Interactive Visualisation for Predictive Modelling with Decision Tree Induction

Tu Bao Ho, Trong Dung Nguyen

Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa, 923-1292 JAPAN

Abstract. In this paper we describe system CABRO for decision tree induction (DTI) that contributes to the combination of machine learning, visualisation and model selection techniques. We first discuss some issues in data mining and briefly introduce R-measure for attribute selection problem in DTI. We then present the DTI interactive visualisation system CABRO, based on R-measure and a combination of several DTI techniques, in which we focus on solutions to two problems: (1) support for understanding of large decision trees, and (2) support for interactive learning and model selection.

1 Introduction

For the KDD predictive modelling task, *decision tree induction* (DTI) is certainly the most applicable technique due to its power and simplicity. The performance of a DTI system depends strongly on its methods to solve three key problems: *attribute selection*, *pruning*, and *discretisation*. For the problem of attribute selection, most measures used in DTI are based on either information theory such as information gain, gain-ratio [12], or statistics such as χ^2 , gini-index [2]. In [5], a measure based on rough set theory, called *R-measure*, was introduced. In this paper we first briefly present R-measure (section 2), and lately by using tools for model selection we show experimentally that R-measure can be used effectively in KDD (subsection 3.1).

Though decision trees are a simple notion it is not easy to understand and analyse large decision trees generated from huge data sets. For example, the widely used program C4.5 [12] produces a decision tree of nearly 18,500 nodes with 2624 leaf nodes from the census bureau database given recently to the KDD community that consists of 199,523 instances described by 40 numeric and symbolic attributes (103 Mbytes). It is extremely difficult for the user to understand and use that big tree in its text form. In such cases, a graphic visualisation of discovered decision trees with different ways of accessing and viewing trees is of great support and recently it receives much attention from the KDD researcher and user. In [7] the authors developed an interactive-graphic environment for automatic generation of decision trees. In [13], the process of inducing bagged decision trees is visualised by CAT scan (classification aggregation tablet). System MineSet provides a 3D visualisation of decision trees [3]. The interactive

visualisation system CABRO described in this paper, associated with a *tree visualiser*, presents an alternative efficient way that allows the user to manipulate graphically and interactively trees in the learning process (subsection 3.1).

There exist various available DTI methods for solving three main DTI key problems mentioned above. It raises in practice the problem of *model selection* that is how to choose the most appropriate DTI methods/models for a given application task. Currently, it is a matter of consensus that there are no universally superior models and methods for DTI. The key question of which DTI method is better than others for a given task has still no answer. Such an answer requires meta-knowledge and/or empirical comparative evaluations of models. Even when meta-knowledge is available, empirical evaluations are always necessary for model selection. Though multiple training-and-testing experiments on a large number of data sets, for empirical evaluations of datamining results, provide reliable evaluations they have not always been done well [14]. Another motivation of this work is to develop in CABRO an *automatic and interactive evaluation* tool based on a well designed benchmark, that supports the user to carry out experimental comparative evaluations of DTI methods/models in order to select the most appropriate ones for a given task (subsection 3.2).

Though this work is about DTI, its general principles can be applied to other KDD tasks, e.g., to clustering as done in clustering system OSHAM [6].

2 R-measure for attribute selection

Rough set theory introduced by Pawlak in early 1980s is a mathematical tool to deal with imprecise and incomplete information [10]. The *lower* and *upper* approximations of any $X \subseteq O$, regarding an equivalence relation E , are defined as

$$E_*(X) = \{o \in O : [o]_E \subseteq X\} \quad (1)$$

$$E^*(X) = \{o \in O : [o]_E \cap X \neq \emptyset\} \quad (2)$$

where $[o]_E$ denotes the equivalence class of objects which are indiscernible with o with respect to the equivalence relation E . A key concept in rough set theory is the *degree of dependency* of a set of attributes Q on a set of attributes P , denoted by $\mu_P(Q)$ ($0 \leq \mu_P(Q) \leq 1$), defined as

$$\mu_P(Q) = \frac{\text{card}(\bigcup_{[o]_Q} P_*([o]_Q))}{\text{card}(O)} = \frac{\text{card}(\{o \in O : [o]_P \subseteq [o]_Q\})}{\text{card}(O)} \quad (3)$$

If $\mu_P(Q) = 1$ then Q totally depends on P ; if $0 < \mu_P(Q) < 1$ then Q partially depends on P ; if $\mu_P(Q) = 0$ then Q is independent of P . The measure of dependency is fundamental in rough set theory as based on it many basic notions are defined, such as reducts and minimal sets of attributes, attribute significance, etc.

The limitation of the deterministic model of rough sets when dealing with uncertain information has been recognised and studied, e.g., by probabilistic model [11] and the variable precision model [15]. However, the former cannot

inherit all useful properties of the original rough set model, and the latter raises a new problem of specifying an appropriate threshold. R-measure, inspired by the attribute dependency measure $\mu_P(Q)$, is an alternative one to deal with uncertain information while preserving properties of the rough set model without requiring thresholds. In brief, R-measure measures the dependency of Q on P in maximizing the predicted membership of an instance in the family of equivalence classes generated by Q given its membership in the family of equivalence classes generated by P . It is initially formulated by

$$\mu'_P(Q) = \frac{1}{\text{card}(O)} \sum_{[o]_P} \max_{[o]_Q} \text{card}([o]_Q \cap [o]_P) \quad (4)$$

Theorem For every sets P and Q of attributes we have

$$\frac{\max_{[o]_Q} \text{card}([o]_Q)}{\text{card}(O)} \leq \mu'_P(Q) \leq 1 \quad (5)$$

An analysis based on the notion of impurity function provides a basis to go from μ'_P to $\tilde{\mu}$ for DTI in order to deal better with large data sets. We have the following general form of the measure for arbitrary attribute sets P and Q and called it *R-measure*

$$\tilde{\mu}_P(Q) = \frac{1}{\text{card}(O)} \sum_{[o]_P} \max_{[o]_Q} \frac{\text{card}([o]_Q \cap [o]_P)^2}{\text{card}([o]_P)} \quad (6)$$

R-measure is applied to selecting attributes in DTI by considering Q as the class attribute and P a descriptive attribute. System CABRO provides the user a set of tools based on techniques to solve three main DTI problems: the gain-ratio [12], gini-index [2], χ^2 [8] and R-measure for attribute selection; the error-complexity [2], reduced-error and pessimistic error [9] for pruning; and the entropy-based and error-based discretisation [4] for discretisation. Generally, these techniques are independent and the user can combine them to obtain different DTI models.

3 An Interactive Visualisation System

KDD is an iterative cycle in which the system provides and receives interactively feedback to and from the user. To support the learning and understanding decision trees, CABRO uses a *Visual Interactive Model* (VIM) through a rich graphical environment. The VIM aims at combining “meaningful pictures and easy interactions to stimulate creativity and insight; promoting a process of ‘generate and test’, it facilitates a rapid cycle of learning” [1]. In CABRO, the VIM offers: (1) a *Tree Visualiser* that supports the understanding and analysing decision trees; (2) an *Interactive Mode* that supports the user participation in the model selection and the learning process. CABRO has been implemented in UNIX workstations under the X Window, and recently in MS Windows 95, particularly on super-power PCs with 384 MB of SDRAM.

CABRO supports the user to load a data set, to select and evaluate DTI models, to induce decision trees, to navigate and analyse generated trees with different views and operations, to match decision trees with unknown cases.

3.1 Tree Visualiser

The *Tree Visualiser* is a graphical interface that displays a tree in a graphical form of a set of nodes and connections. Its most important feature is *the transformation of a decision tree from a static form (text) to a dynamic form (graphic) on which a certain operations can be carried out*. The user can dynamically navigate through the tree and have different views on generated trees, switch among several view modes, choose alternatively parts of the tree or paths to one class to be focused on. The basic views and operations in the Tree Visualiser are:

- *View of the tree structure*: Initially, a tree is displayed with only the root node and its direct sub-nodes (top-left window, Figure 1). The tree then can be *collapsed* or *expanded* partially or fully from the root or from any decision node. It allows the user to manipulate even a large tree by expand it from the root node step by step, either level by level (breadth-first) or branch by branch (depth-first). The user can exploit different *multiple views* on parts of the tree, such as to view only a subtree of interest from a decision node, to collapse some nodes (denoted by a $\boxed{+}$) and/or expand some other ones (denoted by a $\boxed{-}$) in order to hide non-interest subtrees and to focus on interest subtrees, etc.

Fig. 1. CABRO's Tree Visualiser

- *View of decision/leaf nodes*: The user can click on a node to see its information: branching attribute and branched attribute-value, number of covered

cases, the major class, the percentage of major class, the path leading from the root to this decision/leaf node (dark color in boxes for class “brown-spot”, Figure 2), etc.

- *View of a class on the tree*: The user can focus on observing decision paths to leaf nodes of a class. Only these leaf nodes are highlighted and proportions of cases bearing this class label are indicated approximately in decision nodes.

Figure 1 presents two different collapsed and expanded views of a tree.

3.2 Model Selection and Interactive Learning

Model Selection

There are three main criteria for selecting DTI models – size, accuracy and understandability of trees [8]. The understandability of trees is difficult to be quantified or measured and a tree visualiser will be of great support for the judgements of domain experts or users. The tree size and accuracy can be quantitatively evaluated, among them the accuracy is widely considered to be of great importance. The *k-fold stratified cross validation* has been recommended as a more reliable method for evaluating learning models, but it is time-consuming and not always be done well [14]. Doing carefully evaluations of discovered knowledge with good benchmarks has to become a principle in KDD.

Fig. 2. Support for Model Selection

The CABRO’s model selection tool, using the *k-fold stratified cross validation* (by default, $k = 10$), supports the user to select an appropriate DTI

technique/model from the candidate ones by automatically generating and evaluating decision trees induced by the considered techniques/models. In the current version of CABRO, the user can try a new model composed by an attribute selection measure chosen from the gain-ratio, the gini-index, χ^2 and R-measure; a pruning technique from error-complexity, reduced-error and pessimistic error; and a discretisation technique from the entropy-based and error-based techniques (Figure 2). The open structure of program CABRO permits to add new DTI techniques to the set of component techniques. When a model candidate is considered, CABRO generates the corresponding decision tree and evaluates its performance as follows:

- The system carries out automatically a random shuffle of the data set then divides the data set into 10 mutually exclusive subsets (folds) of approximately equal size and the same proportions of labels as in the original data set.
- The system carries out automatically a run of 10-fold cross validation for the considered model, where one run is 10 times of applying the model to training data (the union of 9 subsets) for inducing a decision tree, and to testing data (the rest subset) for obtaining one estimation of the model predictive accuracy and tree size. It evaluates the final estimation of the model accuracy and tree size as the average of those obtained from 10 runs, and adds these into the “result” table (Figure 2).

This process is repeated for different model candidates with the same division of the original data set into 10 folds for the sake of a fair evaluation. For each model candidate, CABRO visualizes graphically the corresponding *pruned tree*, *its size*, *its prediction error rate*. CABRO offers the user a multiple view of these trials and facilitates the user to compare results of trials in order to make his/her final selection of techniques/models of interest.

Interactive Learning

In addition to the ability of interacting with the system by the Model Selection and Tree Visualiser tools, CABRO also provides the user the interaction during the generating-and-test process to obtain more appropriate results. An *Interactive Learning Mode* (ILM) allows the user participate actively in the DTI process. In fact, the user can execute a full automatic learning of decision trees without intervention by pressing the “run” button, or execute a step-by-step learning of decision trees by pressing the “step” button, or combine these two modes (Figure 2). The ILM provides currently two interactive functions:

- *interactively selecting an attribute* to branch decision nodes from several promising evaluated attributes. There are two reasons for doing that: (1) impurity functions are independent at each decision node on a tree then it is reasonable and better to evaluate the impurity of decision nodes by several measures and the selected attribute can be obtained by a plurality

vote among them, and (2) it may happen that a considering measure finds some attributes whose values are slightly different from the best one, and it is not necessarily to select the attribute with biggest value to branch the node. By displaying this information (Figure 3) CABRO supports the user to combine his/her domain knowledge with the automatically evaluated attributes by system. This opportunity is meaningful, particularly at several first levels or at a branch of interest on the decision tree.

- *backtracking and regrow the tree*: while step-by-step generating a decision tree, the user can easily examine intermediate trees, backtracking to re-grow the tree at some node with respect to the induction scheme in changing parameters (e.g. the minimum size of a node or the expected accuracy of a leaf). Thus, the iteration can be done partially on the tree.

Fig. 3. Support for Interactive Learning

Illustration: Evaluation of R-measure

We have used the Model Selection tool to evaluate the performance of R-measure in comparison with three widely used attribute selection measures: the gain-ratio, the gini-index and χ^2 . We carried out an experimental comparative evaluation on 18 data sets from the University California Irvine repository of databases for four selected models. These are formed by combining fixed methods of error-complexity pruning and entropy-based discretisation with gain-ratio (c), gini-index (g), χ^2 (χ) and R-measure (R).

Table 1 presents experimental results on size and error rates (both unpruned and pruned trees) of these four measures obtained automatically. As error rates of pruned trees are of most importance (columns 4 and 8), we indicate the lowest

values on each by bold numbers. The other columns are the pruned tree sizes, error rates and sizes before pruning. It is worth noting that it will be very time-consuming and difficult to carry out precisely all of these experiments without such an automatic tool. This evaluation results show that R-measure is comparable to the best attribute selection measures in the current data mining literature.

	unpruned		pruned			unpruned		pruned	
	size	errors	size	errors		size	errors	size	errors
Vote, 16x300, symbolic					Cancer, 9x700, symbolic				
c	22.6±2.5	7.3±3.6	4.0±0.0	5.0±2.8	87.9±23.1	7.3±2.2	46.1±19.1	7.4±2.9	
g	24.7±2.8	7.0±3.0	7.0±4.2	5.9±2.7	92.3±26.0	6.9±2.3	36.2±11.9	7.4±3.5	
χ	24.7±2.8	7.0±3.0	7.0±4.2	5.9±2.7	92.3±26.0	6.9±2.0	36.2±11.9	7.4±3.5	
R	25.0±3.0	7.5±3.4	5.8±2.9	5.7±2.7	94.5±26.4	7.0±2.2	37.3±11.2	7.1±3.4	
Shuttle, 9x956, symbolic					Promoters, 45x105, symbolic				
c	88.4±10.9	0.2±0.1	53.4±15.8	0.2±0.1	18.2±3.3	25.5±9.8	9.8±4.3	24.5±7.5	
g	144.8±6.6	0.2±0.1	114.2±12.2	0.2±0.1	19.0±4.0	23.6±12.7	9.4±3.7	22.7±10.0	
χ	199.0±17.2	0.3±0.1	162.7±30.4	0.3±0.1	19.0±4.0	23.6±10.9	9.4±3.7	22.7±10.0	
R	165.3±15.6	0.2±0.1	135.3±18.3	0.3±0.1	19.0±4.0	23.6±12.7	9.4±3.7	22.7±10.0	
Solar Flare, 12x1286, symbolic					Diabetes, 8x768, numeric				
c	104.0±11.6	26.8±2.5	26.8±7.8	25.3±1.5	41.2±2.2	24.4±3.1	18.2±9.4	25.3±2.6	
g	150.8±15.0	28.4±2.9	54.4±15.0	27.8±1.3	53.0±4.4	25.3±2.7	22.0±4.8	25.6±2.5	
χ	168.6±24.6	28.3±2.2	45.8±18.1	26.6±2.0	47.6±5.1	25.3±2.7	13.6±6.2	25.5±2.5	
R	155.0±19.4	26.9±1.8	44.6±31.1	25.5±1.0	74.2±8.6	24.7±2.6	27.8±19.8	25.3±2.6	
Splice, 45x3189, numeric					Glass, 9x214, numeric				
c	529.8±68.0	10.2±1.5	245.8±36.8	8.0±1.7	21.0±3.2	33.2±8.6	17.3±5.5	34.5±8.2	
g	565.8±72.0	10.4±2.4	214.6±39.8	8.4±1.8	35.2±4.4	33.2±7.7	22.3±6.9	36.6±6.8	
χ	585.8±76.0	10.5±2.4	253.0±56.8	8.8±1.7	29.6±2.8	35.0±5.6	19.1±7.5	37.3±6.4	
R	569.8±77.4	11.0±2.5	207.4±30.4	8.6±1.9	32.2±5.8	34.1±7.3	18.7±6.8	35.9±6.9	
Waveform, 36x3195, symbolic					Heart Disease, 13x270, mixed				
c	1148.3±179.5	28.9±1.7	223.9±72.9	25.7±1.1	13.8±5.4	25.6±4.1	8.8±3.8	25.6±4.1	
g	1320.5±193.9	27.8±1.3	244.5±69.5	24.4±1.6	33.0±4.0	27.4±4.7	25.8±2.6	25.6±5.6	
χ	1355.7±185.0	28.4±1.6	340.6±191.6	26.8±1.3	26.6±9.7	27.4±4.7	9.0±3.2	26.3±4.9	
R	1432.3±193.5	29.3±1.5	249.6±78.4	25.1±1.1	38.0±13.4	27.4±4.7	8.2±5.1	25.2±4.6	
Vehicle, 18x846, numeric					Hypothyroid, 25x3163, numeric				
c	174.5±35.9	32.4±5.2	131.9±40.7	32.7±5.1	22.6±2.5	1.1±0.4	11.8±1.5	0.9±0.4	
g	222.8±38.4	31.9±3.2	111.4±37.8	32.0±3.7	49.2±5.4	1.3±0.5	16.8±3.7	0.9±0.4	
χ	216.2±40.4	30.2±3.9	111.4±47.4	31.9±3.2	54.8±5.0	1.3±0.5	10.6±0.6	0.9±0.4	
R	218.2±39.6	31.6±3.2	101.5±28.7	31.8±3.5	57.8±6.0	1.4±0.4	18.2±4.9	0.9±0.4	
Audiology, 70x226, symbolic					Cars, 8x392, numeric				
c	49.8±9.0	29.6±13.7	28.4±13.3	30.9±11.0	32.3±2.0	24.8±4.8	17.1±9.5	26.0±2.0	
g	68.2±12.4	29.6±11.5	37.0±16.0	30.9±11.9	44.7±10.3	24.0±4.8	21.4±8.5	26.8±5.2	
χ	93.9±22.3	44.3±8.9	66.9±14.9	45.2±8.7	41.8±8.8	23.8±5.0	17.4±9.1	26.5±5.2	
R	72.1±10.3	28.3±10.9	41.3±13.4	29.1±11.7	44.6±12.5	24.2±5.1	21.8±12.8	25.2±4.8	
Horse-colic, 28x368, numeric					Pima-diabetes, 8x768, numeric				
c	48.9±9.1	16.2±3.8	8.2±4.1	14.3±5.1	34.3±6.7	24.9±4.5	17.6±5.8	23.4±3.6	
g	86.4±19.1	17.8±1.9	30.9±20.3	16.8±3.5	45.4±3.5	24.7±4.2	25.4±8.3	23.5±3.5	
χ	92.0±25.8	18.1±2.3	22.0±22.4	17.0±3.3	40.0±4.2	24.7±4.2	18.5±9.0	23.5±3.5	
R	115.6±22.7	17.0±1.7	15.8±13.7	15.9±4.2	65.1±8.5	24.7±4.2	30.6±17.4	23.9±3.2	
Segmentation, 19x2310, numeric					Iris, 4x150, numeric				
c	327.4±48.2	6.3±1.5	236.4±46.5	6.2±1.6	4.3±0.5	4.0±3.2	4.0±0.0	3.3±3.3	
g	341.3±30.9	5.9±1.7	257.5±81.2	6.1±2.0	4.3±0.5	3.3±3.3	4.0±0.0	2.7±3.2	
χ	373.2±25.2	7.3±1.6	310.7±48.3	7.6±2.0	4.3±0.5	4.7±3.7	4.0±0.0	4.0±4.0	
R	342.5±33.6	6.1±1.8	272.0±90.4	6.1±2.1	4.3±0.5	4.7±3.7	4.0±0.0	4.0±4.0	

Table 1. Experimental comparative evaluation results

4 Conclusion

We have described the interactive visualisation system CABRO for DTI. The system employs several DTI well-known techniques selected from the KDD and

machine learning literature, including our proposed R-measure for attribute selection. Implemented as an interactive-graphic environment, CABRO supports a creative combination of contributing techniques, in particular the tree visualisation, the model selection, the interactive learning which are certainly important principles of data mining and knowledge discovery. Using the model selection tool, we have carried out carefully a comparative evaluation that shows the high application potential of CABRO and R-measure.

References

1. Belton, V. and Elder, M.D.: Decision Support Systems: Learning from Visual Interactive Modelling, *Decision Support Systems*, 12, pp. 355–364 (1994).
2. Breiman, L., Friedman, J., Olshen, R. and Stone, C.: *Classification and Regression Trees*, Belmont, CA: Wadsworth (1984).
3. Brunk, C., Kelly, J. and Kohavi, R.: MineSet: An Integrated System for Data Mining, *Proceedings of Third International Conference on Knowledge Discovery and Data Mining*, pp. 135–138 (1997).
4. Fayyad, U.M. and Irani K.B.: On the Handling of Continuous-Valued Attributes in Decision Tree Generation, *Machine Learning*, pp. 87–102 (1992).
5. Ho, T.B., Nguyen, T.D. and Kimura, M.: Induction of Decision Trees Based on the Rough Set Theory, in *Data Science, Classification and Related Methods*, C. Hayashi et al. (Eds.), Springer, 215–222 (1997).
6. Ho, T.B.: Discovering and Using Knowledge From Unsupervised Data, *Journal Decision Support Systems*, Elsevier Science, Vol. 21, No. 1, 27–41 (1997).
7. Kervahut, T. and Potvin, J.Y.: An Interactive-Graphic Environment for Automatic Generation of Decision Trees, *Decision Support Systems*, Vol. 18, pp. 117–134 (1996).
8. Mingers, J.: An Empirical Comparison of Selection Measures for Decision Tree Induction, *Machine Learning*, 3, pp. 319–342 (1989).
9. Mingers, J.: An Empirical Comparison of Pruning Methods for Decision Tree Induction, *Machine Learning*, 4, pp. 227–243 (1989).
10. Pawlak, Z.: *Rough sets: Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers (1991).
11. Pawlak, Z., Wong, S.K.M. and Ziarko, W.: Rough Sets, Probabilistic versus Deterministic Approach, *International Journal of Man-Machine Studies*, 29, pp. 81–95 (1988).
12. Quinlan, J.R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993).
13. Rao, J.S., Potts, W.J.E.: Visualizing Bagged Decision Trees, *Proceedings of Third International Conference on Knowledge Discovery and Data Mining*, pp. 243–246 (1997).
14. Salzberg, S.L.: On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach, *Data Mining and Knowledge Discovery*, 1, 317–327 (1997).
15. Ziarko, W.: Variable Precision Rough Set Model, *Journal of Computer and System Sciences*, 46, pp. 39–59 (1993).