Construction of Decision Trees based Entropy and Rough Sets under Tolerance Relation

Ning Yang[1] Tianrui Li[2] Jing Song[3]

[1]Department of Mathematics, Southwest Jiaotong University, Chengdu 610031, China
[2]School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China
[3]Research Center for Secure Application in Networks and Communications,
Southwest Jiaotong University, Chengdu 610031, China

Abstract

Decision tree induction is one of the most popular data mining techniques with applications in various fields. Present algorithms for construction decision trees can not deal with missing value in information systems properly. A new concept, rough gain ratio, is first introduced by the aid of tolerance relations in the extended rough sets theory. Then, an approach for inducing decision trees under the rough gain ratio is presented. Examples show that the decision trees generated by the proposed method tend to have simpler structure and more understandable rules than C4.5.

Keywords: Data mining, Decision tree, Rough set, Tolerance relation

1. Introduction

Data mining is the non-trivial process of identifying valid novel potentially useful and ultimately understandable patterns in data. It is currently a fast growing field both from an application and from a research point of view. The reason is that companies see a high chance for deriving valuable information from huge amount of available data that can then be used for improving their business. Decision trees are considered to be one of the most popular data-mining techniques for knowledge discovery. It systematically analyzes the information contained in a large amount of data source to extract valuable rules and relationships and usually is used for the purpose of classifying or prediction. Compared to other data-mining techniques, it is widely applied in various areas since it is robust to data scales or distributions [1]-[3].

A decision tree is a tree structure representation of the given decision problem such that each non-leaf node is associated with one of the decision variables, each branch from a non-leaf node is associated with a subset of the values of the corresponding decision variable, and each leaf node is associated with a value of the target (or dependent) variable [4]. A decision tree is constructed from a training set, which consists of objects. Each object is completely described by a set of attributes and a class label. In addition, to construct a decision tree, it is necessary to find at each internal node a test for splitting the data into subsets, namely, we have to select appropriate attributes as the tree nodes. The concrete process for construction of decision tree by starting with an empty tree and the entire training set is as follows [7].

1. If all the training examples at the current node *t* belong to category *c*, create a leaf node with the class *c*.

2. Otherwise, score each one of the set of possible splits *S*, using a goodness measure.

3. Choose the best split *s** as the test at the current node.

4. Create as many child nodes as there are distinct outcomes of *s**. Label edges between the parent and child nodes with outcomes of *s**, and partition the training data using *s** into the child nodes.

5. A child node *t* is said to be pure if all the training samples at *t* belong to the same class. Repeat the previous steps on all impure child nodes.

Here goodness measures are also known as feature evaluation criteria, feature selection criteria, impurity measures or splitting rules. Presently many goodness measures are available for attribute selection, such as the entropy based measures [2]-[3], gini index measures [7] and Distance measures [5]. Then various decision tree algorithms have been developed for classification. For example, ID3 algorithm for classification uses information gain, an entropy based measure, to select the best splitting attribute. The attribute with the highest information gain is selected as the splitting attribute. One of the main drawbacks of ID3 is that the measure Gain used tends to favor attributes with a large number of distinct values. This drawback was overcome to some extent in C4.5 by introducing a new entropy based measure called Gain Ratio [2].

In real world data sets, it is often the cases that some attribute values are missing from the data [6, 8, 9]. Several researchers have addressed the problem of dealing with missing attribute values in the training as well as testing sets. Friedman suggested that all objects with missing attribute values can be ignored while forming the split at each node. If it is feared that too much discrimination information will be lost due to ignoring, missing values may be substituted by the mean

value of the particular feature in the training subsample in question [6]. On the other hand, Quinlan argues that in case of missing values the splitting criteria should be reduced proportionally as nothing has been learned from these instances [2]. Once a split is formed, all objects with missing values can be passed down to all child nodes, both in the training and testing stages. The classification of an object with missing attribute values will be the largest represented class in the union of all the leaf nodes at which the object ends up. However, the existing approaches for handling missing values do not perform well in real situations.

Rough set theory is a mathematical approach to vagueness [10]. The main advantage of rough set theory in data analysis is that it does not need any preliminary or additional information about data like probability distributions in statistics, basic probability assignments in Dempster–Shafer theory, a grade of membership or the value of possibility in fuzzy set theory. The rough set philosophy is founded on the assumption that with every object of the universe of discourse we associate some information (data, knowledge). Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory [10]. However, the classical rough set approach is based on complete information systems while in many cases we have to deal with incomplete information systems (IIS) since practical data is always incomplete to some extent. Therefore, tolerance, similarity, limited tolerance, characteristic relations have been proposed respectively for extending rough set theory in IIS and they can cope with missing values effectively in many situations [11]-[12].

Various approaches for construction decision trees under rough sets theory have been proposed in the literatures. The core of condition attributes with respect to decision attributes in rough sets theory is used for selection of attributes in the process of construction multivariate decision trees in [13]. Another approach to selection of attributes for construction of decision tree is presented based on the idea that if the size of the implicit region corresponding to one condition attribute is the smallest, then this attribute will be chosen as the node for branching [14]. The weighted mean roughness, a new concept based on rough sets theory, is presented and regarded as the criteria for choosing attributes in decision trees construction [15]. The variable precision rough set model is also employed for inducing decision trees in [16]. This approach is aimed at handling uncertain information during the process of inducing decision trees and generalizes the rough set based approach to decision tree construction by allowing some extent misclassification when classifying objects. However, the existing approaches can not handling information systems with missing values. In this paper, a new concept, rough gain ratio, is first introduced by the aid of tolerance relations in the extended rough sets theory. Then, an approach for inducing decision trees in discrete variable domains under the rough gain ratio is presented.

2. Preliminaries

In this section, we will introduce some basic concepts of ID3, C4.5 as well as rough sets and their extensions [2, 12].

ID3 algorithm uses information gain to decide the splitting attribute. Given a collection S of c outcomes, Entropy is defined as $Entropy(S) = \sum -p(I)\log_2 p(I)$, where $p(I)$ is the proportion of S belonging to class I.

Definition 1[2] The information gain of example set S on attribute A is defined as $Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v)$, where, S_v =subset of S for which attribute A has value v. The attribute value that maximizes the information gain is chosen as the splitting attribute.

C4.5 is an extension of ID3 algorithm. Information Gain used in ID3 algorithm always tends to select attributes that have a large number of values since the gain of such an attribute would be maximal. To overcome this drawback Quinlan suggested the use of Gain Ratio as a measure to select the splitting attribute instead of Information Gain.

Definition 2[2] The gain ratio of example set S on attribute A is defined as $GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}$, where $SplitInfo(S, A) = I\left(\frac{|S_1|}{|S|}, \frac{|S_2|}{|S|}, \cdots, \frac{|S_m|}{|S|}\right)$, where $S_1, S_2, \cdots, S_m$ are the partitions induced by attribute A in S.

Definition 3[10] An information system is defined as a pair $< U, C \cup D >$, where U is a non-empty finite set of objects, $C \cup D$ is a non-empty finite set of attributes, C denotes the set of condition attributes and D denotes the set of decision attributes, $C \cap D = \varnothing$. Each attribute $a \in C$ is associated with a set V_a of its value, called the domain of a.

Definition 4[12] An information system $< U, C \cup D >$ is called as an incomplete information system (IIS) if there exists an a in C and an x in U that satisfy the value $a(x)$ is unknown, denoted as *.

Table 1 is an IIS. Under this definition of IIS, the toleration relation is proposed to deal with unknown data in [11].

Definition 5[11] Let $B \subseteq C$ be a subset of attributes. The similarity relation is defined as:

$$SIM(B) = \{(x, y) \in U \times U \mid \forall a \in B, a(x) = a(y) \text{ or } a(x) = * \text{or } a(y) = *\}$$

Property 1[11] $SIM(B)$ is a tolerance relation: $SIM(B) = \bigcap_{a \in B} SIM(\{a\})$.

	Outlook	Temperature	Humidity	Windy	Play?
1	*	hot	high	false	No
2	sunny	hot	high	true	No
3	overcast	hot	*	false	Yes
4	*	mild	high	*	Yes
5	rain	cool	normal	false	Yes
6	rain	cool	normal	true	No
7	overcast	*	normal	true	Yes
8	sunny	mild	high	*	No
9	sunny	*	normal	false	Yes
10	rain	mild	normal	false	Yes
11	sunny	mild	*	true	Yes
12	overcast	mild	*	true	Yes
13	overcast	hot	normal	false	Yes
14	rain	mild	high	*	No

Table 1: Incomplete information system.

Let $S_B(x)$ denote the object set $\{y \in U \mid (x, y) \in SIM(B)\}$. $S_B(x)$ is the maximal set of objects which are possibly indiscernible by B with x. Let $D_B(x)$ denote the object set $\{y \in U \mid (x, y) \notin SIM(B)\}$. $D_B(x)$ is the maximal set of objects which are definitely discernible by B with x. Of course, $S_B(x) \bigcap D_B(x) = \varnothing$ and $S_B(x) \bigcup D_B(x) = U$ for any $x \in U$. Let $U / SIM(B)$ denote classification, which is the family set $\{S_B(x) \mid x \in U\}$. Any element from $U / SIM(B)$ will be called a *tolerance class.* Tolerance classes in $U / SIM(B)$ do not constitute a partition of U in general. They may be subsets/supersets of each other or may overlap. Of course, $\bigcup U / SIM(B) = U$ [11].

3. Construction of Decision Trees based entropy and rough sets under tolerance relation

3.1. A measure for inducing decision trees under rough sets based tolerance relation

This section will first present two new concepts, rough information gain and rough gain ratio, as the measures for the latter construction of decision trees.

Let $C=\{c_1, \ldots, c_k\}$, $U / SIM(\{c_i\}) = \{U_1^i, \ldots, U_{m_i}^i\}$, $i=1, 2, \ldots, k$, $U / SIM(D) = \{U_1^D, \ldots, U_l^D\}$. Then $U_j^i / SIM(D) = \{U_{i,j}^1, U_{i,j}^2, \cdots, U_{i,j}^l\}$, $i=1, 2, \ldots, k$, $j=1,\ldots,m_i$.

Let $\text{Rinfo}(U_j^i) = -\sum_{n=1}^{l} \frac{|U_{i,j}^n|}{|U_j^i|} \times \log_2 \left(\frac{|U_{i,j}^n|}{|U_j^i|} \right)$, $i=1, 2, \ldots, k$, $j=1,\ldots,m_i$. $p(c_i) = \sum_{j=1}^{m_i} |U_j^i|$, $i=1, 2, \ldots, k$. Then

$$\text{RInfo}(\{c_i\}) = \sum_{j=1}^{m_i} \left(\frac{|U_j^i|}{p(c_i)} \times \text{Rinfo}(U_j^i) \right), \quad i=1, 2, \ldots, k.$$

$$\text{RInfo}(S_i) = -\sum_{n=1}^{l} \left(\sum_{j=1}^{m_i} |U_{i,j}^n| \times \log_2 \left(\sum_{j=1}^{m_i} |U_{i,j}^n| \right) \right), \quad i=1, 2, \ldots, k.$$

Definition 6 The rough information gain is defined as $\text{RGain}(S_i, \{c_i\}) = \text{Rinfo}(S_i) - \text{Rinfo}(\{c_i\})$, $i=1, 2, \ldots, k$.

Definition 7 The rough gain ratio is defined as $\text{RGainRatio}(S_i, \{c_i\}) = \frac{\text{RGain}(S_i, \{c_i\})}{\text{RIntrinsicInfo}(S_i, \{c_i\})}$, where

$$\text{RIntrinsicInfo}(S_i, \{c_i\}) = -\sum_{j=1}^{m_i} \left(\frac{|U_j^i|}{|U|} \times \log_2 \left(\frac{|U_j^i|}{|U|} \right) \right).$$

Example 1. The RInfo, RGain, RIntrinsicInfo, RGainRatio of the attributes "Outlook, Temperature, Humidity, Windy" in Table 1 are listed in Table 2.

	Outlook	Temperature	Humidity	Windy
RInfo	0.883341	0.846951	0.737019	0.956745
RGain	0.080738	0.005454	0.136962	0.020673
RIntrinsicInfo	1.584963	1.530493	0.997503	0.997503
RGainRatio	0.05094	0.00356	0.1373	0.02072

Table 2: The RInfo, RGain, RIntrinsicInfo, RGainRatio of the attributes in Table 1.

3.2. Algorithm for Inducing Decision Trees based entropy and rough sets under tolerance relation

The idea of construction of decision tree based on the extended model of rough sets is to use its tolerance relation and consider unknown values as any values in the corresponding attribute. Then, using the rough gain ratio as a goodness measure, carry on construction of decision tree, the process is similar to that of C4.5. The concrete algorithm denoted as RC4.5 is shown as follows.

Algorithm 3.1 The RC4.5-algorithm

RC4.5(Instances; Decision attribute; Attributes)

Instances are the training objects. Decision attribute is the attribute whose value is to be predicted by the tree. Attributes is a list of other attributes that may be tested by the learned decision tree.

Returns a decision tree that correctly classifies the given instances.

Create a *Root* node for the tree.

if the Decision attribute's values of all the instances are the same, *Return* the single-node tree *Root*, with label = the unique Decision attribute's value.

end

if Attributes is empty, *Return* the single-node tree *Root*, with label = most common value of Decision attribute in Instances

end

Otherwise

begin

$A \leftarrow$ the attribute from Attributes with the greatest rough gain ratio

The decision attribute for $Root \leftarrow A$

for each possible value, v_i, of A,

Add a new tree branch below *Root*, corresponding to the test $A = v_i$

Let Instances(v_i) be the subset of Instances that have value v_i for A

if Instances(v_i) is empty

then below this new branch add a leaf node with label = most common value of Decision attribute in Instances

else below this new branch add the subtree

RC4.5(Instances(v_i); Decision attribute; Attributes-{A})

end

end

end

RETURN *Root*

Example 2. We employ the above approach for construction of decision tree of Table 1. Fig. 1 is the decision tree of Table 1. The numbers of leaves and size of the decision tree without unknown value in it are 8 and 15, respectively. While the numbers of leaves and size are 14 and 19, respectively in the decision tree constructed by C4.5 using J48 in WEKA [17]. Obviously, the decision trees generated by the proposed method tend to have simpler structure and more understandable rules than C4.5.

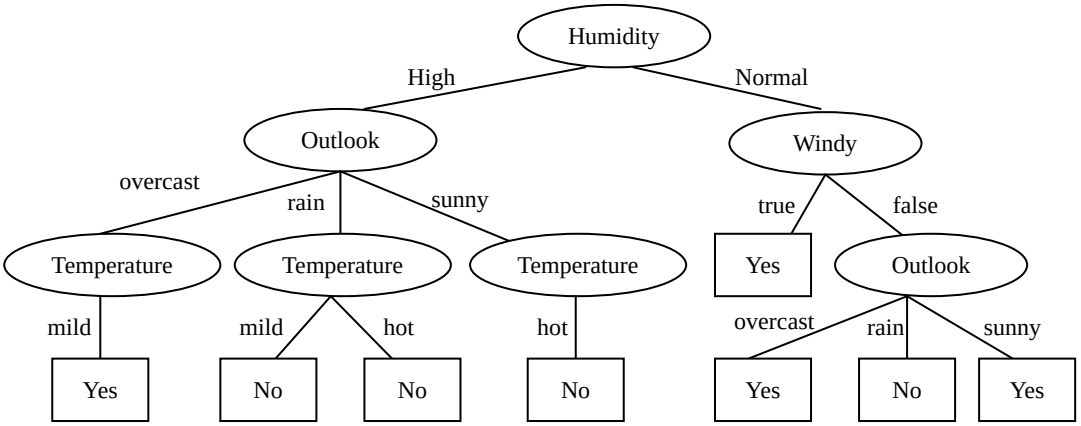

Fig.1: Construction of the decision tree based on rough gain ratio.

We employ a post-pruning process, namely, if all descendants of a node in the decision tree have the same class label, then delete this node and its descendants and create a leaf node with the same class label.

Example 3. Fig. 2 is the decision tree of Fig. 1 after pruning. The numbers of its leaves and size are 7 and 11, respectively.

4. Conclusions

This paper first presents the concept of rough gain ratio. Then an approach for inducing decision tree is proposed in discrete variable domains under the rough gain ratio. Examples show that the decision trees constructed by

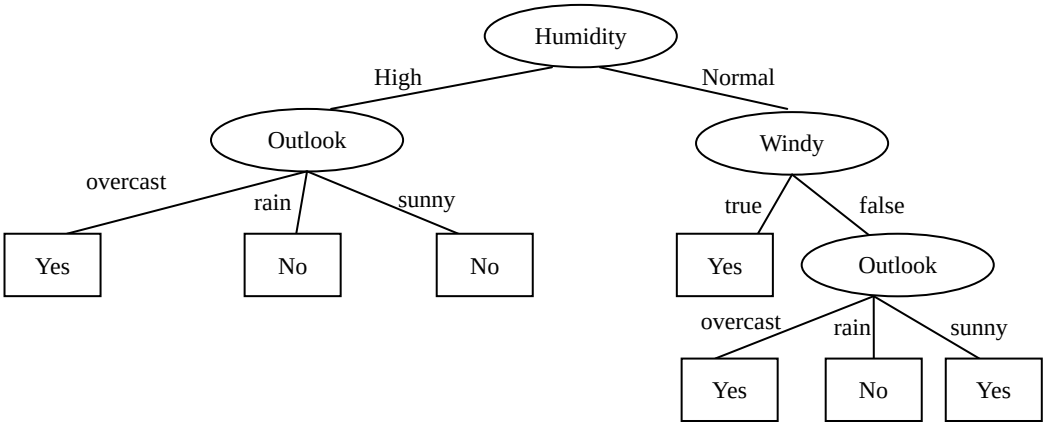

Fig.2: The decision tree after pruning.

this method have a simpler structure and more understandable rules than C4.5. Our future work is to obtain its performance in continuous variable domains and in different databases including different ratios of unknown values in the databases. How to employ the proposed approach to induce the decision tree in a very large database is also an interesting research direction.

Acknowledgement

This work is partially supported by Natural Science Foundation of China (No.60074014) and the Research Fund for the Doctoral Program of Higher Education (No.20060613007).

References

[1] M. J. Berry and G. S. Linoff, *Mastering data mining*. New York: John Wiley & Sons. 2000.

[2] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1993.J. R. Quinlan, Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77-90, 1996.

[3] K-M. Osei-Bryson, Post-pruning in decision tree induction using multiple performance measures, *Computers & Operations Research*, 34: 3331-3345, 2007.

[4] RLd. Mantras, A distance-based attribute selection measure for decision tree induction, *Machine Learning*, 6:81-92, 1991.

[5] E. Acuna and C. Rodriguez, The treatment of missing values and its effect in the classifier accuracy, In Banks D. et al. (eds) *Classification, Clustering and Data Mining Applications*, pages: 639-648, 2004.

[6] L. Rokach and O. Maimon, Top-Down Induction of Decision Trees Classifiers—A Survey. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 35(4):476-487, 2005.

[7] G. Batista and M. C. Monard, An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Applied Artificial Intelligence*, 17(5-6):519-533, 2003.

[8] S. Zhang, Z. Qin, C. X. Ling and S. Sheng, Missing Is Useful: Missing Values in Cost-Sensitive Decision Trees, *IEEE Transactions and Knowledge and Data Engineering*, 17(2):1689-1693, 2005.

[9] Z. Pawlak, *Rough sets: Theoretical aspects of reasoning about data*. Dordrecht, Kluwer, 1991

[10] M. Kryszkiewicz, Rough set approach to incomplete information system, *Information Sciences*, 11(2): 39-49, 1998.

[11] T. Li, D. Ruan and W. Geert, A rough sets based characteristic relation approach for dynamic attribute generalization in data Mining, *Knowledge-Based Systems*, 20(5): 485-494, 2007.

[12] D. Miao and J. Wang, Rough sets based approach for multivariate decision tree construction, *Journal of Software* (Chinese), 8(6):425-431, 1997.

[13] J. Wei, Rough set based approach to selection of node, *International Journal of Computational Cognition*, 1(2):25-40, 2003.

[14] Y. Jiang, Z. Li and Q. Zhang, New method for constructing decision tree based on rough sets theory, *Computer Applications* (Chinese), 24(8):21-23, 2004.

[15] J. Wei, S. Wang and M. Wang, Rough set based approach for inducing decision trees, *Knowledge Based Systems* (2007), doi:10.1016/j.knosys.2006.10.001.

[16] http://www.cs.waikato.ac.nz/ml/weka.