# Learning the attribute selection measures for decision tree

Xiaolin Chen, Jia Wu, Zhihua Cai

Xiaolin Chen, Jia Wu, Zhihua Cai, "Learning the attribute selection measures for decision tree," Proc. SPIE 8784, Fifth International Conference on Machine Vision (ICMV 2012): Algorithms, Pattern Recognition, and Basic Technologies, 87842S (13 March 2013); doi: 10.1117/12.2021251

Event: Fifth International Conference on Machine Vision (ICMV 12), 2012, Wuhan, China

# Learning the Attribute Selection Measures for Decision Tree

Xiaolin Chen[1], Jia Wu[1], Zhihua Cai[1]
[1] Faculty of Computer Science, China University of Geosciences, Wuhan 430074, China

## ABSTRACT

Decision tree has most widely used for classification. However the main influence of decision tree classification performance is attribute selection problem. The paper considers a number of different attribute selection measures and experimentally examines their behavior in classification. The results show that the choice of measure doesn't affect the classification accuracy, but the size of the tree is influenced significantly. The main effect of the new attribute selection measures which base on normal gain and distance is that they generate smaller trees than traditional attribute selection measures.

**Keywords:** Decision tree; Classification; Attribute selection measures; Classification accuracy; Tree size

## 1. INTRODUCTION

Because of its speed and high precision, decision tree has been widely used in the classification. Decision tree learning is instance-based inductive learning algorithm, which forms of decision tree classification rules from a group of no order, no rules inference, and usually used to classify the unknown data .The main influence of decision tree classification performance is how to select each node attributes to be tested, namely attribute selection metric problem.

The measures of the traditional attribute selection most widely used are based on the standards of entropy theory, such as information gain (Information Gain), information gain ratio (Gain Ratio) [1]. For such standards' shortcomings we discuss two new attribute selection measures, the normal gain attribute selection measure which also based on entropy theory and the distance-based attribute selection measure which based on the distance metric selection criteria in this paper. We establish the decision tree with the attribute selection measures introduced above, and compare them from the decision tree size, classification accuracy two aspects.

In this paper we designed several experiments of decision tree with different attribute selection measures. In each experiment, we evaluate the performance of decision tree form two parts which are decision tree size, classification accuracy. The rest of paper is organized as follows. In Section 2, we present several attribute selection measures of decision tree and their improvements. In Section 3, we describe the data and experimental procedure. Finally, we draw a conclusion of this paper.

## 2. SEVERAL ATTRIBUTE SELECTION MEASURES AND THEIR IMPROVEMENTS

The specific method of decision tree classification is: it begins with a set of examples; each example is described in terms of a set of attributes. The attribute selection is to choose the attribute that best divides the examples into their classes and then partition the data according to the values of that attribute. This process is recursively applied to each node subset, and the procedure terminate when all examples in the current subset have the same class [2]. The result of the process is represented as a tree in which each node specifies an attribute and each branch emanating from a node specifies the possible values of that attribute. Terminal nodes (leaves) of the tree correspond to the sets of examples with the same class or no more attributes are available. Therefore, a fundamental step in decision tree algorithm is the selection of the attribute at each node. We select the best attribute by seeing how well each one separates the data into the various classes.

### 2.1. TRADITIONAL SELECTION MEASURES

Information gain and gain ratio are two typical attribute selection measures which are based on information entropy. ID3 algorithm uses information gain as the attribute selection measure, the attribute which has the largest information gain is selected as a split attribute [3]. This method of classification has the minimum of expected tests for given examples and make sure to find a simple tree.

Fifth International Conference on Machine Vision (ICMV 2012): Algorithms, Pattern Recognition, and
Basic Technologies, edited by Yulin Wang, Liansheng Tan, Jianhong Zhou, Proc. of SPIE Vol. 8784,
87842S · © 2013 SPIE · CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2021251

For dataset D, information Gain of attribute A defines as

$$Gain(D,A) = Info(D) - Info(D,A) \tag{1}$$

Where

$$Info(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{2}$$

- $p_i$ is the probability of occurrence of each class $C_i(i=1,2\ ,......n)$ in the set $D$ of examples.

$$Info(D,A) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \tag{3}$$

- $v$ is the number of possible values of attributes A;

- $|D_j|$ is the number of examples in $D$ having value $v_j$ for the attribute A;

- $|D|$ is the number of examples in the node dataset $D$ ;

However, as has already been pointed out in the literature (Hart, 1984; Kononenko et al., 1984; Quinlan, 1986), this attribute selection measure is biased in favor of attribute with a large number of values. In order to compensate for this bias, Quinlan (1986) introduced a modification of the Gain measures. The modification is Gain Ratio which use split information to standardize the information gain. The split information defines as:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{4}$$

The $SplitInfo_A(D)$ value represents the information that the training data set D divided into v partitions by attribute A. Next the Gain Ration can be described as:

$$GainRatio(D,A) = \frac{Gain(D,A)}{SplitInfo_A(D)} \tag{5}$$

## 2.2. THE DISADVANTAGE OF TRADITIONAL SELECTION MEASURES

The information gain and gain ratio regard as two typical attribute selection measures have been used widely. But a notable disadvantage of the information gain is that it is biased towards selecting attributes with many values. This motivated Quinlan to define the Gain Ratio which mitigates this bias but suffers from other disadvantages [4]. The disadvantage of the gain ratio is when one of the $D_j$ is close to dataset $D$ , the denominator may got the value of zero or extraordinary small. That is for the attribute which has the same value for nearly all the instance of dataset $D$ , it will cause the Gain Ratio have not defined or very large. Then we will introduce two new attribute selection measures for these disadvantages.

## 2.3. IMPROVED SELECTION MEASURES

### 2.3.1. NORMAL GAIN

Normal gain (NG) is based on information gain [5]. It is about to replace the base 2 logarithm of information gain by base n logarithm (n is the number of division). Normal gain constraint the gain value when the value of n closes to infinite large, and effectively overcome the bias of smaller partitions for information gain.

$$NormalRatio(D,A) = \frac{Gain(D,A)}{\log_2 n} \tag{6}$$

Noted that in the formula of NG, when the division number of n increases, NG decreases; when the purity of each partition increases (the probability of instance belonging to one class), NG increases. That two opposite effects to NG causes it may not properly distinguish the best division of some special examples. But the experiment has proved that the result of NG is better than GR and IG in majority of attribute selection experiments.

### 2.3.2.DISTANCE-BASED ATTRIBUTE SELECTION MEASURES

R. López De Mántaras introduced a distance between partitions as attribute selection measures and proved that it is not biased towards many-valued attributes [6]. In this selection measure, the chosen attribute in a node will be that whose corresponding partitions is the closet (in terms of distance) to the correct partition of the subset of examples in this node.

In case of, we design two partitions PA, PB on the same set D. Attribute A has n different values { $A_1, A_2, .....A_n$ } and attribute B has m different values { $B_1, B_2, .....B_m$ }. Then the distance-based attribute selection measure can be described as:

$$d_N(P_A, P_B) = \frac{Info(P_B / P_A) + Info(P_A / P_B)}{Info(P_A \cap P_B)} \tag{7}$$

Where

$$Info(P_B / P_A) = -\sum_{i=1}^{n} P_i \sum_{j=1}^{m} P_{j/i} \log_2 P_{j/i} \tag{8}$$

$$Info(P_A \cap P_B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} \log_2 P_{ij} \tag{9}$$

And we define

$$\begin{cases} P_i = P\left(A_i\right)、\ P_i = P\left(B_i\right) \\ P_{\mathrm{ij}} = P\left(A_i \cap B_j\right) \\ P_{j/i} = P(B_j / A_i) \qquad 1 \le i \le n, 1 \le j \le m \end{cases} \tag{10}$$

In order to compare the new distance-based selection measure with the Quinlan's gain, then present the relation with Quinlan's information gain. It has been proved that the distance of the two partitions also can be describes as:

$$1 - d_{\mathrm{N}}(P_C, P_V) = \frac{Gain(A_k, X)}{Info(P_V \cap P_C)} \tag{11}$$

Furthermore, the gain ratio also can be expressed in terms of information measures on partitions as follows:

$$G_R(A_k, X) = \frac{Gain(A_k, X)}{I(P_v)} \tag{12}$$

From the formula above, we can known that the mainly difference between the distance measure and gain ratio is the denominator. The distance measure adopt the $Info(P_V \cap P_C)$ instead of $I(P_v)$, which can associate with the partition generated by attribute $A_K$ .Because the $Info(P_V \cap P_C)$ cannot be zero when the partition $D_j$ is close to dataset $D$ .Therefore, distance can solve the problem of gain ratio always may not be defined. Furthermore, because the distance always has $Info(P_V \cap P_C) > Gain(A_k, X)$ , it will not have the disadvantage of gain ratio that choosing the attribute with very low $I(P_v)$ rather than with high Gain. The distance-based attribute selection measure also solves the problem of choosing attributes with large number of values that has been proved.

## 3. EXPERIMENTAL METHOD AND RESULTS

The previous section described a number of different measures of attribute selection. The main purpose of the current research was to conduct a detailed comparison between these alternatives to determine their effect on the decision tree size and classification accuracy. This section describes the experiment of data used and the methodology, results of the experiment.

### 3.1. EXPERIMENTAL DATA

We run our experiments on 36 standard UCI data sets[7] on Weka[8]. The data represent a wide range of domains and data characteristics listed in Table 1. Before classification, we should handle the data with following three steps:

1. Dealing with the missing attribute values. For the special datasets, we use the unsupervised attribute filter Replace Missing Values in Weka to replace all the missing attribute value.

2. Discretize the numeric attribute values. For the special datasets, we used the unsupervised filter Discretize in Weka to handle all numeric attribute values in each data set.

3. Removing some of the useless attributes [9]. There are three such attributes in the above-mentioned 36 data sets: "Hospital Number" attribute in data set "colic.ORIG", "instance name" attribute in the data set "splice", and the "animal" attribute in data sets "zoo". In order to remove these useless attributes we adopt the unsupervised filter named Remove in Weka.

Table 1. Description of data sets used in the experiments

| No. | Dataset | Instances | Attributes | Classes | Missing | Nume |
|---|---|---|---|---|---|---|
| 1 | anneal | 898 | 39 | 6 | Y | Y |
| 2 | anneal.ORIG | 898 | 39 | 6 | Y | Y |
| 3 | audiology | 226 | 70 | 24 | Y | N |
| 4 | autos | 205 | 26 | 7 | Y | Y |
| 5 | balance-scale | 625 | 5 | 3 | N | Y |
| 6 | breast-cancer | 286 | 10 | 2 | Y | N |
| 7 | breast-w | 699 | 10 | 2 | Y | N |
| 8 | colic | 368 | 23 | 2 | Y | Y |
| 9 | colic.ORIG | 368 | 28 | 2 | Y | Y |
| 10 | credit-a | 690 | 16 | 2 | Y | Y |
| 11 | credit-g | 1000 | 21 | 2 | N | Y |
| 12 | diabetes | 768 | 9 | 2 | N | Y |
| 13 | Glass | 214 | 10 | 7 | N | Y |
| 14 | heart-c | 303 | 14 | 5 | Y | Y |
| 15 | heart-h | 294 | 14 | 5 | Y | Y |
| 16 | heart-statlog | 270 | 14 | 2 | N | Y |
| 17 | hepatitis | 155 | 20 | 2 | Y | Y |
| 18 | hypothyroid | 3772 | 30 | 4 | Y | Y |
| 19 | ionosphere | 351 | 35 | 2 | N | Y |
| 20 | iris | 150 | 5 | 3 | N | Y |

| | | | | | | |
|---|---|---|---|---|---|---|
| 21 | kr-vs-kp | 3196 | 37 | 2 | N | N |
| 22 | labor | 57 | 17 | 2 | Y | Y |
| 23 | letter | 20000 | 17 | 26 | N | Y |
| 24 | lymph | 148 | 19 | 4 | N | Y |
| 25 | mushroom | 8124 | 23 | 2 | Y | N |
| 26 | primary-tumor | 339 | 18 | 21 | Y | N |
| 27 | segment | 2310 | 20 | 7 | N | Y |
| 28 | sick | 3772 | 30 | 2 | Y | Y |
| 29 | sonar | 208 | 61 | 2 | N | Y |
| 30 | soybean | 683 | 36 | 19 | Y | N |
| 31 | splice | 3190 | 62 | 3 | N | N |
| 32 | vehicle | 846 | 19 | 4 | N | Y |
| 33 | vote | 435 | 17 | 2 | Y | N |
| 34 | vowel | 990 | 14 | 11 | N | Y |
| 35 | waveform-5000 | 5000 | 41 | 3 | N | Y |
| 36 | zoo | 101 | 18 | 7 | N | Y |

### 3.2. EXPERIMENTAL METHODS

In all, four different measures were tested on all of the data sets. The particular factors of interest, as mentioned above, were decision tree size and classification accuracy.

In order to obtain independent test data and reliable results, each original data set was split randomly (90/10) into a training and a test data set. The trees were grown and pruned on the training data set and then accuracy was measured on the test data set [10].

To guard against random splits that happened to be untypical, the whole procedure was carried out ten times, giving ten independent pairs of training and test data for each data set. All the methods were run on the same datasets and the results averaged across the ten pairs.

### 3.3. EXPERIMENTAL RESULTS

Classification accuracy is the very important criteria for evaluation a performance of a classifier [11]. Thus, we use the natural dependent measures for experiments on decision tree induction.

Classification accuracy, it equal to the percentage of instances correctly classified, refers to the predictive ability of a decision tree in terms of classifying an independent set of test data.

The comparison results on accuracy values of all four algorithms on each data set are shown in Table 2. The symbols v and * in the table respectively denote statistically significant upgrade or degradation over information gain with a 95% confidence level. The $w/t/l$ values are summarized at the bottom of the table. Each entry $w/t/l$ in the table means that the algorithms win on $w$ data sets, tie on $t$ data sets, and lose on $l$ data sets, compared to information gain measure.

Then we can know from the Table 2 that the classification accuracy of the improved selection measures is almost the same with traditional measures. The classification accuracy of normal gain measure does not have obvious difference with information measure on 33 data sets. And the distance measure does not have obvious difference with information measure on 30 data sets. But in some data sets the improved got high classification accuracy than traditional measures, although the differences are not statistically significant.

Table 2. Experimental results on classification accuracy and standard deviation

| Database | Info Gain | Ratio | Normal Gain | Distance |
|---|---|---|---|---|
| anneal | 99.62±0.66 | 99.48± 0.68 | 99.44± 0.72 | 99.51± 0.66 |
| anneal.ORIG | 89.63±2.93 | 90.77± 2.61 | 89.20± 2.96 | 90.46± 2.72 |
| audiology | 78.05±7.37 | 83.28± 7.59 v | 78.85± 9.29 | 81.32± 7.92 |
| autos | 78.75±8.57 | 79.34± 8.43 | 77.06± 8.26 | 79.61± 8.52 |
| balance-scale | 37.74±4.92 | 37.73± 4.89 | 37.79± 4.96 | 37.73± 4.89 |
| breast-cancer | 58.95±9.22 | 58.56± 8.64 | 62.33± 8.92 | 60.19± 8.78 |
| breast-w | 90.12±3.11 | 90.00± 3.28 | 90.12± 3.11 | 90.10± 3.38 |
| colic | 72.25±6.73 | 71.33± 7.59 | 74.85± 7.04 | 72.10± 8.05 |
| colic.ORIG | 53.02±7.91 | 65.95± 7.46 v | 67.05± 8.01 v | 62.66± 7.62 v |
| credit-a | 73.84±5.32 | 73.84± 4.41 | 75.55± 5.09 | 73.94± 4.18 |
| credit-g | 62.49±4.31 | 59.82± 3.92 | 62.47± 3.72 | 60.43± 4.36 |
| diabetes | 60.31±4.85 | 59.28± 4.71 | 60.31± 4.85 | 59.01± 4.70 |
| glass | 51.28±9.04 | 51.27± 9.04 | 51.28± 9.04 | 50.39± 8.42 |
| heart-c | 62.42±9.04 | 68.45± 8.50 | 67.69± 8.93 | 67.82± 8.62 |
| heart-h | 66.17±9.25 | 67.54± 7.24 | 68.29± 8.31 | 67.16± 7.33 |
| heart-statlog | 62.07±9.12 | 68.37± 9.61 | 62.07± 9.12 | 68.93± 8.54 |
| hepatitis | 71.01±9.72 | 74.80±10.70 | 74.13±10.52 | 74.50±10.04 |
| hypothyroid | 90.25±1.1 | 90.02± 1.06 | 90.43± 1.07 | 90.09± 1.15 |
| ionosphere | 84.67±5.46 | 79.89± 6.12 * | 84.67± 5.46 | 82.16± 6.29 |
| iris | 90.80±7.26 | 89.33± 7.03 | 90.80± 7.26 | 89.13± 7.38 |
| kr-vs-kp | 99.60±0.38 | 99.63± 0.32 | 99.60± 0.40 | 99.63± 0.34 |
| labor | 73.40±16.26 | 82.07±15.41 | 83.03±15.09 | 80.63±14.94 |
| letter | 66.42±2.19 | 70.03± 2.27 v | 66.42± 2.18 | 68.18± 2.19 v |
| lymph | 73.40±10.98 | 71.36±11.12 | 73.27± 8.96 | 69.07±10.23 |
| mushroom | 99.75±0.35 | 99.74± 0.37 | 99.86± 0.31 | 99.72± 0.39 |
| primary-tumor | 34.31±7.68 | 35.40± 6.80 | 33.81± 7.42 | 35.10± 6.96 |
| segment | 92.11±1.77 | 92.58± 1.61 | 92.10± 1.75 | 92.73± 1.55 |
| sick | 97.57±0.85 | 97.50± 0.91 | 97.56± 0.84 | 97.52± 0.90 |
| sonar | 62.90±11.09 | 59.93± 9.99 | 62.90±11.09 | 60.55±10.25 |
| soybean | 88.65±3.42 | 92.60± 3.01 v | 92.36± 2.90 v | 92.22± 3.43 v |
| splice | 89.75±1.65 | 89.22± 1.70 | 90.16± 1.85 | 88.60± 1.70 * |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| vehicle | 61.28±4.25 | 62.80± 4.73 | | 61.28± 4.25 | | 63.92± 4.82 | |
| vote | 93.15±3.32 | 92.96± 3.26 | | 93.15± 3.32 | | 93.54± 3.20 | |
| vowel | 79.81±3.93 | 76.30± 4.33 | * | 76.55± 4.44 | * | 77.52± 4.41 | * |
| waveform-5000 | 64.03±4.86 | 58.90± 4.68 | * | 64.03± 4.86 | | 60.20± 4.96 | * |
| zoo | 97.12±4.96 | 97.65± 4.63 | | 95.45± 6.03 | | 97.12± 4.96 | |
| Average | 75.19 | 76.05 | | 76.28 | | 75.93 | |
| $w/t/l$ | (v/ /*) | (4/29/3) | | (2/33/1) | | (3/30/3) | |

Also in Table 3 we compare the improved measures with each other. From the table we can conclude that they don't have significant difference on classification accuracy.

Table 3. Summary of the experimental results

| Test base | Ratio | NormalGain | Distance |
|---|---|---|---|
| Info | (4/29/3) | (2/33/1) | (3/30/3) |
| Ratio | \ | (2/32/2) | (0/35/1) |
| NormalGain | \ | \ | (2/32/2) |

The above results support that the decision tree classification accuracy is not sensitive to the goodness of the attribute selection measures [12]. The main effect of the new attribute selection measure is to reduce the size of the tree, rather than alter its accuracy. In our experiments we use the number of leaves to measure the size of tree. Although the classification accuracy of normal gain doesn't significant improved, it may generate smaller trees than gain ratio. The number of leaves for decision trees generated by different measures is show in the Fig. 1.

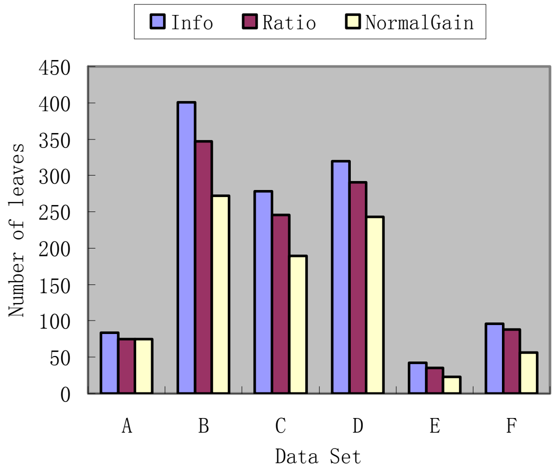

Fig. 1. Number of leaves for different measures. The meaning of letters in the figure is as flows:A is the data set of audiolpgy,B is the data set of breast-cancer,C is the data set of colic,D is the data set of heart-c,E is the data set of labor and F is the data set of lymph

From the above theoretical proof in section 2 we can known that the distance measure will generate smaller tree than gain ratio especially in the domain of the data sets whose attributes have a large number of values. Then we choose the data set 'hepatitis' which comply with the above characteristics for experiment in order to support the theoretical proof. In weka, we build a decision tree with the distance measure which has 99 leaves and it is smaller than the tree of information gain (102 leaves) and gain ratio (107 leaves).

# 4. CONCLUSIOINS

In this paper, first we introduce two traditional attribute selection measures of decision tree. For the traditional measures disadvantages, we present two improved measures. In fact, the results show that classification accuracy is not improved significantly by using different attribute selection measures, but the size of the tree may be influenced significantly. The new attribute selection measures which base on normal gain and distance generate smaller trees than traditional attribute selection measures.

In summary, there exists a number of ways to improve the performance of decision tree classification and future work should explore these methods.

# 5. ACKNOWLEDGEMENTS

This work was supported by the National High Technology Research and Development Program (863 Program) of China (No.2009AA12Z117).

# REFERENCES

[1] J. W. Han, M. Kamber. Data Mining[M].China Machine Press, 2001, pp.279-334..

[2] J. R. Quinlan, C4.5: Programs for Machine Learning [M].San Mateo, California: Morgan Kaufmann, 1993.

[3] J. R. Quinlan, Induction of Decision Trees, Machine Learning, v.1 n.1, pp.81-106.

[4] P. W. Allan, W. Z. Liu, Technical Note: Bias in Information-Based Measures in Decision Tree Induction, Machine Learning, v.15 n.3, pp.321-329, June 1994.

[5] S. J. Hong, Use of Contextual Information for Feature Ranking and Discretization[J], IEEE Transactions on Knowledge and Data Engineering 1997, 9(5):718-730.

[6] R. López De Mántaras, A Distance-Based Attribute Selection Measure for Decision Tree Induction, Machine Learning, v.6 n.1, pp.81-92, Jan. 1991.

[7] C. Merz, P. Murphy, D.Aha, UCI repository of machine learning databases. In: Department of ICS, University of California, Irvine, 1997. http://www.ics.uci.edu/mlearn/MLRepository.html.

[8] H. W. Ian, F. Eibe, Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco. http://prdownloads.sourceforge.net/weka/datasets-UCI.jar

[9] L. Jiang, C. Li, Z. Cai, Learning decision tree for ranking, Knowledge and Information Systems, 2009, 20: 123-135.

[10] M. John, An Empirical Comparison of Selection Measures for Decision-Tree Induction, Machine Learning, v.3 n.4, pp.319-342, March 1989.

[11] F. A. Thabtah, P. Cowling, Y. Peng, Multiple labels associative classification. Knowledge and Information Systems 9(1):109–129

[12] I. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and regressing trees. Belmont, CA: Wadsworth International Group, 1984.