (and others which penalize complexity) should be developed further, with the clear goal to narrow this gap as much as possible. This is a vital question for those who use data mining for informed decision support.

G-REX has in this study extracted Boolean rules and decision trees. The relatively free choice of different representation language is an important property of G-REX. With this in mind different representation languages should be evaluated, both on performance and experienced clarity.

References

[1] R. Davis, B. G. Buchanan and E. Shortliffe, Production rules as a representation for a knowledge-based consultation program, *Artificial Intelligence*, Vol8. No 1, pp. 15-45, 1977.

[2] R. Andrews, J. Diederich and A. B. Tickle, A survey and critique of techniques for extracting rules from trained artificial neural networks, *Knowledge-Based Systems*, 8(6), 1995.

[3] M. J. A. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales and Customer Support*, Wiley, 1997.

[4] M. Craven and J. Shavlik, Rule Extraction: Where Do We Go from Here?, *University of Wisconsin Machine Learning Research Group Working Paper 99-1*, 1999.

[5] U. Johansson, L. Niklasson and R. König, Rule Extraction from Trained Neural Networks using Genetic Programming, *Supplementary proceedings 13th International Conference on Artificial Neural Networks*, Istanbul, Turkey, pp. 13-16, 2003.

[6] J. Dorado, J. R. Rabunãl, A. Santos, A. Pazos and D. Rivero, Automatic Recurrent and Feed-Forward ANN Rule and Expression Extraction with Genetic Programming, *Proc. 7th International Conference on Parallel Problem Solving from Nature*, Granada, Spain, September 2002.

[7] J. Koza, *Genetic Programming - On the Programming of Computers by Natural Selection*, seventh printing, MIT Press, 2000.

[8] J. R. Quinlan, See5 version 1.16, www.rulequest.com, 1998.

[9] U. Johansson, C. Sönströd, R. König and L. Niklasson, Neural Networks and Rule Extraction for Prediction and Explanation in the Marketing Domain, *Proc. The International Joint Conference on Neural Networks,* IEEE Press, Portland, OR, pp. 2866-2871, 2003.

[10] U. Johansson, L. Niklasson and R. König, The Truth is in There - Rule Extraction from Opaque Models Using Genetic Programming, FLAIRS 04, Miami, FL, To appear.

[11] J. A. Hartigan, *Clustering Algorithms*, John Wiley & sons, 1975.

[12] D. Biggs, B. de Ville and E. Suen, A method for choosing multiway partitions for classification and decision trees, *Journal of Applied Statistics*, 18:49-62, 1991.

[13] D. E. Rumelhart and J. McClelland, editors. *Parallel Distributed Processing*, Vol. 1, MIT press, 1998.

[14] D. S. Broomhead and D. Lowe, Multivariate functional interpolation and adaptive networks, *Complex Systems*, 2:321-355, 1998.

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

[16] T.-S. Lim, W.-Y. Loh and Y.-S. Shih, A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms, *Machine Learning*, 40:203-229, 2000.