Decision Tree Construction based on Rough Set Theory under Characteristic Relation

Jing Song[1] Tianrui Li[2] Ying Wang[1,3] Jianhuai Qi[1]

[1] Research Center for Secure Application in Networks and Communications,
Southwest Jiaotong University, Chengdu 610031, P. R. China
[2] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, P. R. China
[3] College of Software Engineering, University of Sciences and Technology of China, Hefei 230052, P. R. China

Abstract

Several approaches based on rough set have been proposed for constructing decision tree in complete information systems. In fact, many information systems are incomplete in practical applications. In this paper, a new algorithm, Decision Tree Construction based on Rough Set Theory under Characteristic Relation (DTCRSCR), is proposed for mining classification knowledge from incomplete information systems. Its idea is that the attribute whose weighted mean roughness under the characteristic relation is the smallest will be selected as current splitting node. Experimental results show the decision trees constructed by DTCRSCR tend to have simpler structures and higher classification accuracy.

Keywords: Rough set, Decision tree, Weighted mean roughness, Characteristic relation

1. Introduction

Data mining, the efficient discovery of previously unknown patterns in large databases, has become a hot topic for decision makers. As an important and popular data mining technique, classification has been widely applied in business, medicine, industry, etc. In the past years, different classification methods have been proposed. None of classification approaches is always superior to the others in terms of classification accuracy. However, there are advantages and disadvantages to the use of each. The KNN technique requires only that the data be such that distances can be calculated. Bayesian classification assumes that the data attributes are independent with discrete values. Thus, although it is easy to use and understand, results may not be satisfactory. Decision tree techniques are easy to understand, but they may lead to overfitting. To avoid this, pruning techniques may be needed [1].

The ID3 algorithm for building the decision tree is based on the information theory and attempts to minimize the expected number of comparisons [2]. The basic idea of the induction algorithm is that the attribute which has maximum gain value of information entropy will be chosen as the current splitting node. Improvements on it, C4.5 [3] and C5.0 [4], allow the use of missing data and improved techniques for splitting. For example, when a decision tree is built by C4.5, missing data will be simply ignored. That is, the gain ratio is calculated by looking only at the other records that have a value for that attribute. To classify a record with a missing attribute value, the value for that item can be predicted based on what is known about the attribute values for the other records [1].

The classical rough set theory, proposed by Z. Pawlak in 1982, has been extensively studied in recent years. It is a new mathematical tool to deal with vagueness and uncertainty and has been applied successfully in data mining [5, 6, 8, 9, 10, 11, 12, 13]. For example, the rough set approach was adopted to classify different types of meteorological storm events responsible for summer severe weather in [13]. The rough approximation-based algorithms which can be used to select splitting node in the construction of decision tree were discussed in [8, 12]. However, these approaches are under the assumption that information systems are complete. In order to deal with incomplete data directly, an extension of conventional rough sets, the characteristic relation-based rough sets, was proposed. This extension better reflects the real conditions of incomplete information systems (IIS). A rule induction algorithm, accepting input data with both lost values and “do not care” conditions, is described in [5]. In [6], a method for incremental updating approximations of a concept in the IIS is proposed under the characteristic relation aiming to a dynamical attribute set.

To overcome the complexity in the structure expression and the difficulty in handling missing data, in this paper, a novel mining algorithm, Decision Tree

Construction based on Rough Set Theory under Characteristic Relation (DTCRSCR), is proposed, which firstly compute the weighted mean roughness of every condition attribute under the characteristic relation. Then, the attribute whose weighted mean roughness is the smallest will be selected as the splitting node. Experimental results show that the decision trees constructed by DTCRSCR tend to have simpler structures and higher classification accuracy.

The material of the paper is organized as follows. Section 2 introduces the basic concepts of characteristic relation-based rough sets and their extensions. The DTCRSCR method for constructing decision tree in the IIS under characteristic relations is illustrated in Section 3. Experimental comparison of the proposed method with C5.0 is given in Section 4. Section 5 concludes the research work of this paper.

2. Preliminaries

The followings are some terms, basic concepts of rough sets under characteristic relation and their extensions.

Definition 2.1 [7] An information system is defined as a pair $< U, A >$ where U is a non-empty finite set of objects, $A = C \cup D$ 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 $a \in A$ is associated with a set V_a of its value, called the domain of a .

Definition 2.2 [7] $< U, A >$ is an IIS if there exists a in A and x in U that satisfy that the value $a(x)$ is missing. All the missing values are denoted by " ? " or " * ", where the lost value is denoted by " ? ", "do not care" condition is denoted by " * ".

In [7], Grzymala-Busse presented that the characteristic set and characteristic relation can be determined by using the idea of blocks of attribute-values pairs which is defined as follow.

Definition 2.3 [7] Let b be an attribute and v be a value of b for some cases. If $t = (b, v)$ is an attribute-value pair, $v \neq ?$ and $*$, then a block of t, denoted $[t]$, is a set of all cases from U that attribute b have value v. If there exists a case x such that $v = b(x) = ?$, then the case x is not included in the block $[(b, v)]$ for any value v of attribute b. If there exists a case x such that $v = b(x) = *$, then the case x is included in the block $[(b, v)]$ for all value v of attribute b.

Definition 2.4 [7] Let $B \subseteq A$ be a subset of attributes. The characteristic set $I_B^C(x)$ is the intersection of blocks of attribute-value pairs (b, v) for all attributes b from B for which $b(x)$ is specified and $b(x) = v$.

Definition 2.5 [7] Let $B \subseteq A$ be a subset of attributes. The characteristic relation, denoted by C_B, is defined as: $(x, y) \in C_B \Leftrightarrow y \in I_B^C(x)$.

The characteristic relation C_B is reflexive but not symmetric and transitive. Obviously, it is a generalization of the indiscernibility relation in complete information systems.

Definition 2.6 [7] The lower and upper approximation of X with regard to B under the characteristic relation are

$X_B^C = \cup \{ I_B^C(x) \mid x \in X, I_B^C(x) \subseteq X \}$,

$X_C^B = \cup \{ I_B^C(x) \mid x \in X, I_B^C(x) \cap X \neq \varnothing \} = \cup \{ I_B^C(x) \mid x \in X \}$,

respectively.

Definition 2.7 Let $< U, A >$ is an IIS, $X \subseteq U$, $B \subseteq A$, $u_B(X) = card(X_B^C) / card(X_C^B)$ is a precision of X with regard to B under the characteristic relation $(0 \leq u_B(X) \leq 1)$. The weighted mean roughness of X with regard to B is defined as:

$$\beta(B) = 1 - (\sum_{j=1}^{m} \omega_j u_B(X_j)) \qquad (1)$$

Where j is the jth decision class of decision attributes. $j = 1, 2, ..., m$, m is the number of decision class; X_j is the jth set of decision class; ω_j , the percent of X_j in U , is defined as: $\omega_j = card(X_j) / card(U)$.

According to the definition of the weighted mean roughness under the characteristic relation, we know the value of $\beta(B)$ ranges from 0 to 1. When $\beta(B) = 0$, there is no uncertainty. When $\beta(B) = 1$, this means the set B leads to the greatest uncertain partition. As $\beta(B) \to 0$, the uncertainty decreases.

3. Decision Tree Construction based on Rough Ser Theory under Characteristic Relation

Based on the above definitions, we develop a novel algorithm which combine the characteristic relation-based rough sets theory for mining classification knowledge from incomplete information system in order to support decision-making effectively. It firstly computes the weighted mean roughness of every condition attribute under the characteristic relation. Then, the attribute whose weighted mean roughness is the smallest will be selected as the splitting node. The algorithm is illustrated in Table 1.

Algorithm: DTCRSCR

Input: Data set $sample$ (all of the values of attributes are discrete), The collection of condition attributes $attribute\_list$.

Output: $decision\_tree$.

Method:

Step1. With respect to $sample$, firstly compute the lower and upper approximation of every condition attribute with regard to every partition set of decision attribute. Then, calculate the weighted mean roughness of every condition attribute.

Step2. The attribute B whose weighted mean roughness under the characteristic relation-based rough sets is smallest will be selected as current splitting node.

Step3. For every value of the selected attribute B, we can obtain a data set Q of corresponding branch by using test $B.value = v$.

Step4. For every branch Q, if it has not reached the leaf then call $\mathrm{DTCRSCR}(Q, attribute\_list \setminus \{B\})$.

Step5. return.

Table 1: The DTCRSCR algorithm.

We have Table 2 to illustrate the above algorithm. Suppose F1、F2、F3 and F4 are the discrete condition attributes, CLASS is the decision attribute. Then, according to the above algorithm, the process of building decision tree is as follows:

ID	F1	F2	F3	F4	CLASS
1	?	0	1	0	1
2	0	1	2	0	1
3	0	1	2	0	2
4	*	0	1	1	1
5	?	0	1	1	1
6	0	1	2	1	1
7	1	0	0	*	2
8	1	0	0	2	2
9	0	1	2	2	1
10	*	0	0	2	1
11	?	1	1	1	1
12	1	0	2	1	2
13	1	1	0	*	1
14	0	1	0	0	2

Table 2: A data set.

Firstly, we compute the lower and upper approximations of every condition attribute with regard to every partition set of decision attribute under the characteristic relation, namely,

$$X_{F1}^{C}(CLASS=1)=\{10,4\}$$
$$X_{F2}^{C}(CLASS=1)=\varnothing$$
$$X_{F3}^{C}(CLASS=1)=\{1,11,4,5\}$$
$$X_{F4}^{C}(CLASS=1)=\varnothing$$
$$X_{C}^{F1}(CLASS=1)=\{1,10,11,12,13,14,2,3,4,5,6,7,8,9\}$$
$$X_{C}^{F2}(CLASS=1)=\{11,13,14,2,3,6,9,1,10,12,4,5,7,8\}$$
$$X_{C}^{F3}(CLASS=1)=\{10,13,14,7,8,12,2,3,6,9,1,11,4,5\}$$
$$X_{C}^{F4}(CLASS=1)=\{13,7,10,8,9,11,12,4,5,6,1,14,2,3\}$$
$$X_{F1}^{C}(CLASS=2)=\varnothing$$
$$X_{F2}^{C}(CLASS=2)=\varnothing$$
$$X_{F3}^{C}(CLASS=2)=\varnothing$$
$$X_{F4}^{C}(CLASS=2)=\varnothing$$
$$X_{C}^{F1}(CLASS=2)=\{12,13,7,8,14,2,3,6,9\}$$
$$X_{C}^{F2}(CLASS=2)=\{1,10,12,4,5,7,8,11,13,14,2,3,6,9\}$$
$$X_{C}^{F3}(CLASS=2)=\{10,13,14,7,8,12,2,3,6,9\}$$
$$X_{C}^{F4}(CLASS=2)=\{11,12,4,5,6,10,8,9,13,7,1,14,2,3\}$$

Next, the weighted mean roughness of every condition attribute under the characteristic relation-based rough sets is obtained by (1).

The weighted mean roughness of F1：

$$1-(\frac{2}{14}\times\frac{9}{14}+\frac{0}{9}\times\frac{5}{14})=\frac{178}{196}.$$

The weighted mean roughness of F2：

$$1-(\frac{0}{14}\times\frac{9}{14}+\frac{0}{14}\times\frac{5}{14})=1.$$

The weighted mean roughness of F3：

$$1-(\frac{4}{14}\times\frac{9}{14}+\frac{0}{14}\times\frac{5}{10})=\frac{160}{196}.$$

The weighted mean roughness of F4：

$$1-(\frac{0}{14}\times\frac{9}{14}+\frac{0}{14}\times\frac{5}{14})=1.$$

Obviously, the weighted mean roughness of F3 is smaller than that of the others. The attribute F3 is selected as the root of the decision tree. Then, the other attributes (F1, F2, F4) will be tested on every branch of the root respectively. The algorithm continues recursively by adding new subtrees to each branching arc.

4. Experimental Evaluation

Experimental comparison of DTCRSCR with C5.0 is given in this Section. Experiments are performed on an 864MHz Pentium Server with 512MB of memory, running windows XP server and SQL server 2000. Algorithms are coded in C#. We choose 10 data sets, publicly available from the UC Irvine Machine Learning Database Repository [14], as benchmark

datasets for the performance tests. The descriptions of experimental data are shown in Table 3.

Data set	Tuples	?	*	Attribute(C/D)
monks-1_learn.tab	124	R	R	6/1
monks-1_test.tab	432	R	R	6/1
monks-2_learn_tab	169	R	R	6/1
monks-3_learn.tab	122	R	R	6/1
monks-3_test.tab	432	R	R	6/1
breast-cancer.tab	286	E	R	9/1
hayes-roth_learn.tab	132	R	R	5/1
flare1.tab	322	R	R	13/1
post-operative.tab	90	E	R	8/1
lymphography_tab	148	R	R	18/1

Table 3: The description table of experimental data.

Where

E: It means the data set contains '?' or '*'.

R: It means we randomly replace some data with '?' or '*' in the data set.

The experimental results are listed in Table 4.

From Table 4, we obtain the following results:

(1) In most of data sets (7 out of 10 data sets), the decision trees (here the number of the leaves and the nodes of the whole tree are listed) constructed by DTCRSCR tend to have simpler structure and higher classification accuracy than C5.0.

(2) Only in the data set "monks-1_test.tab", the decision trees constructed by DTCRSCR have simpler structure and the same classification accuracy than C5.0.

(3) In these 2 data sets "monks-3_test.tab" and "lymphography_tab", the decision tree constructed by DTCRSCR has the more complex structures and the same classification accuracy than C5.0.

Therefore, the decision trees constructed by DTCCRS generally have simpler structure and higher accuracy than that constructed by C5.0.

Data set	C5.0(leaves/nodes)	C5.0(accuracy)	DTCCRS(leaves/nodes)	DTCCRS(accuracy)
monks-1_learn.tab	50/59	91.87%	46/62	98.374%
monks-1_test.tab	220/371	50.926%	220/369	50.926%
monks-2_learn_tab	146/255	90.374%	124/218	94.083%
monks-3_learn.tab	88/203	83.097%	86/161	86.066%
monks-3_test.tab	22/27	100%	25/34	100%
breast-cancer.tab	228/248	96.503%	90/135	97.902%
hayes-roth_learn.tab	33/41	85.610%	29/40	96.212%
flare1.tab	64/132	84.290%	38/75	99.690%
post-operative.tab	61/104	90.412%	56/106	91.111%
lymphography_tab	47/138	98.649%	59/144	98.649%

Table 4: A performance comparison of DTCCRS with C5.0.

5. Conclusions

Decision tree is one of the most significant classification methods applied in data mining. However, most decision tree algorithms can not handle missing data effectively. In this paper, we presented a new algorithm DTCRSCR based on the characteristic relation-based rough sets for construction of the decision tree. It firstly computes the weighted mean roughness of every condition attribute under the characteristic relation. Then, the attribute whose weighted mean roughness is the smallest will be selected as the splitting node. Experimental results show that the decision trees constructed by DTCRSCR generally tend to have simpler structures and higher classification accuracy than C5.0. In addition, because DTCRSCR may be quite time-consuming to use, an interesting direction of our future work is to study how to improve the DTCRSCR algorithms for better performance.

References

[1] M. H. Dunham, *Data Mining – Introductory and Advanced Topics*, Prentice Hill, 2002.

[2] J. R. Quinlan, Induction of decision tree. *Machine Learning*, 11(1): 80-106, 1986.

[3] J. R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann, 1993.

[4] http://www.rulequest.com/see5info.html, 2001.

[5] J. W. Grzymala-Busse, S. Siddhaye, Rough set approaches to rule induction from incomplete data. *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge- Based Systems*, pp. 923–930, 2004.

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

[7] J. W. Grzymala-Busse, Characteristic relations for incomplete data: A generalization of the indiscernibility relation. *Transactions on Rough Sets IV*, pp. 58–68, 2005.

[8] J. Wei, D. Huang, S. Wang, Z. Ma, ."Rough Set Based Decision Tree. *Proceedings of the 4th World Congress on Intelligent Control and Automation*, 7: 426-430, 2002.

[9] M. Beynon, Reducts within the variable precision rough set model: a further investigation. *European Journal of Operational Research*, 134:592-605, 2001.

[10] J. Stefanowski, A. Tsoukias, Incomplete information tables and rough classification. *Computational Intelligence*, 17:545–566, 2001.

[11] J. G. Dy, C. E. Brodley, Feature selection for unsupervised learning. *The Journal of Machine Learning Research archive*, 5:845–889, 2004.

[12] Y. Jiang, Z. Li, Q. Z, Y. Liu, New method for constructing decision tree based on rough sets theory. *Computer Application*, 24(8):21-23, 2004.

[13] J. F. Peters, Z. Suraj, S. Shan, S. Ramanna, W. Pedrycz, N. Pizzi, Classification of meteorological volumetric radar data using rough set methods. *Pattern Recognition Letters*, 24(6):911–920, 2003.

[14] http://www.ics.uci.edu/˜mlearn/MLRepository.html. (UCI Machine Learning Repository).