Research on Decision Tree Based on Rough Set

Wei Wei[1], Mingwei Hui[1], Beibei Zhang[1], Rafal Scherer[2], Robertas Damaševičius[3]

[1] School of Computer Science and Engineering, Xi'an University of Technology, China
[2] Czestochowa University of Technology Al, Poland
[3] Multimedia Engineering Department, Kaunas University of Technology, Lithuania
weiwei@xaut.edu.cn, 1614821053@qq.com, bbzhang115@hotmail.com, rafal.scherer@pcz.pl, robertas.damasevicius@ktu.lt

Abstract

This paper proposes a decision tree generation method based on variable precision rough set theory. The proposed method mainly deals with the uncertain information in the decision tree process and allows a certain degree of noise interference during classification. It mainly summarizes based on entropy and Decision tree construction method based on rough set theory. Two well-known algorithms, ID3 and C4.5, are discussed in terms of entropy. Decision tree based on rough set theory and based on variable precision are introduced in terms of rough set. Decision tree constructed by rough set theory. Then the difference between the method based on rough set theory and basic entropy is discussed. Although the decision tree constructed based on entropy and rough set theory can achieve a good match with the original data set, but it reduces its generalization ability for future data. Compared with the traditional decision tree construction algorithm based on entropy and rough set, the decision tree construction method based on the variable precision rough set theory constructs a simple decision tree structure, which improves the generalization of the decision tree. It also has a certain ability to suppress noise at the same time.

Keywords: Decision tree, Rough set theory, Variable precision rough set theory

1 Introduction

1.1 Research Status of the Combination of Rough Set and Decision Tree Technology

In the 1970s, Polish scholar Z. Pawlak and some logicians from the Polish Academy of Sciences and the University of Warsaw in Poland began to work together on the logical characteristics of information systems, and then the rough set theory was discovered. In 1982, in Z. After Pawlak published the classic paper "Rough Sets", rough set theory was born. After that, the blooming colors of rough set theory attracted many logicians, mathematicians and computer researchers who perfected the theory and practical application of rough set. He has done a lot of research work in this respect. Until Z. Pawlak published a series of articles in 1991 and 1992, he made a complete summary of the results of practical and theoretical research during this period [1]. After that, the rough set is in Many fields have been widely used. With the success of international conferences related to rough sets, rough sets have attracted the attention of more and more researchers, who have begun to engage in research in this field, which makes rough set theory rapidly Development. Nowadays, the research of rough set theory has become a hot spot in the field of artificial intelligence research. It has attracted the attention of researchers in the research fields of process control, decision analysis, knowledge acquisition, and machine learning. Decision tree method is through a series of rules the method of classifying data is simple and clear, which makes it widely used in data mining. After JRQuilan proposed the ID3 decision tree algorithm, some researchers began to improve and perfect the algorithm. In particular, some scholars have introduced rough set theory and variable precision rough set theory into the generation of decision trees. This innovation not only allows the existence of noisy data, but also further improves the generalization ability of decision trees. The current improvement of decision tree algorithms There are two main aspects. On the one hand, it is to construct a better algorithm for node selection. On the other hand, it reduces the complexity of the decision tree while ensuring the readability and fault tolerance of the generated rules.

1.2 The Main Research Work of This Article

The main research work of this paper is a decision tree generation method based on variable precision rough set theory, and a series of examples are compared to prove that this algorithm is superior to other algorithms.

Because rough set theory can only process pre-assumed data, it cannot handle data with noise well.

*Corresponding Author: Wei Wei; E-mail: weiwei@xaut.edu.cn
DOI: 10.53106/160792642021112206015

Decision trees generated based on rough set theory are also prone to inconsistencies with actual data, which cannot be better and accurate. Based on the idea of variable precision rough set theory, this article proposes an improvement to the previous decision tree generation method based on rough set theory, and further proposes a decision tree generation method based on variable precision rough set theory. Variable precision rough set theory A certain degree of error is allowed when dealing with some uncertain data, which makes the generated decision tree can better suppress part of the noisy data, so this method has certain advantages over the construction method of rough set [2].

The main contents of this article are arranged as follows:

Chapter 1: Introduction. Analyzes the research significance of this article. Briefly describes the historical development, background significance and research status of decision trees and rough set theory, and deeply analyzes the current rough set theory-based decision tree generation method produced by the combination of the two. Compared with the advantages of the previous algorithm.

Chapter 2~5: The improvement of the decision tree generation method based on rough set theory. The background and principle of rough set theory are briefly described, and the decision tree generation method based on rough set theory is introduced in detail. After discovering the shortcomings of this method, it is proposed Based on the improved method of variable precision rough set theory, finally compare the algorithms with examples.

Chapter 6: Conclusion. Make a general summary of the full text and make predictions for the next work.

2 Basic Theory of Rough Sets

Rough set theory is based on classification. It does not require any prior knowledge to describe the inaccuracy of things. It only depends on the expression system and is calculated through upper and lower approximations. Rough set theory explains, the inaccuracy of objective affairs is caused by insufficient knowledge, that is, if you have enough knowledge, the classification ability will be greatly improved. An example is given below to explain the related concepts of rough sets [3]. (✓ respectively means excellent, expensive and profitable, × on the contrary.)

The following rules can be obtained from Table 1.

Table 1. Rough set examples (a)

Commodity	1	2	3	4	5	6
Quality	✓	✓	✓	×	×	×
Price	×	✓	✓	✓	✓	×
Profit and loss	✓	×	✓	×	×	×

Determined rules:

(1) Good quality and low price are profitable, such as commodity 1.

(2) Poor quality and expensive or cheap ones are loss-making, such as commodities 4, 5 and 6.

Possible rules: Good quality and high price may be loss or profit, such as goods 2 and 3.

So in summary, the 6 products can be divided into 3 different situations according to the "profit and loss".

Profit: {1}; possible loss and possible profit: {2, 3}; loss: {4, 5, 6};

The lower approximate set: {4, 5, 6} must be a loss; the upper approximate set: {2, 3, 4, 5, 6}, the complement {1} must be profitable.

Boundary set: {2, 3} is the difference between the upper and lower approximate sets, which may be loss or profit. (✓ respectively means excellent, expensive, long and profitable, × on the contrary.)

Table 2. Rough set examples (b)

Commodity	1	2	3	4	5	6
Quality	✓	✓	✓	×	×	×
Price	×	✓	✓	✓	✓	×
Shelf life	✓	×	✓	×	✓	×
Profit and loss	✓	×	✓	×	×	×

Rough set can use upper and lower approximate sets to describe the uncertainty of knowledge. Generally, the uncertainty expressed by rough set is roughness. Roughness is related to the mastery of knowledge, and the roughness can disappear with the increase of knowledge. For example, adding the "shelf life" attribute on the basis of Table 1, as shown in Table 2, it is obvious that the upper and lower approximations of the "loss" are equal, and the roughness disappears.

The following describes some characteristics and related definitions of rough set theory.

Indistinguishable relationship:

The indistinguishable relationship is the most basic concept in rough set theory. In rough set, the information to be classified is often divided into different categories according to the difference of various attributes. If there are exactly two pieces of information that have the same attribute value, then they are indistinguishable. In this case, they cannot be distinguished based on the existing knowledge, thus forming an equivalence relationship. For example, the goods 2 and 3 in Table 1, in the "quality Under the premise that the two conditional attributes of" and "price" are the same, their decision-making attributes "profit and loss" are different. This is obviously an indistinguishable relationship [4].

Boundary:

In our understanding of some things, some can be accurately separated and some cannot be divided into boundaries, such as height and length. Everyone has their own criteria for determining these vague concepts, and there is no way to unify them. Rough sets the boundary in the theory is a fuzzy concept, that is, a

fuzzy concept has no clear boundary. In order to illustrate the ambiguity, for each inaccurate concept, the more accurate concepts of lower approximation and upper approximation are often used to divide the set. Among them, the lower approximation of set A represents the elements in the set that can be clearly divided into set A, and the upper approximation is those elements that may belong to set A. The difference between the two is the boundary of this set. This boundary is caused by the incompleteness It is classified into all elements in this set or its complement [5]. If the boundary is not empty, then A is an exact set.

Definition 2.1 An information system is a four-tuple $IP=(U, A, V, f)$, U is a collection of finite objects, called the universe of discourse. A represents the attribute set, which is usually divided into two subsets C and D, which respectively represent the condition attribute set and the decision attribute set. $f: U\times A\to V$ is an information function, where $V=U_{a\in A}V_a$ and V_a are the domains of attribute $a\in A$.

Definition 2.2 Given an information system $IP=(U, A, V, f)$, Let $x, y\in U, M\subseteq A$, if and only if $\forall a\in M, f(x,a)=f(y,a)$, it is said that A and B are indistinguishable from C. For $\forall M\in A$, define an equivalence relation on U, denoted as $IND(M)$.

Any subset X of U can be described by $<\underline{B}(X), \overline{B}(X)>$, which are the lower approximation and upper approximation of B with respect to X. As shown in Figure 1, the lower approximation $\underline{B}(X)$ is the union of all equivalence classes that must belong to X based on the knowledge in B. The upper approximation $\overline{B}(X)$ is based on the knowledge in B to judge the union of all equivalence classes that may belong to X. Among them, $NEG_B(X)$ is the negative region of set X with respect to B, which means that it certainly does not belong to the union of equivalence classes of X. $BN_B(X)$ is called the boundary of set X with respect to B, that is, the difference between the upper approximation and the lower approximation. If $BN_B(X)$ is empty, then it can be said that the set X is clear about B, otherwise it is rough. X can also be defined by B relatively similar to the clarity $a_B(X)=\frac{|\underline{B}(X)|}{|\overline{B}(X)|}$, if $a_B(X)=1$, then it can be said that the set X is clear with respect to B; if $a_B(X)<1$, then it is rough [6].

3 Improvement of Decision Tree Generation Method Based on Rough Set Theory

Definition 3.1 Let $A\subseteq C, B\subseteq D$, $A^*=\{X_1, X_2, \cdots, X_n\}$

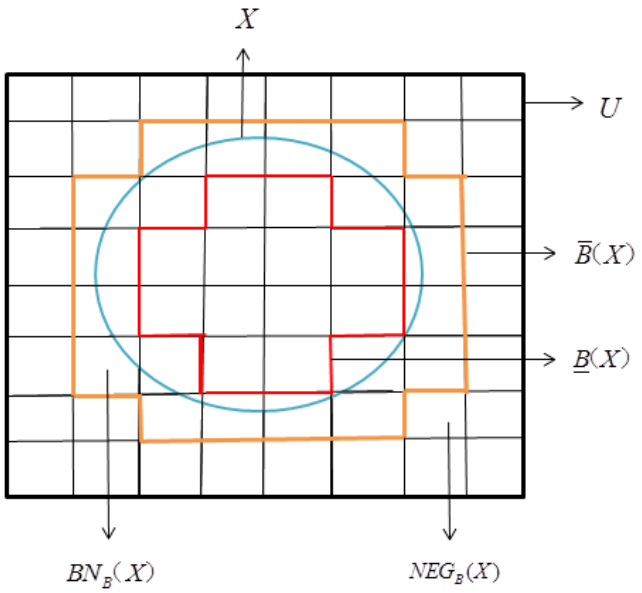

Figure 1. The approximation set

and $B^*=\{Y_1, Y_1, ..., Y_1\}$ denote the partition of U derived from the equivalence relations $IND(A)$ and $IND(B)$, where $IND(A)$ and $IND(B)$ are the equivalence relations derived from A and B. The clear area is defined as

$$Exp_A(B^*)=\cup_{Y_i\in B^*}\underline{A}(Y_i) \quad (1)$$

Where $\underline{A}_\beta(Y_i)$ represents the lower approximation of $IND(A)$ relative to Y_i. The definition of the clear zone is to combine the negative zone and the lower approximation. In this way, two more clear branches will be produced when making decisions.

Definition 3.2 Let $A\subseteq C, B\subseteq D$, $A^*=\{X_1, X_2, \cdots, X_n\}$ and $B^*=\{Y_1, Y_1, ..., Y_1\}$ denote the partition of U derived from the equivalence relations $IND(A)$ and $IND(B)$, where $IND(A)$ and $IND(B)$ are the equivalence relations derived from A and B. The non-explicit area is defined as

$$Imp_A(B^*)=\cup_{Y_i\in B^*}(\overline{A}(Y_i)-\underline{A}(Y_i)) \quad (2)$$

Where $\underline{A}_\beta(Y_i)$ represents the upper approximation of $IND(A)$ relative to Y_i, and the definition of the non-clear area represents the union of the boundaries of each division. So, we can get:

$$Exp_A(B^*)\cup Imp_A(B^*)=U.$$

The original intention of building a decision tree is to obtain more useful knowledge, that is, the knowledge will become clearer and clearer through the tool of the decision tree, so that people are more easily familiar with the system [7]. Therefore, the method of node selection based on rough sets It can be simply

described as that when we evaluate an attribute, we divide the data set into two parts according to the value of the attribute: one is the clear area and the other is the non-clear area. After this, we can calculate the size of the two parts of all attributes, and pass Compare, select all the attributes with the largest clear area or the smallest non-clear area as the sub-nodes. Comparing the entropy-based method with the rough set-based method, it is found that the clear area does not contribute to the information gain, so the final information gain is only It is related to the non-explicit area. When the entropy of the non-explicit area is the smallest, the current minimum attribute will be selected. However, the proposed method also has certain limitations. The rough set-based decision tree generation method can only be used when the object is strictly classified an accurate decision tree can be constructed only under circumstances, and it cannot allow real data to be noisy. Therefore, we propose a decision tree generation method based on variable precision rough sets.

4 Decision Tree Generation Method Based on Variable Precision Rough Set Theory

Because rough set theory cannot deal with inaccurate or uncertain original data sets in the process of constructing decision trees, the use of rough set theory cannot well describe some uncertain or inaccurate practical problems. If the original data set is completely acceptable in the process of constructing a decision tree using rough set theory, the division of the instance by the generation method must be completely accurate. The instance either belongs to a certain approximate interval or does not belong to it, which makes the final generated decision tree accurate reflect the original data set. Although the decision tree generated in this way can accurately express the relevant information of the original data set, it cannot guarantee good generalization ability for unknown data. For data in the real world, most of them will be affected by noise Impact, if the generated decision tree can be accurate enough that even noise can be accurately matched, then it will not be able to effectively express the characteristics of the original data, and it will not be able to effectively predict the unknown data category [8].

In order to better deal with the problem of noise, we need to introduce variable precision rough set theory. Variable precision rough set theory does not strictly define the boundary, but defines a relative threshold so that a part of valuable noise can also become a sample decision tree, the decision tree model established in this way greatly improves the generalization ability.

4.1 Variable Precision Rough Set Theory

Assuming that $(U, \tilde{R})$ is an approximate space, $R^* = \{E_1, E_2, \cdots, E_n\}$ represents the set of equivalence classes contained in $\tilde{R}$. For any subset $X \subseteq U$, the approximate β of $\tilde{R}$ with respect to X is defined as

$$\underline{R}_\beta X = \bigcup \{E_i \in R^* \mid c(E_i, X) \le \beta\} \quad (3)$$

$\underline{R}_\beta X$ is also called the positive region of β, denoted as $Pos_\beta(X)$. The upper β of X about $\tilde{R}$ is approximately defined as

$$\overline{R}_\beta X = \bigcup \{E_i \in R^* \mid c(E_i, X) < 1-\beta\} \quad (4)$$

The β boundary of X with respect to $\tilde{R}$ is defined as

$$BN_\beta X = \bigcup \{E_i \in R^* \mid \beta < c(E_i, X) < 1-\beta\} \quad (5)$$

The difference between the upper and lower approximations is the β boundary of X with respect to $\tilde{R}$.

The β negative area of X with respect to $\tilde{R}$ is defined as

$$NEG_\beta X = \bigcup \{E_i \in R^* \mid c(E_i, X) \ge 1-\beta\} \quad (6)$$

Where

$$c(E_i, X) = \begin{cases} 1-\frac{|X \cap E|}{|E|}, & |X|>0 \\ 0, & |X|=0 \end{cases} \quad (7)$$

Where $\frac{|X \cap E|}{|E|}$ is the inclusion degree, which indicates the degree to which the equivalence class Y of X formed by the relationship R is included in the set X. The value range of β is $0 \le \beta \le 0.5$. The variable precision rough set model can be compared with the original rough set model. It is obtained that when $\beta = 0$ time the variable precision rough set model will become the rough set model.

4.2 Decision Tree generation Method Based on Variable Precision Rough Set Theory

First of all, for the decision tree generation method of variable precision rough set theory, we have to propose two new concepts, that is, the improved variable precision clear area and variable precision non-clear area. They will introduce an error threshold β to replace the rough set theory the clear area and non-clear area of, become the rule for selecting attributes in the decision tree induction [9].

Definition 4.2.1 Let $A \subseteq C, B \subseteq D$, $A^* = \{X_1, X_2, \cdots, X_n\}$ and $B^* = \{Y_1, Y_1, ..., Y_1\}$ denote the partition of U derived from the equivalence relations $IND(A)$ and $IND(B)$, where $IND(A)$ and $IND(B)$ are the equivalence

relations derived from A and B. The variable precision area is defined as

$$Exp_{A\beta}\left(B^{*}\right)=\bigcup_{Y_i\in B^{*}}\underline{A}_{\beta}\left(Y_i\right) \quad (8)$$

Among them, $\underline{A}_{\beta}\left(Y_i\right)$ is the β lower approximation of Y_i relative to $IND(A)$.

After the variable precision clear area is expanded, the variable precision unclear area will become as follows.

Definition 4.2.2 Let $A\subseteq C, B\subseteq D$, $A^{*}=\{X_1, X_2, ..., X_n\}$ and $B^{*}=\{Y_1, Y_1, ..., Y_1\}$ denote the partition of U derived from the equivalence relations $IND(A)$ and $IND(B)$, where $IND(A)$ and $IND(B)$ are the equivalence relations derived from A and B. The variable precision area is defined as

$$Imp_{A\beta}\left(B^{*}\right)=\bigcup_{Y_i\in B^{*}}\left(\overline{A}_{\beta}\left(Y_i\right)-\underline{A}_{\beta}\left(Y_i\right)\right) \quad (9)$$

Among them, $\underline{A}_{\beta}\left(Y_i\right)$ is the β lower approximation of Y_i relative to $IND(A)$, $\underline{A}_{\beta}\left(Y_i\right)$ is the β upper approximation of Y_i relative to $IND(A)$.

Take an example to illustrate the advantages of the classic rough set theory after introducing the error threshold β. As shown in Figure 2, U is the universe of discourse, X is a non-empty subset $E_i, E_j \in \{E_1, E_2, ..., E_n\}$. $\underline{R}(X)$ is the lower approximation of X with respect to the equivalence relation R. Take the range of the error threshold β as $0\le\beta<0.5$. First look at E_i. In rough set theory, it is divided into the boundary, but after taking the range of the error threshold β as $0<\beta<0.5$, $c(E_i, X)\le\beta$, that is, under the variable precision rough set It belongs to the lower approximation. In reality, there may be noise in E_i, but because the rough set theory cannot be divided into the lower approximation, the established decision tree will have a gap with the actual data, which affects the generalization ability. Introducing errors After the threshold β, if E_i satisfies the error range, it can be classified into the lower approximation of the variable precision rough set, and become the data that can be actually operated, avoiding the abandonment of this part of data due to noise and making the established decision tree not In the same way, E_j is divided into the lower approximation of $U-X$ [10-11].

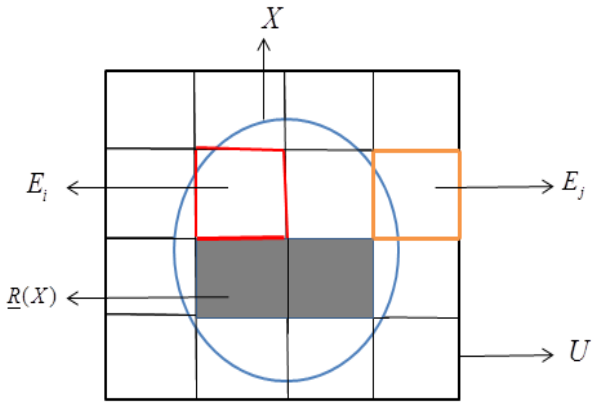

Figure 2. An improved example

It is very clear that after the decision tree is established and the error threshold β is introduced, the range of the variable precision area will be significantly improved. Because in the actual data processing, there will always be more or less noise, so After processing, the tolerance range of the lower approximation can be expanded, and then the range of the clear area is also expanded. Then when processing realistic data, the decision tree will be closer to the reality, improve the generalization ability, and reduce the complexity of the tree. Make the decision tree can handle more kinds of data types, instead of only dealing with a single data type like the rough set-based decision tree construction method, without more selectivity.

5 Comparison of Examples of Decision Trees Generated by Various Algorithms

5.1 Examples and Comparisons of Decision Tree Generation Methods Based on Basic Entropy

ID3 Algorithm

As shown in Figure 3, this example has 14 samples, 4 condition attributes, and 1 decision attribute.

Calculate the information entropy of each category

$$Entropy(Out)=\frac{5}{14}\times\left[-\frac{3}{5}log_2\frac{3}{5}-\frac{2}{5}log_2\frac{2}{5}\right]+\frac{4}{14}\times\left[-log_2 1\right]+\frac{5}{14}\times\left[-\frac{3}{5}log_2\frac{3}{5}-\frac{2}{5}log_2\frac{2}{5}\right]=0.6935$$

$$Entropy(Tem)=\frac{4}{14}\times\left[-\frac{1}{2}log_2\frac{1}{2}-\frac{1}{2}log_2\frac{1}{2}\right]+\frac{6}{14}\times\left[-\frac{4}{6}log_2\frac{4}{6}-\frac{2}{6}log_2\frac{2}{6}\right]+\frac{4}{14}\times\left[-\frac{3}{4}log_2\frac{3}{4}-\frac{1}{4}log_2\frac{1}{4}\right]=0.9111$$

Sample	Outlook	Temperature	Humidity	Wind	PlavTennis
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rain	Mild	High	Weak	Yes
D_5	Rain	Cool	Normal	Weak	Yes
D_6	Rain	Cool	Normal	Strong	No
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rain	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rain	Mild	High	Strong	No

Figure 3. Basic entropy dataset

$$Entropy(Hum)=\frac{7}{14}\times\left[-\frac{4}{7}log_2\frac{4}{7}-\frac{3}{7}log_2\frac{3}{7}\right]+\frac{7}{14}\times\left[-\frac{6}{7}log_2\frac{6}{7}-\frac{1}{7}log_2\frac{1}{7}\right]=0.7885$$

$$Entropy(Win)=\frac{8}{14}\times\left[-\frac{2}{8}log_2\frac{2}{8}-\frac{6}{8}log_2\frac{6}{8}\right]+\frac{6}{14}\times\left[-\frac{3}{6}log_2\frac{3}{6}-\frac{3}{6}log_2\frac{3}{6}\right]=0.8922$$

$$Entropy(PT)=-\frac{5}{14}\times log_2\frac{5}{14}-\frac{9}{14}\times log_2\frac{9}{14}=0.9403$$

Then calculate the information gain of each conditional attribute

$$Gain(PT,Out)=0.2468\quad Gain(PT,Tem)=0.0292$$
$$Gain(PT,Win)=0.0481\quad Gain(PT,Hum)=0.1518$$

Compare the information gain of each conditional attribute and find the one with the largest information gain as the root node. The information gain is the largest. Taking it as the root node, the instance can be divided into 3 parts.

$$O_1=\{D_1,D_2,D_8,D_9,D_{11}\}$$
$$O_2=\{D_3,D_7,D_{12},D_{13}\}$$
$$O_3=\{D_4,D_5,D_6,D_{10},D_{14}\}$$

Same as above, calculate the information gain of the remaining conditional attributes, and then *Humidity* and *wind* should be selected as the root node.

For node *Humidity*, the instance can be divided into two parts.

$$H_1=\{D_1,D_2,D_8\}\quad H_2=\{D_9,D_{11}\}$$

For node *Wind*, it can also be divided into two parts

$$W_1=\{D_4,D_5,D_{10}\}\quad W_2=\{D_6,D_{14}\}$$

Then a decision tree is established, as shown in the figure below.

C4.5 Algorithm

Assuming that there is only one possibility for each category in each attribute, its information entropy is zero, so the information gain has no way to make a classification effect, so there is an information gain rate. Still use the data set in Figure 3, now calculate the classification information, the previous data can be used [12].

$$Split(PT,Out)=-\frac{5}{14}\times log_2\frac{5}{14}-\frac{5}{14}\times log_2\frac{5}{14}-\frac{4}{14}\times log_2\frac{4}{14}=1.5774$$

$$Split(PT,Tem)=-\frac{4}{14}\times log_2\frac{4}{14}-\frac{4}{14}\times log_2\frac{4}{14}-\frac{6}{14}\times log_2\frac{6}{14}=1.5567$$

$$Split(PT,Hum)=-2\times\frac{7}{14}\times log_2\frac{7}{14}=1$$

$$Split(PT,Win)=-\frac{8}{14}\times log_2\frac{8}{14}-\frac{6}{14}\times log_2\frac{6}{14}=0.9852$$

Next, calculate the information gain rate

$$GainRatio(PT,Out)=\frac{Gain(PT,Out)}{Split(PT,Out)}=\frac{0.2468}{1.5774}=0.1565$$

$$GainRatio(PT,Tem)=\frac{Gain(PT,Tem)}{Split(PT,Tem)}=\frac{0.0292}{1.5567}=0.0188$$

$$GainRatio(PT, Hum) = \frac{Gain(PT, Hum)}{Split(PT, Hum)} = \frac{0.1518}{1} = 0.1518$$

$$GainRatio(PT, Win) = \frac{Gain(PT, Win)}{Split(PT, Win)} = \frac{0.0481}{0.9852} = 0.0488$$

Outlook has the largest information gain rate, so choose *Outlook* as the root node. Same as above, can be divided into three parts.

$$O^{\wedge}_1 = \{D_1, D_2, D_8, D_9, D_{11}\}$$
$$O^{\wedge}_2 = \{D_3, D_7, D_{12}, D_{13}\}$$
$$O^{\wedge}_3 = \{D_4, D_5, D_6, D_{10}, D_{14}\}$$

Repeat the above process, taking *Sunny* child node as an example.

$$Entropy^{\wedge}(PT) = -\frac{2}{5}\times log_2\frac{2}{5} - \frac{3}{5}\times log_2\frac{3}{5} = 0.9710$$

$$Entropy^{\wedge}(Tem) = \frac{2}{5}\times log_2 1 + \frac{2}{5}\times\left[-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}\right] + \frac{1}{5}\times log_2 1 = 0.4$$

$$Entropy^{\wedge}(Hum) = \frac{2}{5}\times log_2 1 + \frac{3}{5}\times log_2 1 = 0$$

$$Entropy^{\wedge}(Win) = \frac{2}{5}\times\left[-\frac{1}{2}\times log_2\frac{1}{2} - \frac{1}{2}\times log_2\frac{1}{2}\right] + \frac{3}{5}\times\left[-\frac{2}{3}\times log_2\frac{2}{3} - \frac{1}{3}\times log_2\frac{1}{3}\right] = 0.9510$$

Calculate information gain

$$Gain^{\wedge}(PT, Tem) = 0.571$$
$$Gain^{\wedge}(PT, Win) = 0.9710$$
$$Gain^{\wedge}(PT, Hum) = 0.02$$

Calculate classification information:

$$Split^{\wedge}(PT, Tem) = -\frac{2}{5}\times log_2\frac{2}{5} - \frac{2}{5}\times log_2\frac{2}{5} - \frac{1}{5}\times log_2\frac{1}{5} = 1.5220$$

$$Split^{\wedge}(PT, Hum) = -\frac{3}{5}\times log_2\frac{3}{5} - \frac{2}{5}\times log_2\frac{2}{5} = 0.9710$$

$$Split^{\wedge}(PT, Win) = -\frac{3}{5}\times log_2\frac{3}{5} - \frac{2}{5}\times log_2\frac{2}{5} = 0.9710$$

Finally calculate the information gain rate

$$GainRatio^{\wedge}(PT, Tem) = \frac{Gain^{\wedge}(PT, Tem)}{Split^{\wedge}(PT, Tem)} = \frac{0.571}{1.5220} = 0.3752$$

$$GainRatio^{\wedge}(PT, Hum) = \frac{Gain^{\wedge}(PT, Hum)}{Split^{\wedge}(PT, Hum)} = \frac{0.02}{0.9710} = 0.0206$$

$$GainRatio^{\wedge}(PT, Win) = \frac{Gain^{\wedge}(PT, Win)}{Split^{\wedge}(PT, Win)} = \frac{0.9710}{0.9710} = 1$$

The information gain rate of *Wind* is the largest, so next select *Wind* as the node. Because the decision attributes of *Overcast* are pure, so this one ends. Same as above, after selecting *Rain*, calculate the information gain rate and find the information gain rate of *Temperature* and *Humidity* The same, then we choose one arbitrarily and choose *Humidity* as the node, then the decision tree is actually similar to Figure 4.

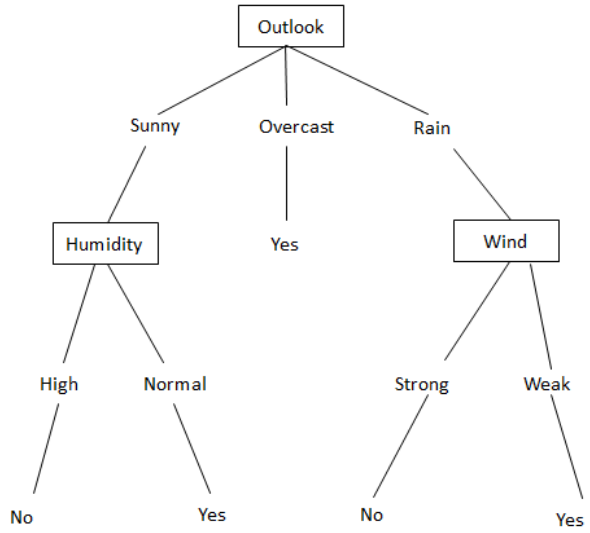

Figure 4. Decision tree based on entropy

Summarizing these two methods of constructing decision trees based on entropy, it is found that they completely match the training set, which is the example, but they cannot allow noise in the training set, and they are only limited to the case where the number of training sets is small, so this greatly affects the practicality of this type of algorithm [13-14].

5.2 Examples and Comparisons of Decision Tree Generation Methods Based on Rough Set Theory

The following will compare the algorithm based on the decision tree generation method based on rough set and variable precision rough set. Through the comparison, we will understand the advantages of the decision tree constructed after step-by-step improvement and the superiority of the algorithm [15].

Table 3 is a training set with 10 instances, including 4 attributes, where A, B, and C are conditional attributes, and D is a decision attribute.

Table 3. Rough set data set

ID	A	B	C	D
1	1	2	1	0
2	2	1	1	0
3	2	2	1	0
4	1	1	1	0
5	1	1	2	0
6	3	2	1	0
7	3	2	2	1
8	2	1	2	1
9	1	2	1	1
10	1	2	2	2

First divide each attribute.

The conditional attribute A divides the complete set U into:

$$A^* = \{A_1, A_2, A_3\} = \{\{1,4,5,9,10\},\{2,3,8\},\{6,7\}\}$$

The conditional attribute B divides the complete set U into:

$$B^* = \{B_1, B_2\} = \{\{2,4,5,8\},\{1,3,6,7,9,10\}\}$$

The conditional attribute C divides the complete set U into:

$$C^* = \{C_1, C_2\} = \{\{1,2,3,4,6,9\},\{5,7,8,10\}\}$$

The conditional attribute D divides the complete set U into:

$$D^* = \{D_1, D_2, D_3\} = \{\{1,2,3,4,5,6\},\{7,8,9\},\{10\}\}$$

Then calculate the clear area of each condition attribute relative to the decision attribute.

$$card\left(Exp_A\left(D^*\right)\right) = card\left(Exp_B\left(D^*\right)\right)$$
$$= card\left(Exp_C\left(D^*\right)\right) = 0$$

It is found that they do not have a clear area, so no matter which one of the three we choose, it will not help the classification. However, through comparison, we find that there are fewer types of B and C, and we can choose B as the root node arbitrarily. The established decision the tree is shown in Figure 5.

Through the decision tree established based on rough set, it can be found that there is no way to determine whether the leaf node (0, 1) is 0 or 1, so there is inconsistency, which requires variable precision rough set to further improve.

Assuming that the error threshold is taken as, then we calculate the clear range of the variable precision of the decision attribute relative to each condition attribute.

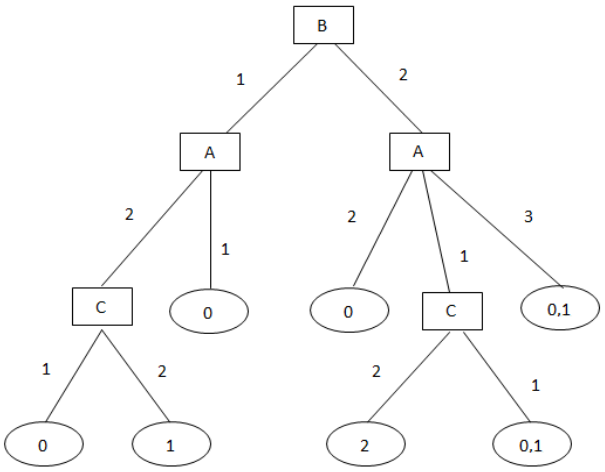

Figure 5. Decision tree based on rough set theory

$$card\left(Exp_{A\beta}\left(D^*\right)\right) = card\left(Exp_{B\beta}\left(D^*\right)\right) = 0,$$
$$card\left(Exp_{C\beta}\left(D^*\right)\right) = card\left(\bigcup_{D_i \in D^*} \underline{C}_\beta\left(D_i\right)\right)$$
$$= card\left(C_1\right) = 6$$

The variable precision of the conditional attribute C is the largest relative to the other two, so C is selected as the root node. Since C is divided into two parts, there are two forks from the root node. Same as above, after calculation, the following decision tree can be obtained as shown in Figure 6 [16].

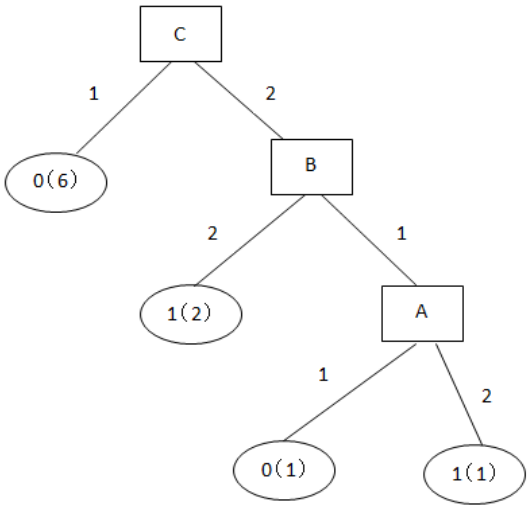

Figure 6. Decision tree generated based on variable precision rough set theory

The decision tree generated based on the variable precision rough set theory is easier than the decision tree generated based on the rough set theory. The decision tree generation method based on the rough set theory completely matches the original data when generating the decision tree. This kind of decision the classification accuracy of the tree for the original data is very high, but the excessive matching of the original data will inevitably reduce the generalization ability of the future data. The variable precision rough set theory

introduces an error threshold β when selecting nodes, which increases the clear area the size reduces the impact of a few special instances, and allows the generated decision tree to have a certain error on the original data. While avoiding excessive matching of the original data, it also enhances the generalization ability of the decision tree for future data. This is in practical applications Middle is very important [17].

6 Conclusion

Based on the rough set theory, this paper proposes the variable precision clear area and the variable precision non-clear area, which makes up for the ID3 algorithm, the C4.5 algorithm and the decision tree generated based on the rough set theory. The three algorithms cannot deal with noise well. The problem of data. Through the comparison of the generated decision tree, the decision tree generated based on the variable precision theory can reduce the misclassification, thereby reducing the size of the decision tree, and then improving the generalization ability of the decision tree. On the whole, based on entropy the decision tree generation method can only handle data with less data and no noise, while the decision tree generation method based on rough set theory pays more attention to the size of the clear area, and can handle some noisy data and some indivisible data. For rough sets, the clear zone is sometimes not "clear", but the emergence of the variable precision clear zone makes it easier to find the clear zone, so that the data can be processed more simply. However, the variable precision rough set theory still has some problems to be solved. Although we set the error threshold is not ruled out, but it is not ruled out that there will be no noise and error in that range. On the other hand, practical applications have stricter requirements on the error threshold, but the range of the error threshold is more difficult to determine, so the next work should Determine the range of the error threshold according to the actual application, so as to establish a more accurate decision. The better you look, the better we all look. Thank you for your cooperation and contribution. We are looking forward to seeing you at the conference.

Acknowledgements

This work was supported by Supported by Natural Science Foundation of Shaanxi Province of China (2021JM-344) and Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data (No. IPBED7).

References

[1] W. Wei, X. Fan, H. Song, X. Fan, J. Yang, Imperfect information dynamic stackelberg game based resource allocation using hidden Markov for cloud computing, *IEEE Transactions on Services Computing*, Vol. 11, No. 1, pp. 78-89, January-February, 2018.

[2] X.-H. Wu, The Application of Attribute Reduction Based on Rough Set in Higher Education Assessment, *Journal of Langfang Teachers College (Natural Science Edition)*, Vol. 13, No. 3, pp. 33-36, June, 2013.

[3] F. Wang, Z. Wang, Y. Zeng, Improved Algorithm of Decision Trees Construction Based on Variable Precision Rough Set, *Computer and Digital Engineering*, Vol. 41, No. 3, pp. 337-339, March, 2013.

[4] S. Liu, C. Guo, F. Al-Turjman, K. Muhammad, V. H. C. de Albuquerque, Reliability of response region: A novel mechanism in visual tracking by edge computing for IIoT environments, *Mechanical systems and signal processing*, Vol. 138, pp. 106537.1-106537.15, April, 2020.

[5] W. Wei, B. Zhou, D. Połap, M. Wozniak, A regional adaptive variational PDE model for computed tomography image reconstruction, *Pattern Recognition*, Vol. 92, pp. 64-81, August, 2019.

[6] S. Liu, S. Wang, X. Liu, C.-T. Lin, Z. Lv, Fuzzy Detection aided Real-time and Robust Visual Tracking under Complex Environments, *IEEE Transactions on Fuzzy Systems*, Vol. 29, No. 1, pp. 90-102, Janaury, 2021.

[7] S. Liu, S. Wang, X. Liu, A. H. Gandomi, M. Daneshmand, K. Muhammad, V. H. C. de Albuquerque, Human Memory Update Strategy: A Multi-Layer Template Update Mechanism for Remote Visual Monitoring, *IEEE Transactions on Multimedia*, pp. 1-1, March, 2021.

[8] W. Wei, X. Xia, M. Wozniak, X. Fan, R. Damasevicius, Y. Li, Multi-sink distributed power control algorithm for Cyber-physical-systems in coal mine tunnels, *Computer Networks*, Vol. 161, pp. 210-219, October, 2019.

[9] J. Q. Quinlan, Induction of Decision Trees, *Machine Learning*, Vol. 1, No. 1, pp. 81-106, March, 1986.

[10] Y.-L. Zhang, J. Zhou, H.-X. Liu, New Decision Tree Generation Algorithm Based on Rough Set, *Computer Applications and Software*, Vol. 27, No. 6, pp. 95-97, June, 2010.

[11] L. Shi, Q. Duan, J. Zhang, M. Xiong, L. Xi, X. Ma, Decision tree ensemble learning algorithm based on rough set, *Guangxi Sciences*, Vol. 25, No. 4, pp. 423-427, August, 2018.

[12] C.-R. Ding, L.-S. Li, B.-H. Yang, Decision tree constructing algorithm based on rough set, *Computer Engineering*, No. 11, pp. 75-77, June, 2010.

[13] T. Zhang, H.-P. Pan, Formal algorithm of decision tree and its application in geographic information, *Bulletin of Surveying and Mapping*, No. 7, pp. 51-53, 2002.

[14] F.-H. Meng, Decision tree analysis of horizontal drilling, *Petroleum Drilling Techniques*, Vol. 23, No. 1, pp. 56-58, 1995.

[15] R. Rastogi, K. Shim, PUBLIC: A Decision Tree Classifier

that Integrates Building and Pruning, *Proceedings of the 24th VLDB Conference*, New York City, New York, USA, 1998, pp. 404-415.

[16] B. Liu, W. Hsu, Y. Ma, Integrating classification and association rule mining, *Proceedings of KDD-98*, New York City, New York, USA, 1998, pp. 80-86.

[17] R.-J. Lewis, An Introduction to Classification and Regression Tree (CART) Analysis, *Annual Meeting of the Society for Academic Emergency Medicine*, San Francisco, California, USA, 2000, pp. 1-14.

Biographies

Wei Wei is an associate professor of School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China. He is a senior member of IEEE, CCF. He received his Ph.D. and M.S. degrees from Xian Jiaotong University in 2011 and 2005, respectively.

Mingwei Hui is a graduate student of School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China. He is a senior member of CCF. He obtained his undergraduate degree from Xi'an Shiyou University in 2020.

Beibei Zhang received the M.S. and Ph.D. degrees from the Xian Jiaotong University, respectively. He is currently an assistant professor at Xian University of Technology. His research interests include wireless networks and wireless sensor networks applications, mobile computing, distributed computing, and pervasive computing.

Rafał Scherer received his MSc degree in computer science from the Czestochowa University of Technology, Poland, in 1997 and his PhD in 2002 from the same university. Nowadays, he is an associate professor at Czestochowa University of Technology.

Robertas Damaševičius graduated at the Faculty of Informatics, Kaunas University of Technology (KTU) in Kaunas, Lithuania in 1999, where he received a B.Sc. degree in Informatics. He finished his M.Sc. studies in 2001 (cum laude), and he defended his Ph.D.