Computing and Informatics, Vol. 26, 2007, 663–677

VPRSM BASED DECISION TREE CLASSIFIER

Jin-Mao Wei, Ming-Yang Wang, Jun-Ping You

Institute of Computational Intelligence
Key Laboratory for Applied Statistics of MOE, Northeast Normal University
Changchun, Jilin 130024, P. R. China
e-mail: weijm374@nenu.edu.cn

Shu-Qin Wang

School of Mathematics and Statistics
Key Laboratory for Applied Statistics of MOE, Northeast Normal University
Changchun, Jilin 130024, P. R. China

Da-You Liu

Open Symbol Computation and Knowledge Engineering
Laboratory of State Education, Jilin University
Changchun, Jilin 130024, P. R. China

Manuscript received 17 May 2005; revised 22 May 2007
Communicated by Peter Vojtáš

Abstract. A new approach for inducing decision trees is proposed based on the Variable Precision Rough Set Model. From the rough set theory point of view, in the process of inducing decision trees with evaluations of candidate attributes, some methods based on purity measurements, such as information entropy based methods, emphasize the effect of class distribution. The more unbalanced the class distribution is, the more favorable it is. The rough set based approaches emphasize the effect of certainty. The more certain it is, the better. The criterion for node selection in the new method is based on the measurement of the variable precision explicit regions corresponding to candidate attributes. We compared the presented approach with C4.5 on some data sets from the UCI machine learning repository, which instantiates the feasibility of the proposed method.

Keywords: Rough sets, variable precision explicit region, variable precision implicit region, and decision tree

1 INTRODUCTION

Rough set theory, introduced by Polish mathematician Pawlak in 1982, is a new mathematical tool to deal with vagueness and uncertainty [1]. It has been widely studied and applied in many fields such as machine learning, data mining and pattern recognition [1, 2, 3, 4], etc. In [5], the authors proposed a new approach based on rough set theory for inducing decision trees. The approach was testified to be a simple and feasible way for inducing decision trees. However, the proposed approach also has its limitations in applications. It works well only in accurate classification where objects are strictly classified according to equivalence classes. Hence the induced classifiers are sensitive to noisy data and lack the ability to tolerate possible noises in real world data sets. This is an important issue to tackle in applications [6, 7, 8, 9]. Furthermore, the rough set based approach tends to partition instances too excessively, and thus constructs a large decision tree and reveal trivial details in the data. In consequence, some leaf nodes' comprehensive ability will decrease because they contain too few instances. This results in overfitting when inducing a classifier enforces the pruning of the constructed decision tree to enhance the generalization ability. Variable Precision Rough Set Model (VPRSM) [10, 11, 12, 13] is the expansion to the basic rough set model, which allows some misclassification when classifying instances on the basis of rough set model.

This paper proposes two concepts of variable precision explicit and implicit regions based on VPRSM, then ameliorates the rough set based approach for inducing decision trees by utilizing the new concepts. We also discuss the differences between the rough set based methods and the methods based on purity measurements. The new approach has the advantage of allowing misclassification when partitioning instances into explicit regions. This will consequently enhance the generalization ability of the induced decision trees, increase the ability of predicting and classifying future data. Simultaneously, VPRSM based approach for inducing decision trees can also do well when the explicit regions of all candidate attributes have the same size.

2 ROUGH SET BASED DECISION TREES

Detailed description of some basic concepts such as equivalence relation, equivalence class, upper approximation, lower approximation, boundary, negative region in the rough set theory can be found in the literature [1, 4]. Given a knowledge representation system:

$$S = (U, Q, V, \rho)$$

U is a certain set of objects called the universe. Q denotes the set of attributes. It is usually divided into two subsets C and D, which denote the set of condition attributes and the set of decision attributes, respectively.

$\rho : U \times Q \to V$ is an information function, where $V = \bigcup_{a \in Q} V_a$ and V_a is the domain of attribute $a \in Q$.

For any subset G of C or D, an equivalence relation $\widetilde{G}$ on U can be defined such that a partition of U induced by it can be obtained. Denote the partition as $G^* = \{X_1, X_2, \ldots, X_n\}$, where X_i is an equivalence class of $\widetilde{G}$. To be simple, we usually call $(U, \widetilde{G})$ an approximation space.

Definition 1. Let $A \subseteq C$, $B \subseteq D$. $A^* = \{X_1, X_2, \ldots, X_n\}$ and $B^* = \{Y_1, Y_2, \ldots, Y_m\}$ denote the partitions of U induced by equivalence relation $\widetilde{A}$ and $\widetilde{B}$, respectively, where equivalence relations $\widetilde{A}$ and $\widetilde{B}$ are induced from A and B. The explicit region is defined as:

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

where $\underline{A}(Y_i)$ denotes the lower approximation of Y_i with respect to $\widetilde{A}$.

Definition 2. Let $A \subseteq C$, $B \subseteq D$. $A^* = \{X_1, X_2, \ldots, X_n\}$ and $B^* = \{Y_1, Y_2, \ldots, Y_m\}$ denote the partitions of U induced by equivalence relation $\widetilde{A}$ and $\widetilde{B}$, respectively, where equivalence relations $\widetilde{A}$ and $\widetilde{B}$ are induced from A and B. The implicit region is defined as:

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

where $\overline{A}(Y_i)$ denotes the upper approximation of Y_i with respect to $\widetilde{A}$.

Obviously, we have:

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

The initial idea of the rough set based approach to selection of decision tree nodes lies in the following process: From an original data set to the final decision tree, the knowledge about the system tends to gradually become explicit. Consequently one will gradually learn much about the system. Hence in the process of constructing a decision tree, from the root to the leaves of the decision tree one condition attribute will be picked as the node of the decision tree, if the explicit region corresponding to it is greater than that of all other available condition attributes and thus one can learn more about the system depicted by the data. In the approach, when we evaluate a possible condition attribute, the data set is partitioned into two parts according to the values of the attribute: one part is the explicit region; the other is the implicit region. After partition we can obtain the sizes of these two parts. Similarly, we can obtain the explicit and implicit regions and their sizes corresponding to all other condition attributes. We compare the sizes of the explicit or implicit regions of all condition attributes. We choose the attribute with the greatest explicit region or the least implicit region as the branch node.

For example, we construct a decision tree for a given data set. In the partially constructed tree, each node corresponds to a condition attribute. The path from the root node to node nl and to a data subset Dl is a partial branch of the final tree. Under branch nl there is a data subset Dl to be partitioned. Each attribute within the available condition attributes is evaluated for partitioning Dl by computing its explicit region. For an instance, attribute A is evaluated, and data subset Dl is partitioned into two parts, Exp and Imp, which denote the explicit region and implicit regions, respectively. The path from the root node towards Exp implies that when the corresponding conditions are satisfied, the classification of all tuples in Exp is explicit, or a unique class label will be assigned to this leaf node unambiguously. From the definition, in an Exp, there may be more than two subsets that have different class labels; however, each subset corresponding to one possible value of the evaluated condition attribute can be assigned a unique class label. From the root node to Imp, the class labels of the tuples in Imp are different. It is apparent that the Exp of the greatest size is preferred and hence the corresponding attribute should be chosen for partitioning the data subset Dl.

In real applications, however, it is always the case that the data to be handled contains noises. It is not difficult to find out that even a small perturbation may totally reverse the result of choice of branch attribute. For example, when evaluating a condition attribute, one data subset after partition has 100 instances in it, and all instances have the same class label. The size of the explicit region corresponding to this attribute will then be at least 100. If we add a small perturbation to the data set by changing the class label of one instance within the 100 instances, all the 100 instances will be partitioned into the implicit region, and thus will reduce the size of the explicit region by 100. Apparently this may consequently change the choice of branch attribute when comparing to the other attributes. From the above discussion, the approach tends to classify instances excessively due to accurate classification in the rough set theory and cannot avoid some negative effects brought by some minority instances. So the rough set based approach for inducing decision trees tends to create large decision trees, and thus the induced decision trees need further elaborate pruning.

3 VPRSM BASED DECISION TREES

3.1 Basic Concepts

VPRSM extends the rough set model by allowing for some degree of misclassification in the largely correct classification.

Definition 3 (13)**.** Assume U denotes the universe to be learned. X and Y denote the non-empty subsets of U. Let:

$$c(X,Y) = \begin{cases} 1 - \frac{|X \bigcap Y|}{|X|}, & |X| > 0 \\ 0, & |X| = 0 \end{cases} \tag{3}$$

where $|X|$ is the cardinality of X and $c(X, Y)$ is the relative classification error of the set X with respect to set Y. That is to say, if all elements of the set X were partitioned into set Y then in $c(X, Y) \times 100\%$ of the cases we would make a classification error. Generally, the admissible classification error β must be within the range $0 \leq \beta < 0.5$.

Suppose $(U, \widetilde{R})$ is an approximation space, $R^* = \{E_1, E_2, \ldots, E_n\}$ denotes the set containing the equivalence classes in $\widetilde{R}$.

For any subset $X \subseteq U$, the β lower approximation of X with respect to $\widetilde{R}$ is defined as:

$$\underline{R}_\beta(X) = \bigcup\{E_i \in R^* | C(E_i, X) \leq \beta\}. \tag{4}$$

The β upper approximation of X with respect to $\widetilde{R}$ is defined as:

$$\overline{R}_\beta(X) = \bigcup\{E_i \in R^* | C(E_i, X) < 1 - \beta\}. \tag{5}$$

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

$$BNR_\beta(X) = \bigcup\{E_i \in R^* | \beta < C(E_i, X) < 1 - \beta\}. \tag{6}$$

Comparing VPRSM with the initial rough set model, we can easily obtain that VPRSM will turn to be the rough set model when $\beta = 0$.

3.2 VPRSM Based Approach for Inducing Decision Trees

First, we give two concepts of variable precision explicit and implicit regions to substitute those of explicit and implicit regions in literature [5] as the criteria for selecting attributes in the process of inducing decision trees.

Definition 4. Let $A \subseteq C$, $B \subseteq D$. $A^* = \{X_1, X_2, \ldots, X_n\}$ and $B^* = \{Y_1, Y_2, \ldots, Y_m\}$ denote the partitions of U induced by equivalence relations $\widetilde{A}$ and $\widetilde{B}$, respectively, where equivalence relations $\widetilde{A}$ and $\widetilde{B}$ are induced from A and B. The variable precision explicit region is defined as

$$Exp_{A\beta}(B^*) = \bigcup_{Y_i \in B^*} \underline{A}_\beta(Y_i) \tag{7}$$

where $\underline{A}_\beta(Y_i)$ is the β lower approximation of Y_i with respect to $\widetilde{A}$.

After the explicit region is enlarged, the implicit region turns to be as follows.

Definition 5. Let $A \subseteq C$, $B \subseteq D$. $A^* = \{X_1, X_2, \ldots, X_n\}$ and $B^* = \{Y_1, Y_2, \ldots, Y_m\}$ denote the partitions of U induced by equivalence relations $\widetilde{A}$ and $\widetilde{B}$, respectively, where equivalence relations $\widetilde{A}$ and $\widetilde{B}$ are induced from A and B. Then the variable precision implicit region is defined as:

$$Imp_{A\beta}(B^*) = \bigcup_{Y_i \in B^*} (\overline{A}_\beta(Y_i) - \underline{A}_\beta(Y_i)) \tag{8}$$

where $\underline{A}_\beta(Y_i)$ is the β lower approximation of Y_i and $\overline{A}_\beta(Y_i)$ is the β upper approximation with respect to $\widetilde{A}$.

Obviously, the concept of variable precision explicit region is the expansion of the explicit region in classification. Classification error β has weakened the rigid requirements to lower approximation in the rough set model to ensure that lower approximation has some tolerance with inconsistent data, which will surely expand the range of explicit region to contain more instances.

It should be noticed that there is a difference between the definitions of explicit and implicit regions and the definitions of positive and negative regions defined in [4] and some discussions in [13]. In order to avoid misunderstanding, we chose to use explicit and implicit regions to distinguish two cases when we deal with data sets. One case is that we can determine the class label of the data under some conditions, whereas the other case is that the class label of the data cannot be assigned unambiguously. The data may belong to one class while it may belong to other possible classes.

In the process of inducing a decision tree based on variable precision explicit region, the inducing approach will select the attribute with the largest size of variable precision explicit region as the node of current branch. The above discussion suggests that complexity of the tree will be reduced, and consequently the tree's generalization ability will be enhanced. Simultaneously, it is also valid to deal with the situation when the explicit regions of all candidate attributes have the same size, which remedies the limitation of the rough set based approach for inducing decision trees.

4 AN EXAMPLE

This section gives a simple example explaining how to construct a decision tree based on VPRSM.

Table 1 is selected from literature [14] that contains 24 records, every one of which corresponds to five attributes. The first four records are condition attributes and the last one "Class" is decision attribute. For simplification, the attribute "Outlook" "Temperature" "Humidity" "Windy" are rewritten as "A" "B" "C" "D" and the decision attribute as "E". Assume $\beta = 0.2$.

We evaluate each of the four condition attributes. The partition with respect to equivalence relation $\widetilde{A}$ is:

$$A^* = \{A_1, A_2, A_3\}$$

$$= \{\{1,2,3,13,14,15,16,19,20\},\{4,5,11,12,21,22,23\},\{6,7,8,9,10,17,18,24\}\}$$

The partition with respect to equivalence relation $\widetilde{B}$ is:

$$B^* = \{B_1, B_2, B_3\}$$

$$= \{\{1,2,3,4,5,8,10,23\},\{6,7,13,14,17,18,19,20,21,22,24\},\{9,11,12,15,16\}\}$$

The partition with respect to equivalence relation $\widetilde{C}$ is:

$$C^* = \{C_1, C_2\}$$

$$= \{\{1, 2, 3, 4, 5, 6, 7, 13, 14, 21, 22, 24\}, \{8, 9, 10, 11, 12, 15, 16, 17, 18, 19, 20, 23\}\}$$

The partition with respect to equivalence relation $\widetilde{D}$ is:

$$D^* = \{D_1, D_2, D_3\}$$

$$= \{\{1, 4, 6, 8, 13, 17, 23\}, \{2, 10, 11, 20, 21, 24\}, \{3, 5, 7, 9, 12, 14, 16, 18, 19, 22\}\}$$

The partition with respect to equivalence relation $\widetilde{E}$ is:

$$E^* = \{E_1, E_2\}$$

$$= \{\{1, 2, 3, 6, 7, 9, 10, 13, 14, 17, 18, 24\}, \{4, 5, 8, 11, 12, 15, 16, 19, 20, 21, 22, 23\}\}$$

The sizes of the variable precision explicit regions with respect to the four attributes are calculated as follows:

$$|Exp_{A\beta}(E^*)| = |\bigcup_{E_i \in E^*} \underline{A}_\beta(E_i)| = |A_2 \bigcup A_3| = 15$$

$$|Exp_{B\beta}(E^*)| = |\bigcup_{E_i \in E^*} \underline{B}_\beta(E_i)| = |B_3| = 5$$

$$|Exp_{C\beta}(E^*)| = |\bigcup_{E_i \in E^*} \underline{C}_\beta(E_i)| = 0$$

$$|Exp_{D\beta}(E^*)| = |\bigcup_{E_i \in E^*} \underline{D}_\beta(E_i)| = 0.$$

It is apparent that the size of the variable precision explicit region with respect to attribute A, i.e. attribute “Outlook”, is the greatest comparing to the size of all the other three attributes. Therefore, attribute “Outlook” is chosen as the root node of the decision tree. Consequently we partition the whole data set into three subsets, which correspond to the three branches of the decision tree, see a) in Figure 1.

The “Sunny” branch has seven tuples, each tuple has class label of “P” that means ‘Play”. This data subset needs no further partition, and of course “P” is assigned as the class label for this leaf node. The “Rain” branch has eight tuples in total, one tuple, No. 8, takes the class label “P”, and the other seven tuples take the class label “N”. However, we do not further partition the subset either, and assign the class label “N” to this leaf node for $c(A_2, E_2) = 0.125 \leq \beta = 0.2$; or we say that A_2 is partitioned into the variable precision explicit region with respect to attribute “Outlook” (see the above calculation of the size of the variable precision explicit region with respect to this attribute). Then we only need to partition the subset corresponding to branch “Overcast”.

We evaluate each of the rest condition attributes.

No.	Outlook	Temperature	Humidity	Windy	Class
1	Overcast	Hot	High	Not	N
2	Overcast	Hot	High	Very	N
3	Overcast	Hot	High	Medium	N
4	Sunny	Hot	High	Not	P
5	Sunny	Hot	High	Medium	P
6	Rain	Mild	High	Not	N
7	Rain	Mild	High	Medium	N
8	Rain	Hot	Normal	Not	P
9	Rain	Cool	Normal	Medium	N
10	Rain	Hot	Normal	Very	N
11	Sunny	Cool	Normal	Very	P
12	Sunny	Cool	Normal	Medium	P
13	Overcast	Mild	High	Not	N
14	Overcast	Mild	High	Medium	N
15	Overcast	Cool	Normal	Not	P
16	Overcast	Cool	Normal	Medium	P
17	Rain	Mild	Normal	Not	N
18	Rain	Mild	Normal	Medium	N
19	Overcast	Mild	Normal	Medium	P
20	Overcast	Mild	Normal	Very	P
21	Sunny	Mild	High	Very	P
22	Sunny	Mild	High	Medium	P
23	Sunny	Hot	Normal	Not	P
24	Rain	Mild	High	Very	N

Table 1. Data Set

The partition with respect to equivalence relation $\widetilde{B}$ is:

$$B^* = \{B_1, B_2, B_3\} = \{\{1, 2, 3\}, \{13, 14, 19, 20\}, \{15, 16\}\}.$$

It should be noticed that here $U = \{1, 2, 3, 13, 14, 15, 16, 19, 20\}$, i.e. the "Overcast" branch, and it is the same hereunder.

The partition with respect to equivalence relation $\widetilde{C}$ is:

$$C^* = \{C_1, C_2\} = \{\{1, 2, 3, 13, 14\}, \{15, 16, 19, 20\}\}.$$

The partition with respect to equivalence relation $\widetilde{D}$ is:

$$D^* = \{D_1, D_2, D_3\} = \{\{1, 13, 15\}, \{2, 20\}, \{3, 14, 16, 19\}\}.$$

The partition with respect to equivalence relation $\widetilde{E}$ is:

$$E^* = \{E_1, E_2\} = \{\{1, 2, 3, 13, 14\}, \{15, 16, 19, 20\}.$$

The sizes of the variable precision explicit regions with respect to the three attributes are calculated as follows:

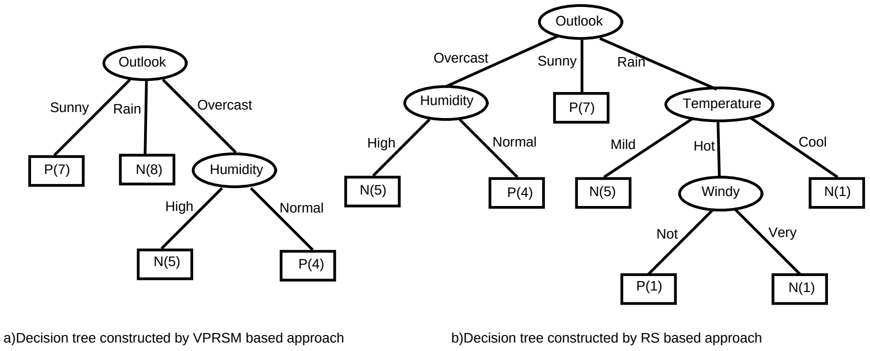
$$|Exp_{B\beta}(E^*)| = |\bigcup_{E_i \in E^*} \underline{B}_\beta(E_i)| = |B_1 \bigcup B_3| = 5$$

$$|Exp_{C\beta}(E^*)| = |\bigcup_{E_i \in E^*} \underline{C}_\beta(E_i)| = |C_1 \bigcup C_2| = 9$$

$$|Exp_{D\beta}(E^*)| = |\bigcup_{E_i \in E^*} \underline{D}_\beta(E_i)| = 0.$$

It is apparent that the size of the variable precision explicit region with respect to attribute C, i.e. attribute "Humidity", is the greatest comparing to the size of the other two attributes. Therefore, attribute "Humidity" is chosen as the node of this branch, and we need no further partition for the growing branches. "N" and "P" are assigned to the branch "High" and "Normal", respectively.

The final decision tree constructed by VPRSM($\beta = 0.2$) is shown as a) in Figure 1.



a)Decision tree constructed by VPRSM based approach

b)Decision tree constructed by RS based approach

Fig. 1. Comparison between the decision trees induced by VPRSM and RS based approaches

The decision tree constructed by the rough set based approach is shown as b) in Figure 1. In the figure, the number in parenthesis denotes the number of instances reaching each leaf node. Comparing the decision trees in Figure 1, we can easily get that the tree constructed by VPRSM is briefer than that constructed by the rough set based approach.

For further comparison with the methods based on purity measurements, we take the fundamental entropy based method for example. The initial idea of the entropy based method is to observe the information gain ($Info\_Gain$) when a data set is split by the possible values of condition attributes. $Info\_Gain$ is defined [8, 15] as:

$$Info\_Gain(A, U) = Info(U) - Info(A, U)$$

where U is the set of objects. A is a condition attribute under evaluation. If the set U of objects is partitioned into disjoint exhaustive classes $\{Y_1, Y_2, \ldots, Y_k\}$ on the basis of the value of decision attribute, then the information needed to identify the class of an element of U is:

$$Info(U) = I(P) = -\sum_{i=1}^{k} p_i log(p_i)$$

P is the probability distribution of the partition $\{Y_1, Y_2, \ldots, Y_k\}$

$$P = (\frac{|Y_1|}{|U|}, \frac{|Y_2|}{|U|}, \ldots, \frac{|Y_k|}{|U|}), p_i = \frac{|Y_i|}{|U|}.$$

If a condition attribute has the greatest $Info\_Gain$, this attribute will be chosen to split the data set. When comparing the entropy based method to the rough set based approach, we find that the former pays attention to the distribution of classes before and after data partition, whereas the rough set based approach pays attention to how much certainty one will confirmatively observe after data partition. When a condition attribute is evaluated, the data subset(Dl for example) is split into two parts, i.e. Exp and Imp. In the fundamental entropy based method, the $Info\_gain$ can be calculated as:

$$Info\_Gain(A, Dl) = Info(Dl) - Info(A, Dl).$$

It can be further calculated as:

$$Info\_Gain(A, Dl) = Info(Dl) - Info(A, Exp) - Info(A, Imp).$$

According to the definition of information entropy, we have $Info(A, Exp) = 0$. Hence the information gain is:

$$Info\_Gain(A, Dl) = Info(Dl) - Info(A, Imp)$$

This implies that Exp doesn't make contribution to the information gain, or at least, Exp does not make contribution to information gain directly. In fact, in practice, if $Info(A, Imp)$ is the smallest, attribute A will be chosen.

From this point, the fundamental entropy based method pays attention to the distribution of classes of Imp. In contrast to the fundamental entropy based method, the rough set based approach pays attention to the size of Exp. If the explicit region $Exp_A(B^*) = \bigcup_{Y_i \in B^*} \underline{A}(Y_i)$ of attribute A, for example, is the greatest, attribute A will be chosen at last.

5 COMPARISON ON SOME REAL DATA SETS

Comparison between the rudimental rough set based approach and the fundamental information theory based method can be found in [5]. In this paper, we compare the VPRSM based approach (to be simple, we note it as Ver4) with the popular algorithm C4.5. We utilize some data sets from the UCI machine learning repository, which are accessible and suitable for constructing decision trees. We compare the decision trees constructed by Ver4 with that obtained by C4.5. The names of all data sets and the results are shown in Table 2. Classification accuracy before and after pruning is shown in Figures 2 and 3.

We use 17 kinds of data sets from the UCI machine learning repository. In the table, "β" indicates the threshold of classification error used in Ver4 when dealing with different data sets, "data size" indicates the sizes of the data sets, "size" indicates the induced tree size, "errors" indicates the learning error of the induced decision tree. The value out of parentheses is the number of tuples that were misclassified by the induced tree. The value within parentheses is the rate of misclassification. It is computed by dividing the number of misclassified tuples by the number of total tuples to be learned. For the breast-cancer data set, we dealt with the attributes in two manners. One is to treat them as continuous attributes, the other as discontinuous attributes.

As to the rough set based approach Ver4, the possible condition attributes were evaluated by computing their corresponding explicit regions. An attribute was chosen as the branch node if its explicit region was the largest comparing to the other candidate attributes. When the explicit regions of all of the available condition attributes were identical, the first processed attribute was chosen as the node of the current branch. In C4.5, all possible attributes were evaluated by calculating their corresponding $Info\_Gain$.

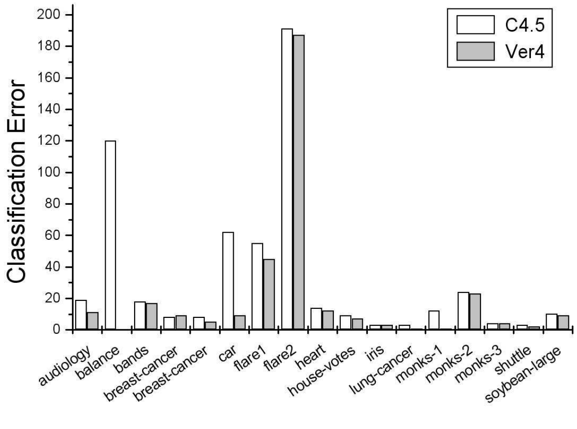


Fig. 2. Comparison between C4.5 and Ver4 before pruning

We can see from Table 2, Figure 2 and Figure 3, that Ver4 shows to be more competent than C4.5, especially before pruning. Figures 2 and 3 reveal that the classification accuracy of Ver4 descends comparing to that before pruning, but the

Prog.	data set	data size	β	Before Pruning		After Pruning	
				size	errors	size	errors
C4.5	audiology	200		434	18 (9.0)	283	32 (16.0)
Ver4	audiology		0.085	691	9 (4.5)	364	28 (14.0)
C4.5	balance	625		111	120 (19.2)	41	156 (25.0)
Ver4	balance		0	226	0	51	150 (24)
C4.5	bands	540		217	18 (3.3)	135	25 (4.6)
Ver4	bands		0.06	304	17 (3.1)	156	34 (6.3)
C4.5	breast-cancer (Cont)	699		45	8 (1.1)	31	11 (1.6)
Ver4	breast-cancer		0	55	9 (1.3)	29	14 (2)
C4.5	breast-cancer (discr)	699		151	8 (1.1)	31	29 (4.1)
Ver4	breast-cancer		0.005	171	5 (0.7)	31	29 (4.1)
C4.5	car	1728		186	62 (3.6)	182	64 (3.7)
Ver4	car		0	442	9 (0.5)	190	92 (5.3)
C4.5	flare1	323		74	55 (17)	36	64 (19.8)
Ver4	flare1		0.19	137	45 (13.9)	43	62 (19.2)
C4.5	flare2	1066		179	191 (17.9)	48	235 (22)
Ver4	flare2		0.33	198	187 (17.5)	84	220 (20.6)
C4.5	heart	270		62	14 (5.2)	43	19 (7.0)
Ver4	heart		0.09	66	12 (4.4)	64	12 (4.4)
C4.5	house-votes	435		37	9 (2.1)	11	12 (2.8)
Ver4	house-votes		0.014	45	7 (1.6)	13	12 (2.8)
C4.5	iris	150		9	3 (2.0)	9	3 (2.0)
Ver4	iris		0.05	9	3 (2.0)	9	3 (2.0)
C4.5	lung-cancer	32		29	3 (9.4)	25	4 (12.5)
Ver4	lung-cancer		0.15	25	1 (3.1)	25	1 (3.1)
C4.5	monks-1	124		45	12 (9.7)	19	20 (16.1)
Ver4	monks-1		0.1	38	1 (0.8)	38	1 (0.8)
C4.5	monks-2	169		79	24 (14.2)	33	40 (23.7)
Ver4	monks-2		0	84	24 (14.2)	35	41 (24.3)
C4.5	monks-3	122		25	4 (3.3)	12	8 (6.6)
Ver4	monks-3		0	27	4 (3.3)	12	8 (6.6)
C4.5	shuttle	15		9	3 (20.0)	1	6 (40.0)
Ver4	shuttle		0.25	9	2 (13.3)	3	5 (33.3)
C4.5	soybean-large	307		166	10 (3.3)	104	15 (4.9)
Ver4	soybean-large		0.03	307	9 (2.9)	154	14 (4.6)

Table 2. Comparison of the rough set based approach and C4.5

sizes of the decision trees reduce for most data sets. It is a consequent result in the stage of pruning. Reduction of the sizes and hence of the description length of decision trees aims at enhancing the generalization abilities of the classifiers for predicting unseen data. In the experiment, there are three problems that have data sets for test, namely the monk problem, the soybean problem and the audiology problem. The results for these problems are listed in Table 3. From the table, for the audiology problem, the size of the decision tree constructed by Ver4 is bigger

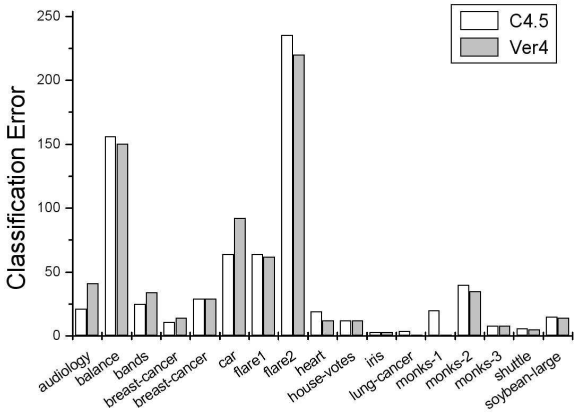


Fig. 3. Comparison between C4.5 and Ver4 after pruning

than that by C4.5, the prediction error being the same. For the monk problems, Ver4 is not worse than C4.5. For the soybean problem, C4.5 shows to be better than Ver4 in both the size and prediction accuracy of the decision trees. In the experiment, we also used some other data sets, such as soybean-small, the results of which are the same for the two algorithms. Hence we do not report all of the results in the paper. In the variable precision rough set based approach, the accuracy and the size of decision trees can be adjusted by choosing suitable thresholds of classification error for the problems. It should be mentioned that a prepruning effect exits in Ver4 comparing with the basic rough set based approach. If the threshold of classification error is too large, the decision tree will stop growing fast. Such pruning problems will be further discussed in our other works.

In the process of constructing decision trees, the same methods as those used in C4.5 for pruning, for dealing with the attributes with missing values and for discretizing continuous attributes are utilized for parallel comparison in Ver4.

6 CONCLUSIONS

Two concepts of variable precision explicit and implicit regions are proposed based on Variable Precision Rough Set Model. A decision tree inducing approach using the concepts for selecting attributes as the current nodes is given. The approach allows some misclassification when partitioning instances into explicit regions in the process of inducing decision trees, which will reduce the scales of the decision trees and thus enhance the generalization ability of the constructed decision trees. Experiments on some data sets from the UCI machine learning repository show that the presented approach is feasible for constructing decision trees. The problems

Prog.	data set	data size	β	Before Pruning		After Pruning	
				size	errors	size	errors
C4.5	audiology	200		434	18(9.0)	283	32(16.0)
	audiology.test	26		434	13(50.0)	283	13(50.0)
Ver4	audiology	200	0.085	691	9(4.5)	364	28(14.0)
	audiology.test	26		691	13(50.0)	364	13(50.0)
C4.5	monks-1	124		45	12(9.7)	19	20(16.1)
	monks1.test	432		45	101(23.4)	19	105(24.3)
Ver4	monks-1	124	0.1	38	1(0.8)	38	1(0.8)
	monks1.test	432		38	24(5.6)	38	24(5.6)
C4.5	monks-2	169		79	24(14.2)	33	40(23.7)
	monks2.test	432		79	150(34.7)	33	151(35.0)
Ver4	monks-2	169	0	84	24(14.2)	35	41(24.3)
	monks2.test	432		84	153(35.4)	35	150(34.7)
C4.5	monks-3	122		25	4(3.3)	12	8(6.6)
	monks3.test	432		25	32(7.4)	12	12(2.8)
Ver4	monks-3	122	0	27	4(3.3)	12	8(6.6)
	monks3.test	432		27	16(3.7)	12	12(2.8)
C4.5	soybean-large	307		166	10(3.3)	104	15(4.9)
	soybean.test	376		166	54(14.4)	104	50(13.3)
Ver4	soybean-large	307	0.03	307	9(2.9)	154	14(4.6)
	soybean.test	376		307	83(22.1)	154	63(16.8)

Table 3. Comparison of prediction accuracy

about pruning, such as how to enhance the performance after pruning, need further investigation.

Acknowledgement

This work is supported by the Science Foundation of Jilin Province, Grant 20040529, and the Science foundation for Young Scholars of Northeast Normal University, Grant 111420000. We used some data sets from the UCI Machine Learning Data Repository (available at `http://www.ics.uci.edu/~mlearn/MLSummary.html`). This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

REFERENCES

[1] Pawlak, Z.: Rough sets. International Journal of Computer and Information Science, Vol. 11, 1982, pp. 341–356.

[2] GrZymala-Busse, J. W.—Ziarko, W.: Data Mining and Rough Set Theory. Communications of the ACM, Vol. 43, 2000, No. 4, pp. 108–109.

[3] Pawlak, Z.: Rough Set Approach to Multi-Attribute Decision Analysis. European Journal of Operational Research, Vol. 72, 1994, No. 3, pp. 443–459.

[4] Pawlak, Z.—Wang, S. K. M.—Ziarko, W.: Rough Sets: Probabilistic Versus Deterministic Approach. Int. J. Man-Machine Studies, Vol. 29-1, 1988, pp. 81–95.

[5] Wei, J.: Rough Set Based Approach to Selection of Node. International Journal of Computational Cognition, Vol. 1, 2003, No. 2, pp. 25–40.

[6] Mingers, J.: An Empirical Comparison of Pruning Methods for Decision-Tree Induction. Machine Learning, Vol. 4, 1989, No. 2, pp. 319–342.

[7] Quinlan, J. R.—Rivest, R.: Inferring Decision Trees Using the Minimum Description Length Principle. Information and Computation, Vol. 80, 1989, No. 3, pp. 227–248.

[8] Quinlan J. R.: Introduction of Decision Trees. Machine Learning, Vol. 3, 1986, pp. 81–106.

[9] Ras, Z. W.—Zemankova, M.: Imprecise Concept Learning within a Growing Language. Proceedings of the Sixth International Workshop on Machine Learning 1989, Ithaca, New York, United States, 1989, pp. 314–319.

[10] Jian, L.—Da, Q.—Chen, W.: Variable Precision Rough Set and a Fuzzy Measure of Knowledge Based on Variable Precision Rough Set. Journal of Southeast University (English Edition), Vol. 18, 2002, No. 4, pp. 351–355.

[11] Kryszkiewicz, M.: Maintenance of Reducts in the Variable Precision Rough Set Model. In 1995 ACM Computer Science Conference (CSS '95), 1995, pp. 355–372.

[12] Ziarko, W.: Probabilistic Decision Tables in the Variable Precision Rough Set Model. Computational Intelligence, Vol. 17, 2001, No. 3, 2001, pp. 593–603.

[13] Ziarko, W.: Variable Precision Rough Set Model. Journal of Computer and System Sciences, Vol. 46, 1993, No. 1, pp. 39–59.

[14] Michalski, R. S.—Carbonell, J. G.—Mitchell, T. M.: Machine Learning – An Artificial Intelligence Approach. Springer-Verlag, printed in Germany, 1983.

[15] Quinlan, J. R.: C4.5 Programs for Machine Learning. Morgan Kaufmann, 1993.

[16] Wolberg, W. H.—Mangasarian, O. L.: Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology. Proceedings of the National Academy of Sciences, U. S. A., Vol. 87, December 1990, pp. 9193–9196.

Jin-Mao Wei received the B. Sc. degree in computer science from Jilin University of Science and Technology, Changchun, China in 1990, the M. Sc. degree in electronics from Northeast Normal University, Changchun, China in 1993, and the Ph. D. degree in automatics from East China University of Science and Technology, Shanghai, China in 2001. He is now a professor in the College of Physics and a researcher at the Key Laboratory for Applied Statistics of MOE, Northeast Normal University. His research interests include data mining, web mining and bioinformatics.