VPRSM based Improved Decision Tree Induction Approach for Handling Uncertain Data

Surekha Samsani, G.Jaya Suma

Abstract*: During past decades, several computational intelligence paradigms have been incorporated in modeling intelligent decision tree classifiers. Variable Precision Rough Set Model (VPRSM) is one of the popular theories developed to deal with uncertain data. Incorporating VPRSM concepts in Decision Tree(DT) classification handles uncertainties in the training data and results in generation of more precise decision trees. VPRSM based DT induction approach choses the attribute with most promising variable precision explicit region value as the splitting attribute. But when multiple candidate attributes are qualified with same leading variable precision explicit region value may lead to ambiguity in selecting the splitting attribute and also randomly selecting the one among them may sometimes degrades the efficiency of the induced classifier. This paper gives a solution to handle this ambiguous situation and proposes an Improved VPRSM Decision Tree Induction approach based on β-Significance (β-IDTA). The efficiency of the β-IDTA approach of DT classification is 10-fold cross validated on different bench mark publicly available medical datasets of UCI ML repository. The experimental results divulge that the proposed approach is very promisingly generating optimal comprehensible trees with improved generalization ability when compared against the trees induced by Rough Set Theory and VPRSM based approaches.

***Index Terms*: Computational Intelligence, Decision Tree Classification, Variable Precision Rough Set Model, β-Significance**

I. INTRODUCTION

During several decades, Rough Set Theory[1] is being applied successfully in the knowledge discovery process for dealing with inconsistencies, feature subset selection, discretization, and in many data mining algorithms. As noise and uncertainties in the real-world data are unavoidable, Wei [2] proposed a novel RST based DT classification by incorporating the concept of RST in DT induction algorithm. Even though RST based decision trees are classifying the data accurately, it too suffers with limitations [3]. Because, RST reveals the complete details about the data; Wei's approach is inducing more accurate decision trees and sometimes results Model Overfitting. To overcome this limitation, Ziarko[4] enhanced the RST by allowing some level of uncertainty and proposed a new theory called Variable Precision Rough Set Model(VPRSM). Ziarko's VPRSM is an enhanced model of RST which allows flexibility in dealing with uncertainties. VPRSM excuses uncertainty to some extent and this degree of allowable uncertainty is introduced as Misclassification Error(MCE) which is represented by β. As per literature, the acceptable value of β is in range from 0 to 1. In 2007 Wei[5] incorporated the Ziarko's concept of flexible MCE β in attribute selection criteria and proposed a new measure called the Variable Precision Explicit Region (VPER) which allows inducing more generalized decision trees. Wei's DT induction approach considers the attribute with promising VPER as the splitting attribute. The existing VPRSM based DT classification techniques not discussed the issue of dealing the conflicting situations such as multiple attributes with same cardinal Variable Precision Explicit Region. This paper illustrates this ambiguous context for a sample dataset and also discusses the drawbacks of the existing algorithms and gives a solution to deal with this uncertainty. This paper enhances the existing VPRSM based tree inducing algorithms to handle the ambiguous situation by selecting most significant attribute as the splitting attribute and the proposed concept (β-IDTA) is clearly explained by inducing DT for the same sample data set. The performance of the proposed β-IDTA approach is measured by conducting series of tests on public medical datasets taken from the Irvine's UC machine learning repository [6].

Revised Manuscript Received on June 05 ,2019.

Surekha Samsani, Research Scholar & Assistant Professor, University College of Engineering Kakinada(A), JNTUK, Kakinada, India.

G. Jaya Suma, Professor & HOD, Department of Information Technology, UCEV(A), JNTUK, Vizianagaram, India.

II. VPRSM BASED DECISION TREE CLASSIFICATION

This section presents the existing VPRSM approach of decision tree classification. The detailed explanations, basic concepts and definitions of VPRSM can be found in literature [4],[5].

Relative Classification Error(RCE)

Let A and B denote two subsets of the given Universe of objects U. Then the RCE of the set A with respect to set B is denoted as C(U,A,B) and is defined as[5],

$$C(U,A,B) = \begin{cases} 1 - \dfrac{|A \cap B|}{|A|}, & |A| > 0 \\ 0, & |A| = 0 \end{cases} \quad (1)$$

Where, |A| is the cardinality of A.

β-Lower Approximation

Suppose (U, E^*) is an approximation space [5], $E^* = \{E_1, E_2, \ldots E_n\}$ denotes the set containing the Equivalence Classes(EC) in $IND(U,E)$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

and β is the admissible classification error.

For any subset T ⊆ U, the β- lower approximation of T with respect to (U, E^*) is given by,

$$\underline{E}_{\beta}(U,T) = \bigcup \{E_i \in E^* | C(U, E_i, T) \leq \beta\} \quad (2)$$

Variable Precision Explicit Region (VPER) [5]

Let A ⊆ C, B ⊆ D and (U,A^*)= $\{A_1, A_2, \ldots, A_n\}$ and (U,B^*)= $\{B_1, B_2, \ldots, B_m\}$ denote the approximation spaces of U induced by equivalence relation $IND(U,A)$ and $IND(U,B)$, respectively, then the VPER of B with respect to attribute set A is defined as,

$$\beta EXP_A(U,B^*) = \bigcup_{B_i \in B^*} \underline{A}_{\beta}(U,B_i) \quad (3)$$

Where, $\underline{A}_{\beta}(U,B_i)$ is the β-Lower approximation of B_i with respect to (U,A^*).

In order to explain the existing technique, the data in Table I is taken into consideration.

TABLE I. TRAINING DATASET

U	CONDITIONAL ATTRIBUTES				D
	A_1	A_2	A_3	A_4	
1	F	N	I	F	Y
2	S	D	N	L	N
3	F	N	I	L	Y
4	S	D	N	N	N
5	F	N	I	N	Y
6	F	K	I	F	Y
7	S	N	I	F	N
8	F	K	I	N	Y
9	S	N	I	N	N
10	F	D	N	F	N
11	M	K	I	F	Y
12	F	D	N	N	N
13	M	K	I	N	Y
14	M	K	N	N	Y
15	M	N	N	F	N
16	M	K	N	N	Y
17	M	N	N	L	Y

VPER for a given β- Value: Worked Out Example

The VPER for the conditional attribute A_1 of data given in Table I is calculated as shown below.

The Equivalence Classes generated by A_1 and D are obtained as given below,

$$(U, A_1^*) = \{S_1, S_2, S_3\}$$

$$= \{\{1,3,5,6,8,10,12\}, \{2,4,7,9\}, \{11,13,14,15,16,17\}\}$$

$$(U, D^*) = \{D_1, D_2\} =$$

$$\{\{1,3,5,6,8,11,13,14,16,17\}, \{2,4,7,9,10,12,15\}\}$$

The VPER of A_1 for the flexible MCE β=0.2 can be calculated using (3) and is shown below.

$$\beta - EXP_{A1}(U,D) = \underline{A_1}_{\beta}(U,D_1) \cup \underline{A_1}_{\beta}(U,D_2)$$

The $\underline{A_1}_{\beta}(U,D_1)$ and $\underline{A_1}_{\beta}(U,D_2)$ can be calculated using (2) as follows.

The RCE for the set S_1 with respect to set D_1 is calculated as per (1) is obtained as follows,

$|S_1 \cap D_1| = |\{1,3,5,6,8\}| = 5$ and $|S_1| = 7$.

Therefore, $C(U, S_1, D_1) = 1 - (5/7) = 0.29$

The RCE for remaining partitions are obtained as,

$C(U, S_1, D_1) = 1-5/7 = 0.29$	$C(U, S_1, D_2) = 1-2/7 = 0.72$
$C(U, S_2, D_1) = 1-0 = 1$	**$C(U, S_2, D_2) = 1-4/4 = 0$**
$C(U, S_3, D_1) = 1-1/6 = 0.12$	$C(U, S_3, D_2) = 1-1/6 = 0.84$

The RCE for partitions S_2 and S_3 are less than the given MCE β=0.2.

Therefore, $\underline{A_1}_{\beta}(U,D_1) = S_3$ and $\underline{A_1}_{\beta}(U,D_2) = S_2$

Then, $\beta EXP_{A1}(U,D) = \underline{A_1}_{\beta}(U,D_1) \cup \underline{A_1}_{\beta}(U,D_2)$

$= S_2 \cup S_3 = \{2, 4, 7, 9, 11, 13, 14, 15, 16, 17\}$

The VPER values of all conditional attributes mentioned in Table I are obtained as follows,

$\beta EXP_{A1}(U,D) = \{2, 4, 7, 9, 11, 13, 14, 15, 16, 17\}$

$\beta EXP_{A2}(U,D) = \{2, 4, 6, 8, 10, 11, 12, 13, 14, 16\}$

$\beta EXP_{A3}(U,D) = \Phi$

$\beta EXP_{A4}(U,D) = \Phi$

VPRSM Based Decision Tree Induction: An Example

In VPRSM based DT classification, the attribute with premier VPER is opted as the splitting attribute. If multiple attributes are qualified with same cardinal VPER, then the attribute with highest Gain-Ratio [7] is opted as splitting attribute. The calculated cardinalities of VPERs for the four attributes mentioned in Table I are obtained as,

$$|\beta EXP_{A1}(U,D)| = 10, \quad |\beta EXP_{A2}(U,D)| = 10$$

$$|\beta EXP_{A3}(U,D)| = 0$$

$$|\beta EXP_{A4}(U,D)| = 0$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The attributes A_1 and A_2 are having the same highest Variable Precision Explicit Region values.

To resolve this ambiguous situation, choose the attribute with highest Gain-Ratio. The Gain-Ratio values for A_1 and A_2 are obtained as,

$Gain - Ratio(U,A_1) =$ 0.130865,

$\boldsymbol{Gain - Ratio(U,A_2) =}$ **0.190578**

The attribute A_2 with highest Gain-Ratio is qualified as root node and the final classification tree induced by VPRSM approach is obtained as shown in figure 1 with 7 leaf nodes and 3 internal nodes.

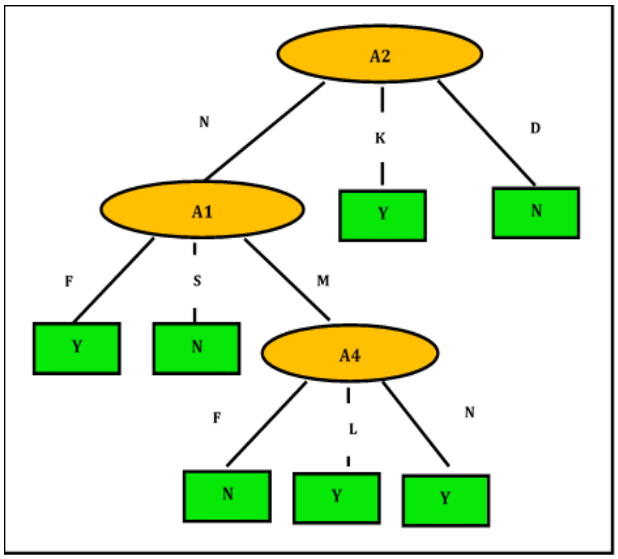

Fig. 1. Final DT induced by VPRSM approach

III. PROPOSED APPROACH

As illustrated in section II, when multiple attributes are qualifying with same maximum Variable Precision Explicit Region value, there is no improvement in the generalization ability of the tree induced by VPRSM approach when compared with that of Entropy and RST based methods. This paper proposes a solution to handle this uncertainty in selecting a best splitting attribute due to the ambiguity raised by VPER measure.

This section proposes an improved VPRSM based decision tree induction approach for handling uncertain data. When multiple attributes are qualified with same maximum VPER values, the proposed approach decides on the attribute with elevated β-Significance[8] as the splitting attribute. As β-Significance of an attribute in a set gives the importance of that attribute in arriving at a decision, selecting the attribute with highest β-Significance for splitting of samples generates more accurate decision trees.

A. β- Dependency

Let A ⊆ C, B ⊆ D, then the degree of dependency of the attribute set *A* with respect to the set of decision classes B, for an allowable misclassification β is represented by $\gamma^{\beta}(U,A,B)$ and is defined as,

$$\gamma^{\beta}(U,A,B) = \frac{|\beta EXP_A(U,B^*)|}{|U|} \quad (4)$$

Where, $\beta EXP_A(U,B^*)$ is the VPER of A and |U| denotes the cardinality of the set U.

The decision B and attribute set A are related to each other as follows.

If $\gamma^{\beta}(U,A,B)$ = 0, then the decision B is independent on attribute set A

$\gamma^{\beta}(U,A,B)$ = 1, then the decision B is completely dependent on attribute set A

$0 \leq \gamma^{\beta}(U,A,B) \leq 1$, then the decision B is partially dependent on attribute set A

B. β-Significance of Attribute

Significance of an attribute '*a*' in a candidate attribute set '*C*', gives the importance of attribute '*a*' in the set and can be observed as the change in the degree of dependency after removing '*a*' from the set '*C*'. The formal definition of β-Significance of an attribute is given below.

Definition

Let $a \in C$ be an attribute, then the β-Significance of attribute '*a*' for the decision class *D* with respect to candidate attributes *C* for a given admissible misclassification error β is represented by $\zeta(a,U,C,D,\beta)$ and is defined as[8],

$$\zeta(a,U,C,D,\beta) = \gamma^{\beta}(U,C,D) - \gamma^{\beta}(U,C-\{a\},D) \quad (5)$$

Where, $\gamma^{\beta}(U,C,D)$ is the β-Dependency of set C with respect to set D for a given β

β-Significance: Example

The β-Significance of all the conditional attributes for the data represented in Table I can be calculated using (5) as shown below.

$$\zeta(A_1,U,C,D,\beta) = \gamma^{\beta}(U,C,D) - \gamma^{\beta}(U,C-\{A_1\},D)$$

$\gamma^{\beta}(U,C,D)$ can be calculated using (4) as given below.

$$\gamma^{\beta}(U,C,D) = \frac{|\beta EXP_C(U,D^*)|}{|U|}$$

$\beta EXP_C(U,D^*)$ is obtained using (3) as follows.

(U,C^*)
={{1},{2},{3},{4},{5},{6},{7},{8},{9},{10},{11},{12},
{13},{14,16},{15},{17}}

$\beta EXP_C(U,D^*) = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17\}$

$\gamma^{\beta}(U,C,D) =$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

$$\frac{|\beta EXP_C(U,D^*)|}{|U|} = \frac{|\{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17\}|}{17} = \frac{17}{17} = 1$$

$\gamma^{\beta}(U, C-\{A_1\}, D)$ can be calculated as given below.

$$\gamma^{\beta}(U, C-\{A_1\}, D) = \frac{|\beta EXP_{C-\{A1\}}(U,D^*)|}{|U|}$$

$\beta EXP_{C-\{A1\}}(U,D^*)$ is obtained as follows,

Let W=C-{A_1} then

(U,W^*) = {{1,7},{2},{3},{4,12},{5,9},{6,11},{8,13},{10},{14,16},{15}, {17}}

$$\beta EXP_W(U,D^*) = \{2,3,4,6,8,9,10,11,12,13,14,15,16,17\}$$

$$\gamma^{\beta}(U, C-\{A_1\}, D) = \frac{|\beta - EXP_W(U,D^*)|}{|U|}$$

$$= \frac{|\{2,3,4,6,8,9,10,11,12,13,14,15,16,17\}|}{17} = \frac{13}{17} = 0.764705$$

Now,

$\zeta(A_1, U, C, D, \beta) = \gamma^{\beta}(U, C, D) - \gamma^{\beta}(U, C-\{A_1\}, D)$

= **0.235294**.

The β-Significance of all conditional attributes mentioned in Table I are obtained as follows,

$\zeta(A_1, U, C, D, \beta)$ = **0.235294**
$\zeta(A_2, U, C, D, \beta)$ = 0
$\zeta(A_3, U, C, D, \beta)$ = 0
$\zeta(A_4, U, C, D, \beta)$ = 0.117647

C. β-IDTA approach of inducing decision trees

The proposed β-IDTA approach of constructing decision tree choses the attribute with highest β-Significance in case of uncertainty. From section II, the VPERs of all four conditional attributes for the data in Table I are obtained as follows.

$|\beta EXP_{A1}(U,D^*)|$ **=10**
$|\beta EXP_{A2}(U,D^*)|$ **= 10**
$|\beta EXP_{A3}(U,D^*)|$ = 0
$|\beta EXP_{A4}(U,D^*)|$ = 0

The attributes A_1 and A_2 are having the same VPER values and hence to select the right one among A_1 and A_2, the β-Significance values of A_1 and A_2 have been calculated and the obtained β-Significance values are given below,

$\zeta(A_1, U, C, D, \beta)$= **0.235294**
$\zeta(A_2, U, C, D, \beta)$ =0

The attribute A_1 is most significant than A_2 and hence A_1 is selected as the splitting attribute at root level. The final decision tree induced by the proposed approach is consisting of only 4 leaf nodes and 2 internal nodes and is shown in figure 2.

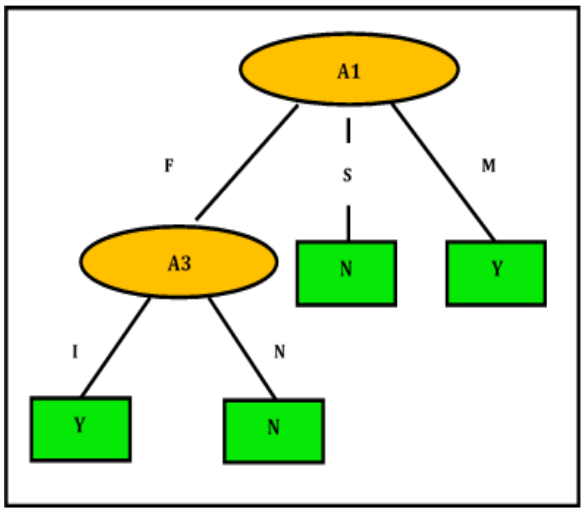

Fig. 2. Final DT induced by Proposed approach

From figures 1 and 2, it is observed that the proposed approach handles uncertainties very well and generates a comprehensive decision tree when compared with basic VPRSM approach.

IV. EVALUATION AND COMPARISON OF THE CLASSIFICATION MODELS

The performance of a classification model can be evaluated by its accuracy. The classification accuracy gives the probability of predicting the correct outcome of the classifying object. The prediction accuracy of any classifier can be calculated using the confusion matrix (or Error matrix) [9]. To evaluate the decision tree induced by the proposed approach, the test data represented in Table II is taken into consideration.

TABLE II. TESTING DATA

U	CONDITIONAL ATTRIBUTES				D
	A_1	A_2	A_3	A_4	
1	F	N	N	L	N
2	S	D	N	L	N
3	M	N	N	F	Y
4	F	K	N	N	N

The confusion matrix and the computed accuracies of the proposed and basic VPRSM approaches on the sample test data is given below.

VPRSM Approach

Confusion Matrix:

		PREDICTED CLASS	
		Y	N
ACTUAL CLASS	Y	**0**	1
	N	2	**1**

$$\text{Accuracy} = \frac{1}{4} \times 100 = 25\%$$

β-IDTA approach

Confusion Matrix:

		PREDICTED CLASS	
		Y	N
ACTUAL CLASS	Y	**1**	0
	N	0	**3**

$$\text{Accuracy} = \frac{4}{4} \times 100 = 100\%$$

Fig. 3. Prediction accuracies of Proposed and VPRSM approaches on test data

Figure 3 clearly depicts that the generalization ability of the proposed approach is better than the basic VPRSM approach of DT classification.

V. EXPERIMENTS AND RESULTS

This section gives an analysis of the experimental results conducted on proposed approach. The proposed approach is 10-fold cross validated [10] on ten different public bench mark medical datasets of the Irvine's UC machine learning repository. The information about the ten datasets used in experimentation is given in Table III.

TABLE III. DATASETS DESCRIPTION

DataSet name	No. of Instances	No. of Conditional Attributes	No. of Decision Classes
Diabetes	769	8	2
Breast Tissue	106	9	4
HeartDisease	303	13	3
Primary Tumor	339	16	20
Thoraric Surgery	470	16	2
Chronic Kidney	400	24	2
Thyroid	500	26	3
Liver Disorders	345	6	2
Hepatitis	112	18	2
Dermatology	359	34	6

The average values of all 10-fold experiment results are tabulated. The performance of decision trees induced by VPRSM and Proposed approaches given in Tables IV&V is expressed in terms of Tree size and prediction accuracy.

TABLE IV. PERFORMANCE OF VPRSM BASED DECISION TREE

Dataset	β	TreeSize	Accuracy(%)
Diabetes	0.15	799	74.03
Breast Tissue	0.025	256	48
Heart Disease	0.1	799	55.86
Primary Tumor	0.05	733	38.66
Thoraric Surgery	0.1	101	86.12
Chronic Kidney	0.01	5	97.05
Thyroid	0.025	19	98.18
Liver Disorders	0.35	45	66.92
Hepatitis	0.1	214	64.54
Dermatology	0.2	564	88.28

TABLE V. PERFORMANCE OF DECISION TREE INDUCED BY PROPOSED APPROACH

Dataset	β	TreeSize	Accuracy(%)
Diabetes	0.25	296	***74.83***
Breast Tissue	0.15	216	***52***
Heart Disease	0.1	758	***57.93***
Primary Tumor	0	774	***40***
Thoraric Surgery	0.1	104	***86.77***
Chronic Kidney	0	7	***98.23***
Thyroid	0.15	8	***98.63***
Liver Disorders	0.3	79	***68.46***
Hepatitis	0.12	208	***65.45***
Dermatology	0.15	470	***89.14***

By observing the experimental results, it is very clear that the proposed approach outperformed the basic VPRSM approach of DT classification in all aspects i.e. the proposed approach generated an optimal decision tree with minimum number of leaves with very less conditions to arrive at a decision and most importantly the prediction accuracy of the tree induced by the proposed approach is also increased when compared against the existing approaches. The comparison of the average classification accuracies and average number of leaf nodes generated by RST, VPRSM, and the presented approaches on different medical datasets is shown in Figure 4 and 5 respectively.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

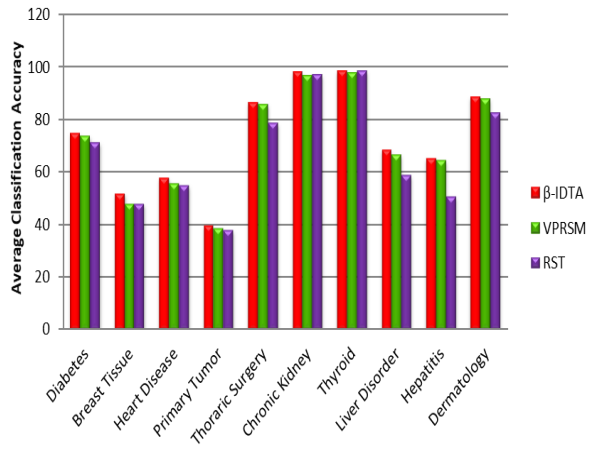

Fig. 4. Classification accuracies of the proposed, VPRSM and RST approaches on 10 different medical datasets

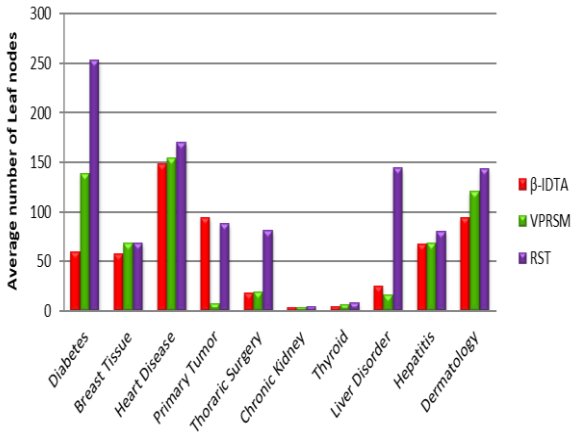

Fig. 5. Average number of Leaf nodes generated by proposed, VPRSM and RST approaches on 10 different medical datasets

From Figure 4 and 5, it is very clear that the proposed approach is better in all aspects i.e., generated optimal comprehensible trees with lesser number of leaf nodes and reduced tree size for an acceptable rate of increase in classification accuracy.

VI. CONCLUSION

This paper analyzed and illustrated the limitations of the basic VPRSM based decision tree classification. This work extends the existing VPRSM based tree inducing algorithm by proposing a solution to handle the ambiguity raised by VPER measure. The proposed approach can handle uncertainties and noise in data very efficiently by using β-Significance measure in ambiguous contexts. By comparing the experimental results of the proposed β-IDTA approach with RST and VPRSM approaches on public medical datasets mentioned in the paper unveils the improvement achieved by the proposed approach in inducing more precise and optimal decision trees.

REFERENCES

1. Z. Pawlak, "Rough Sets", International Journal of Computer and Information Sciences, 1982, 11(5), pp.341-356.
2. J. M. Wei, "Rough Set based approach to selection of node", International Journal of computational Cognition, 2003, 1(2), pp. 25-40
3. J. M. Wei, S. Q. Wang, M. Y. Wang, J. P. You and D. Y. Liu, "Rough set based approach for inducing decision trees", Knowledge-Based Systems, 2007, 20(8), pp. 695-702.
4. W. Ziarko, "Variable Precision Rough Set Model", Journal of Computer and System Sciences, 1993, 46(1), pp. 39-59.
5. J. M. Wei, M. Y. Wang, J. P. You, S. Q. Wang and D. Y. Liu, "VPRSM based decision tree Classifier", Computing and Informatics, 2007, 26(6), pp. 663-677.
6. A. Asuncion & D. J. Newman. (2019, January) "UCI Machine Learning Repository", 2007, Irvine, CA: University of California, Department of Information and Computer Science. [Online] Available:
7. http://www.ics.uci.edu/~mlearn/MLRepository.html.
8. T. Mitchell, Machine Learning. McGraw Hill, New York, 1997.
9. L. Rokach, and O. Maimon, Datamining with decision trees: theory and applications, World Scientific Publishing Co. Pte. Ltd., Singapore, 2008.
10. V. M.Patro, and M. R. Patra, "A novel approach to compute confusion matrix for classification of n-class attributes with feature selection", Transactions on Machine Learning and Artificial Intelligence, 2015, 3(2), pp. 52-64.
11. Y. Bengio and Y. Grandvalet, "No Unbiased Estimator of the Variance of K-Fold Cross-Validation", Journal of Machine Learning Research, 2004, Vol.5, pp. 1089–1105.

AUTHORS PROFILE

Surekha Samsani is presently pursuing Ph.D in Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India. She obtained her B.Tech in Computer Science and Engineering in 2003 and M.Tech in Computer Science from University College of Engineering Kakinada, JNTUH in 2008. She has 13 years of teaching experience and currently working as Assistant Professor in the department of Computer Science and Engineering, University College of Engineering Kakinada(A), JNTUK. Her research includes Data Mining, Big Data Analytics, Machine Learning, Pattern Recognition and Soft computing.

Dr. G.Jaya Suma is Professor, HOD of Department of Information Technology & Smart Campus In charge, University College of Engineering Vizianagaram, JNTUK, Andhra Pradesh, India. She has received her Ph.D. in the area of Data Mining from Andhra University, Visakhapatnam, India. She has 16 years of Teaching experience and 2 years Industrial experience. She has over 60 publications in reputed International Journals and Conferences. She is currently supervising 8 Doctoral candidates in the areas of Data Mining, Soft computing, Machine Learning, Big Data Analytics, Image Processing, Pattern Recognition and Mobile Computing.