Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)

Effects of attribute selection measures and sampling policies on functional structures of decision trees

H. Du, S. Jassim, M. F. Obatusin

Department of Computer Science
University of Buckingham, UK

Abstract

This paper is concerned with a comparative study of the most commonly used attribute selection measures in the construction of decision trees. We examine the effect of these measures on the resulting tree structures against various sampling policies. The emphasis of earlier works in this field has been on the overall size of the tree in terms of the number of levels and the number of leaf nodes. We take a more informative view, encompassing the functionality of decision trees into tree structures. The proposed evaluation criterion combines classification proportion with the combinatorial structure. Our experiments demonstrate that the information-based measures outperform the non-information based ones for unpruned trees against classification proportion thresholds and most sampling policies. Among the information-based measures, the information gain appears to be the best. Pruning improves the performance of statistics-based measures. We also show that there are optimal combinations between attribute selection measures and sampling policies regarding to the best achievable classification thresholds.

1. Introduction

Data mining refers to the process of discovering useful information patterns implicitly embedded in databases. Various data mining approaches and solutions have been developed [1]. Machine learning methods are widely used for constructing models to classify data items into a group of predefined classes. A well-known category of machine learning methods build decision trees as the classification model.

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)

A decision tree takes attributes as its non-leaf internal nodes, class labels as leaf nodes and attribute values as links between the nodes. A decision tree is built with a set of examples, each of which consists of a list of descriptive attribute values and a known class label. The set of examples is divided into a *training set*, a *test set* and an *evaluation set* according to certain sampling policy. An initial tree is constructed using the examples in the training set. In order not to *overfit* the tree to the training examples [2], the tree is tested and pruned using the examples in the test set. The accuracy of the decision tree is measured by error rate in classifying the examples from the evaluation set.

Understandingly, accuracy of classification has been the main focus of comparative studies of the effect attribute selection measures on the resulting decision trees ([3], [4], [5], [6]). However, there has been a considerable interest on the effect of these measures on the combinatorial structure of decision trees evaluated in terms of the total number of levels and the total number of leaf nodes [4, 7]. This traditional concept of structure cannot capture the true desirable features of decision trees. Taking any two trees with different number of levels, balanced or unbalanced, it is difficult to determine which tree is more desirable for classification purposes. A balanced tree with less number of levels may not necessarily be better performing than an unbalanced tree with more number of levels. Decision trees are constructed for the sole purpose of classification, and it seems strange not to encapsulate classification into a meaningful concept of tree structure.

In this paper, we propose a concept for *functionality structure* that combines the functionality of a decision tree with its underlying combinatorial structure. This concept associates with each level of a decision tree a classification value which represents the proportion of examples belonging to the evaluation set that have been classified down to that level. Our underlying philosophy is that a tree that can classify most of the records at levels nearer to the root should be more desirable than other trees. For instance, a decision tree for medical diagnostics, whose attributes relate to medical tests, that has this property helps in reducing the number of tests for the majority of patients, and thereby reducing the cost of decision making. We investigate the combined effects of selection measures and sampling policies on the functionality structure of the resulting tree as determined number of classification thresholds.

The rest of this paper is organised as follows. Section 2 briefly reviews some common decision tree construction algorithms and attribute selection measures. Section 3 describes the design of our experiments. The results are presented and analysed in section 4. Section 5 concludes with major findings and future work.

2. Background

2.1. Decision tree construction algorithms

There are many algorithms for constructing decision trees. Among the most common are CART [8], CHAID [9], ID3 [10] and C4.5 [11]. Although they differ on certain detailed steps and/or the orders of the result trees, these

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)

algorithms have the same procedural framework. According to a certain *selection measure*, an attribute is selected as the root of the tree. The training set is then partitioned into subsets by the values of the selected attribute. For each subset, if all examples are of the same class, a leaf node with the class label is created and a branch from the root to the leaf, labelled with an appropriate value, is formed. If not all examples in the subset are of the same class, the same selection step is repeated for the subset. The whole construction process terminates when all examples in the original training set have been "classified" with leaf nodes, and a complete tree is built.

The single most important aspect that determines the specific behaviour of a decision tree construction algorithm is the attribute selection measure [12]. CART uses the *Gini index* of diversity, ID3 uses the *information gain*, and C4.5 uses the *gain ratio* while CHAID uses the χ^2*-test*. These measures are expected to result in more concise trees than that by a simple random selection of attributes.

The problem of overfitting is dealt with by *pruning* the decision tree. A number of pruning methods have been developed [2, 13]. In most cases, pruning is done after a decision tree is constructed (*post-pruning*) although it is possible to prune the tree as it grows (*pre-pruning*). Post-pruning is considered to be better than pre-pruning because pre-pruning may prevent the inclusion of potentially important attributes [6]. We use the reduced error pruning, a simple post-pruning method. The empirical results show that this method is comparable with other pruning methods with respect to the accuracy of the pruned tree.

2.2. Attribute selection measures

Existing attribute selection measures are either information-based using concepts of probability and information theory, or statistics-based using hypotheses testing frameworks. Below, we briefly describe four measures covering both categories.

2.2.1 Information gain, an information-based measure

An information system is a system of events and associated probabilities. In a relational table, an attribute A is an information system with m distinct values a_1, a_2, ..., a_m as the possible events. For each i ($1 \le i \le m$), the probability of a_i, $P(a_i)$, is the proportional frequency of a_i to the total number of tuples in the table. The average of the self-information of all events within the information system A is called the *entropy* or the *expected information*, and defined as follows:

$$H(A) = \sum_{i=1}^{m} P(a_i) \log_2 \frac{1}{P(a_i)} = -\sum_{i=1}^{m} P(a_i) \log_2 P(a_i)$$

Hence, the entropy for the Class attribute with w classes (C_1, C_2, ..., C_w) is

$$H(Class) = \sum_{i=1}^{w} P(C_i) \log_2 \frac{1}{P(C_i)} = -\sum_{i=1}^{w} P(C_i) \log_2 P(C_i)$$

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)

The expected information of classification when an attribute A with values a_1, a_2, ..., a_m is chosen as the root of the current decision tree is the *conditional entropy* $H(Class|A)$ and defined as:

$$H(Class|A) = \sum_{i=1}^{w}\sum_{j=1}^{m} P(C_i \cap a_j)\log_2 \frac{1}{P(C_i \mid a_j)} = -\sum_{i=1}^{w}\sum_{j=1}^{m} P(C_i \cap a_j)\log_2 \frac{P(C_i \cap a_j)}{P(a_j)}$$

where $P(C_i \cap a_j)$ is the probability of attribute A having the value a_j and class C_i. The information gain on attribute A is the mutual information that exists between the attribute Class and the attribute A. It is calculated as

$$Gain(A) = H(Class) - H(Class \mid A).$$

2.2.2 Information gain ratio, an information-based measure

Information gain ratio on attribute A is the ratio of the information gain on A over the expected information of A, normalising uncertainty across attributes. It is defined as

$$Gain\ Ratio(A) = \frac{H(Class) - H(Class|A)}{H(A)}.$$

2.2.3 Gini index, a statistics-based measure

Gini function measures the impurity of an attribute with respect to classes. The impurity function is defined as:

$$Gini(t) = 1 - \Sigma\, p_i^2$$

where t refers to an attribute or the Class attribute, and p_i is the frequency of a specific class. For the Class attribute with w classes, the impurity is:

$$Gini(Class) = 1 - \sum_{i=1}^{w} P(C_i)^2$$

For an attribute A with m distinct values a_2, ..., a_m, the impurity of $A = a_j$ with respect to the classes is

$$Gini(A = a_j) = 1 - \sum_{i=1}^{w} P(C_i \cap a_j)^2$$

The Gini index of A, defined below, is the difference between the impurity of Class and the average impurity of A regarding to the classes, representing reduction of impurity over the choice of attribute A.

$$GiniIndex(A) = Gini(Class) - \sum_{j=1}^{m} P(a_j)Gini(A = a_j)$$

2.2.4 Chi-square (χ^2) statistic, a statistic-based measure

Chi-square (χ^2) statistic is a measure of the degree of association or dependence between attribute values and classes. The χ^2 function calculates the differences between the actual frequencies of classes in an attribute with the expected frequencies when no association between that attribute and class is assumed. The greater the differences, the stronger the association between the classification and the chosen attribute. The basic expression is given as follows:

$$\chi^2 = \sum\sum \frac{(x_{ij} - E_{ij})^2}{E_{ij}}$$

where x_{ij} represents the actual frequency that examples have attribute value a_j and class c_i, and E_{ij} represents the expected frequency.

3. Experiments

Our experiments aim to assess the effects of different attribute selection measures on the classification proportions at each level of the decision trees. We do so for both pruned and unpruned trees.

The data sets used are collected from the UCI repository for machine learning [14]. We have selected 4 data sets of different sizes and from different application domains. For simplicity of implementation, all examples in the data sets have only categorical attributes without missing values. Table 1 summarises the main characteristics of these data sets.

The examples are selected from a given data set by using the proportional random sampling technique according to a sampling policy. Different sampling policies have been reported in the literature ([2, 4, 5]). Table 2 lists details of all the policies that have been used in our experiments.

Table 1. Example Data Sets for Experiments

Name	No. of Classes	No. of Attributes	No. of Examples
Vote	2	16	435
Tic-tac-toe	2	9	958
Car	4	6	1728
Nursery	5	8	8683

Table 2. Sampling Policies

Training Set	Testing Set	Evaluation Set
10%	10%	80%
15%	35%	50%
25%	25%	50%
35%	15%	50%
60%	20%	20%

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)

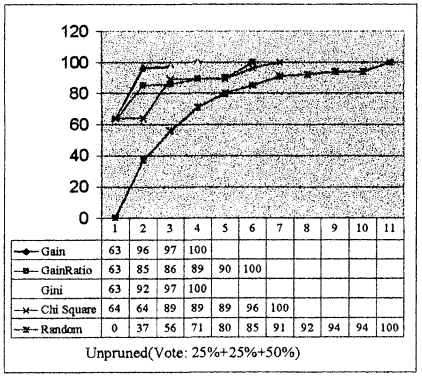
We use a generic implementation of the ID3 algorithm as the procedural framework for constructing a decision tree. The functions for the selection measures are called within the same program, and different decision trees are produced. Data sets are stored in Microsoft Access tables. All selection measures mentioned in section 2 are under investigation. We also use random selection as a default measure against which all others are compared.

For each data set, we follow the given 5 sampling policies. To obtain more accurate results, 3 trials were conducted for each sampling policy. In each trial, we obtain a separate sample of the chosen proportion for a training set, a test set and an evaluation set. For each trial, a decision tree is constructed using each of the selection measures. The trees are then pruned.

4. Experiment Results and Analysis

The experimental work resulted in a large amount of raw data. For each combination of parameters, we recorded a structural description of the decision tree in terms of accumulated classification proportions against the tree levels. We finally obtained 40 tables by averaging results over different trials and grouping them according to the sampling policies and data sets. In most cases, this was straightforward as all the three trials resulted in trees with equal number of levels. In the exceptional cases where different trials produce different numbers of levels, we chose the representative number of levels to be the number nearest to the median of all the numbers of levels in the three trials. There was hardly any distortion on data as a result.

We plotted the content of each table into a chart where each selection measure is presented by a graph showing the accumulated classification proportions against the tree levels. Figure 1 shows two such charts.



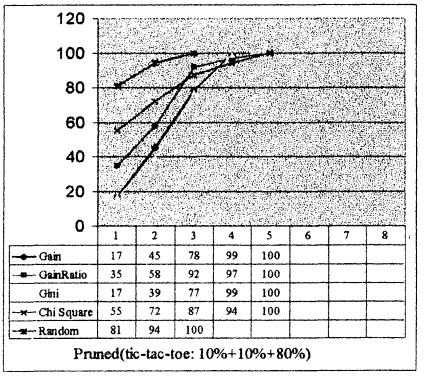


Figure 1. Example Charts of the Experimental Work

Close examinations of these charts reveal interesting and consistent trends and patterns. It is extremely difficult to present and interpret these patterns

without showing all the charts. To overcome the space limitation of this paper, we present our results in an alternative but more convenient and comprehensible way. We define a number of classification thresholds (TH), and for each tree we obtain the levels at which these thresholds are reached. The levels corresponding to each threshold are then grouped in a vector (L_{vote}, $L_{Tic\text{-}Tac\text{-}Toe}$, L_{car}, $L_{nursery}$), where L_{dname} refers to the level at which the given threshold was reached for the data set *dname*. Figure 2 presents 5 tables, one for each sampling policy in the order as specified in Table 2. Each table contains the result data for both the pruned and unpruned trees.

On each row of the unpruned / pruned section of a table, the entry vectors that achieve the minimal total of co-ordinates are marked by bold and underlined fonts. The marking is taken to mean that the given measure has the best performance with respect to the given threshold. The selection measure that has the maximum number of marked vectors in its column is declared as the best overall performing measure for the given sampling policy. The analyses for pruned and unpruned trees are done separately.

TH	Unpruned					Pruned				
	Gain	G. Ratio	Gini	χ^2	Random	Gain	G. Ratio	Gini	χ^2	Random
60%	**(2,3,2,3)**	(2,3,2,4)	**(1,3,2,4)**	**(1,4,2,3)**	(3,4,3,4)	(1,3,2,3)	(1,3,2,3)	(1,3,2,3)	(1,2,2,4)	**(2,4,1,1)**
75%	(3,4,3,4)	**(2,4,3,4)**	(2,4,3,5)	(2,4,3,5)	(4,4,3,5)	(1,3,2,4)	(1,3,2,4)	(1,3,2,4)	(1,3,3,5)	**(2,4,2,1)**
85%	**(3,4,4,4)**	**(2,4,4,5)**	(2,4,4,6)	(3,5,4,5)	(7,5,4,5)	(1,4,4,4)	**(1,3,3,5)**	(1,4,4,4)	**(1,3,3,5)**	**(3,5,3,1)**
90%	**(4,4,4,5)**	**(4,4,4,5)**	**(3,4,4,6)**	**(3,5,4,5)**	(7,5,4,6)	(1,4,4,4)	(1,3,4,5)	(1,4,4,4)	(1,4,4,6)	**(3,5,3,1)**
95%	**(4,4,4,5)**	(6,4,4,5)	(3,4,4,7)	(3,5,4,6)	(8,5,4,6)	**(1,4,4,5)**	**(1,4,4,5)**	**(1,4,4,5)**	(1,5,4,6)	(5,5,4,1)

TH	Unpruned					Pruned				
	Gain	G. Ratio	Gini	χ^2	Random	Gain	G. Ratio	Gini	χ^2	Random
60%	**(2,4,2,3)**	(2,4,2,4)	**(2,4,2,3)**	(2,4,2,4)	(4,4,3,6)	(1,3,2,3)	(1,3,2,4)	(1,3,2,3)	(1,3,2,4)	**(2,2,1,1)**
75%	**(2,4,3,4)**	(2,4,3,5)	**(2,4,3,4)**	(2,5,2,5)	(5,5,4,6)	(1,4,3,4)	(1,4,3,5)	(2,4,3,3)	(1,3,2,5)	**(3,4,1,1)**
85%	**(2,4,4,5)**	**(2,4,4,5)**	**(3,4,4,4)**	(2,5,4,6)	(5,5,4,6)	(1,4,4,4)	(1,4,4,5)	(3,4,4,4)	(1,4,2,6)	**(5,4,1,1)**
90%	**(2,5,4,5)**	**(2,5,4,5)**	(3,5,4,5)	(3,5,4,6)	(6,5,4,6)	(2,4,4,5)	(2,4,4,5)	(3,4,4,5)	(1,4,4,6)	**(5,4,1,1)**
95%	**(3,5,4,5)**	**(2,5,4,6)**	**(3,5,4,5)**	(5,5,4,6)	(6,5,5,6)	(2,4,4,5)	(2,4,4,6)	(3,4,4,5)	(1,5,4,6)	**(6,5,1,2)**

TH	Unpruned					Pruned				
	Gain	G. Ratio	Gini	χ^2	Random	Gain	G. Ratio	Gini	χ^2	Random
60%	(1,4,2,4)	(1,4,2,4)	(1,4,2,4)	**(1,4,2,3)**	(4,5,4,6)	(1,4,2,3)	(1,3,2,4)	(1,4,2,3)	(1,3,2,4)	**(3,1,1,1)**
75%	**(2,4,2,4)**	(2,4,4,5)	(2,4,4,4)	(3,5,2,5)	(5,5,4,6)	(1,4,4,4)	(1,3,4,5)	(1,4,4,4)	(1,4,2,5)	**(3,1,1,1)**
85%	**(2,5,4,5)**	**(2,5,4,5)**	**(2,5,4,5)**	(3,5,4,5)	(6,5,4,6)	(2,4,4,5)	(1,4,4,5)	(2,4,4,5)	(2,5,2,6)	**(4,1,1,1)**
90%	**(2,5,4,5)**	(5,5,4,6)	**(2,5,4,5)**	(6,5,4,6)	(7,6,5,6)	(2,5,4,5)	(1,5,4,6)	(2,4,4,5)	(2,5,4,6)	**(5,2,3,1)**
95%	**(2,5,5,6)**	(6,5,5,6)	(3,5,5,6)	(6,6,5,6)	(11,6,5,7)	(2,5,4,6)	(1,5,4,6)	(3,5,4,6)	(2,5,5,6)	**(5,4,3,1)**

TH	Unpruned					Pruned				
	Gain	G. Ratio	Gini	χ^2	Random	Gain	G. Ratio	Gini	χ^2	Random
60%	**(2,4,2,4)**	**(2,4,2,4)**	**(2,4,2,4)**	(3,5,2,4)	(5,5,4,6)	**(1,3,2,3)**	(1,4,2,4)	**(1,3,2,3)**	(1,3,2,5)	(4,1,1,6)
75%	**(3,5,4,4)**	(3,5,4,5)	(3,5,4,5)	**(4,5,2,5)**	(7,5,4,6)	**(1,4,4,4)**	(1,4,4,5)	**(1,4,4,4)**	**(1,4,2,6)**	(6,4,1,6)
85%	**(3,5,4,5)**	(7,5,4,6)	**(3,5,4,5)**	(5,5,4,6)	(8,6,5,7)	**(1,4,4,5)**	(1,5,4,6)	**(1,4,4,5)**	**(1,5,2,6)**	(7,4,1,7)
90%	**(3,5,5,5)**	(7,5,5,6)	**(3,5,5,5)**	(6,5,5,6)	(8,6,5,7)	(2,5,5,5)	(1,5,5,6)	(2,5,4,5)	**(1,5,3,6)**	(8,5,1,7)
95%	**(3,5,5,6)**	(9,5,5,6)	**(3,5,5,6)**	(7,6,5,6)	(9,6,5,7)	(2,5,5,6)	(2,5,5,7)	(3,5,5,6)	**(1,5,4,6)**	(9,6,1,7)

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)

TH	Unpruned					Pruned				
	Gain	G. Ratio	Gini	χ^2	Random	Gain	G. Ratio	Gini	χ^2	Random
60%	**(3,5,2,4)**	**(2,5,2,5)**	**(3,5,2,4)**	(4,5,2,4)	(4,5,4,7)	(2,4,2,4)	(1,4,2,4)	(2,4,2,4)	(1,3,2,5)	**(3,2,1,1)**
75%	**(3,5,4,5)**	(6,5,4,5)	**(3,5,4,5)**	(6,6,2,6)	(6,6,5,7)	(2,5,4,5)	(1,5,3,5)	(2,5,4,5)	(1,5,2,6)	**(4,4,1,1)**
85%	**(3,5,5,5)**	(8,5,5,6)	**(3,5,5,5)**	(7,6,2,6)	(6,6,5,7)	(2,5,4,5)	(1,5,4,6)	(2,5,5,5)	(1,5,2,6)	**(4,5,1,1)**
90%	**(3,6,5,6)**	(8,6,5,6)	**(3,6,5,6)**	(7,6,4,6)	(7,7,5,7)	(3,6,5,6)	(2,5,4,6)	(3,5,5,6)	(1,5,2,7)	**(4,5,1,1)**
95%	**(3,6,5,6)**	(8,6,5,7)	**(3,6,5,6)**	(8,7,5,7)	(11,7,5,7)	(3,6,5,6)	(2,6,5,7)	(3,6,5,6)	(1,6,5,7)	**(5,6,3,1)**

Figure 2. Performance of Selection Measures Against Classfication Thresholds

Considerations of these tables lead to the following conclusions:

Unpruned Trees

- Information based selection measures outperform the non-information based ones across all sampling policies. This confirms the significance of information contents over statistical values for the tree construction.
- Among the information based measures, information gain achieves the best performance in 4 out of 5 sampling policies, followed closely by Gini index of diversity. It may seem somewhat surprising to conclude that the Gain Ratio selection measure is outperformed by the Information Gain, when in the literature the former has been presented as an improvement on the latter. There are a number of plausible explanations for this discrepancy.
  - At any stage of the tree construction, the Information gain measure is biased towards the attribute that has the highest number of values in the remaining example subset [5]. This means that the effect of attribute choice at any stage on subsequent stages depends very much on the relative attribute-values distributions. On the other hand, when the Gain ratio measure is used, the effect of attribute choice at any stage on subsequent stages is much more complex. It is possible that the relative attribute-values distributions for the data sets, in our experiments, resulted in greater number of examples being classified at higher levels of the decision tree. Hence, while our conclusion goes against the expectation when the traditional concept of structure is used it may not be surprising when our new concept of structure is used to compare performances. This conclusion in fact strengthens the argument for the use of our new concept of structure.
  - A close look at our results shows that except for the Nursery data set, the performance of the Information Gain is almost identical to that of the Gain Ratio measure. The Nursery data set is the largest among the chosen list.

  These explanations and remarks do suggest that the nature and the size of the data set may effect the outcomes. There is a need for more research in this direction.
- Among the non-information based selection measures, χ^2 outperforms the random selection. In fact, random selection has a rather poor performance across all sampling policies, not only resulting in large trees with many levels, but also failing to classify the majority of examples close to the root of the tree.

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

- Classification performance does not only depend on selection measures, but also on sampling policy. The following is a list of best sampling policies for each selection measure:
  - Information gain - 10%-10%-80%, 15%-35%-50%, and 25%-25%-50%
  - Gain ratio - 10%-10%-80%, and 15%-35%-50%
  - Gini index - 15%-35%-50%
  - χ^2 - 10%-10%-80%
  - Random selection - 15%-35%-50%

Pruned Trees

- Non-information based selection measures outperform the information based ones across all sampling policies. This is probably a reflection of the fact that pruning procedure uses statistical calculations and does not depend on any information theory function.
- Except for the 35%-15%-50% policy, random selection appears to have the best classification performance. However, this seemingly good performance is offset by poor accuracy rates ([15]). In fact, the result merely indicates that random selection leads to the construction of a large, complex and erroneous tree, most parts of which are pruned away. In other words, much effort of tree construction is wasted.
- The best sampling policy for χ^2 is 35%-15%-50%.
- Excluding random selection, the information based selection measures do have reasonably good performances with information gain performing well especially with its preferred sampling policies as indicated above in the unpruned case.

5. Conclusion and Future Work

In this paper, we have conducted a comparative study on the effects of different attribute selection measures on decision tree structures by using a number of data sets sampled with various sampling policies. We have used an effective analysis method based on classification thresholds to cope with the large volume of experimental data and clarify the results. We introduced a more informative concept of decision tree structures by combining classification proportions with the combinatorial structures of decision trees. It incorporates the functionality of decision trees into the combinatorial structures.

Our findings indicate that the information-based measures outperform the non-information based ones for unpruned trees on classification proportion thresholds, with the information gain having the best performance. Pruning improves the performance of statistics-based measures. Existing pruning methods use statistical calculations rather than any information content-based functions. Our findings for unpruned trees suggest that designing information based pruning methods is an interesting task, which may lead to desirable consequences for decision tree. It is likely that such a method would be a pre-pruning method. We also show that classification performance is not only related to attribute selection measures but also to the sampling policy. In fact, for each measure there exists a

sampling policy that bring up its best potential.

More studies are needed to unfold the combined effects of sampling policies, attribute selection measures and pruning. The variation in the numbers of attributes of the selected data sets in our studies are not significant enough for us to make creditable conclusions to relate the classification performance to the number of attributes. This is another subject of future work.

References

[1] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. From data mining to knowledge discovery: an overview (Chapter 1), *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy, AAAI/MIT Press: Cambridge, pp. 1-34, 1996.

[2] Mingers, J. An empirical comparison of pruning methods for decision Tree induction, *Machine Learning*, **4**, pp. 227-243, 1989.

[3] Oats, T. & Jensen, D. The effects of training set size on decision tree complexity, *Proc of the 14th Int. Conf. on Machine Learning*, ??, ??. 1997

[4] Mingers, J. An empirical comparison of selection measures for decision tree induction, *Machine Learning*, **3**, pp. 319-342, 1989.

[5] Buntine, W. & Niblett, T. A further comparison of splitting rules for decision tree induction, *Machine Learning*, **8**, pp. 75-85, 1992.

[6] Liu, W.Z & White, A.P. The importance of attribute selection measures in decision tree induction, *Machine Learning*, **15**, pp. 25-41, 1994.

[7] Utgoff, P. E., Berkman, N.C. & Clouse, J. A. Decision tree induction based on efficient tree restructuring, *Machine Learning*, **29**, pp. 5-44, 1997

[8] Breiman, L., Friedman, J., Olshen, R. & Stone, C. *Classification and Regression Trees*, Wadsworth and Brooks, 1984.

[9] Hartigan, J. A. *Clustering Algorithms*, John Wiley & Sons: New York, 1975

[10] Quinlan, J. R. Induction of decision trees, *Machine Learning*, **1**, pp. 81-106, 1986.

[11] Quinlan, J. R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann: Santa Mateo, 1993.

[12] Fayyad, U. & Irani, K. The attribute selection problem in decision tree generation, *Proc. of the 10th National Conf. on Artificial Intelligence*, AAAI Press: San Jose, pp. 104-110, 1992.

[13] Quinlan, J. R. Simplifying decision trees, *International Journal of Man-Machine Studies*, **27**, pp. 221-234, 1987.

[14] *UCI Repository for Machine Learning Databases*, http://www.ics.uci.edu/~mlearn/MLRepositiry.html

[15] Obatusin, M. F. *Measures of Probabilistic Rules for Data Mining*, MSc Thesis, Dept. of Computer Science, Univ. of Buckingham, submitted in November 1999.