# *A Comparative Study of Rough Set Theoretic Decision Tree Induction Algorithms*

*S.Surekha*
*JNTUK University College of Engineering*
*Vizianagaram*
*Vizianagaram, Andhra Pradesh, India*
*surekha.cse@jntukucev.ac.in*

***Abstract*—This paper gives a comparative study of Univariate and Multivariate decision tree classification based on Rough Set Theory.**

***Keywords— Decision Tree Classification; Rough Set Theory; Univariate; Multivariate.***

## I.INTRODUCTION

Decision Tree Classification[1] is an important white box approach used for classification of data. The ultimate goal of a decision tree classification algorithm is to produce a small scale tree with high accuracy in classification of unknown records. The core concept of any decision tree induction algorithm is the criteria used for selecting a best splitting attribute[2]. If only one attribute is selected as the splitting attribute in decision tree construction then it is called as Univariate[3] decision tree induction. Univariate tree construction algorithms often result in generating complex trees because making a decision of split based on the values represented by a single attribute often makes the induced tree model less efficient. So, making a decision based on the values of multiple attributes helps in inducing more efficient and scalable trees resulted in developing Multivariate[4,5] decision trees. In a multivariate decision tree induction algorithm, the attribute selection measure considers the values of multiple attributes and hence results in more scalable trees. Due to the existence of uncertainties and inconsistencies[6] in real world data the most popular decision tree induction algorithms often fail to induce efficient trees. Hence, incorporating the concepts of several set theories that deal with uncertainties may result in generating efficient trees. Rough Set Theory(RST)[7,8] is the most popular intelligent technique for dealing with uncertainties in the data. So, the concepts of Rough Set theory helps in selecting a best splitting attribute[9,10], which in turn results in inducing efficient decision trees. In this paper, the construction of Univariate and Multivariate decision tree induction algorithms based on rough set approach is explained through examples. The RST based Univariate and Multivariate decision tree algorithms are implemented in MATLAB on a computer with Intel Core i5 1.8GHz CPU and 4GB RAM. The accuracies of the algorithms are measured by conducting 10-fold cross validation tests on the datasets of UCI machine learning repository[11] and the results are tabulated in Experimental observations section.

## II.RST BASED DECISION TREE CONSTRUCTION

Decision Tree classification is one of the machine learning[12] techniques represent the classification knowledge in the form of a tree like structure. The decision tree structure acquires the knowledge by thoroughly training on the available training data. The obtained tree like model contains internal nodes and leaf nodes. Each internal node is used to test an unknown record and the outcome at that level indicates the branch to arrive at a leaf node and the label of the leaf node is assigned as the class of the record being examined.

The basic decision tree induction[13] algorithm is a top-down recursive algorithm. Based on the number of attributes considered in the attribute selection measure, decision tree algorithms are categorized into two types. They are

1. Univariate Decision Tree Classification
2. Multivariate Decision Tree Classification

*RST based Univariate Decision Tree*

The basic concepts of RST can be found in [6-10]. Let U represents a finite set of objects characterized into C classes by the feature set A. In RST based Univariate decision tree construction, the attribute with highest positive region[14] is selected as the splitting attribute.

Let P and Q represents two equivalence classes generated on U, then the P-positive region[10,13,14 ] of 'Q' is defined as,

$$POS_A(D) = \bigcup_{T \in U/D} \underline{B}T \quad (1)$$

Where, $\underline{P}T$ is the lower approximation[14] of P with respect to the target set T.

*Illustration:*

Let us consider the Information System(IS)[6] given in TABLE I as an example dataset all the 17 objects are categorized into two classes represented as C={C1,C2} characterized by the feature set A={ P,Q,R,S}. The set of equivalence classes generated by C is denoted as,

U/C= { {1,3,4,5,10,11,12,13,14,16} {2,6,7,8,9,15,17} }.

TABLE I. Sample DataSet

| U | P | Q | R | S | C |
|---|---|---|---|---|---|
| 1 | P1 | Q1 | R1 | S1 | C1 |
| 2 | P2 | Q2 | R2 | S2 | C2 |
| 3 | P1 | Q1 | R3 | S1 | C1 |
| 4 | P1 | Q1 | R1 | S3 | C1 |
| 5 | P1 | Q1 | R2 | S1 | C1 |
| 6 | P1 | Q2 | R1 | S1 | C2 |
| 7 | P2 | Q1 | R1 | S2 | C2 |
| 8 | P2 | Q1 | R2 | S2 | C2 |
| .9 | P1 | Q2 | R2 | S1 | C2 |
| 10 | P1 | Q1 | R2 | S3 | C1 |
| 11 | P1 | Q3 | R1 | S3 | C1 |
| 12 | P2 | Q3 | R3 | S3 | C1 |
| 13 | P1 | Q3 | R2 | S3 | C1 |
| 14 | P2 | Q3 | R2 | S3 | C1 |
| 15 | P2 | Q3 | R1 | S1 | C2 |
| 16 | P2 | Q3 | R3 | S1 | C1 |
| 17 | P2 | Q2 | R3 | S2 | C2 |

The set of equivalence classes generated by each attribute in the feature set A is obtained as,

U/P={ {1,3,4,5,6,9,10,11,13} {2,7,8,12,14,15,16,17} }
U/Q={ {1,3,4,5,7,8,10} {2,6,9,17} {11,12,13,14,15,16} }
U/R={ {1,4,6,7,11,15} {2,5,8,9,10,13,14} {3,12,16,17} }
U/S={ {1,3,5,6,9,15,16} {2,7,8,17} {4,10,11,12,13,14} }

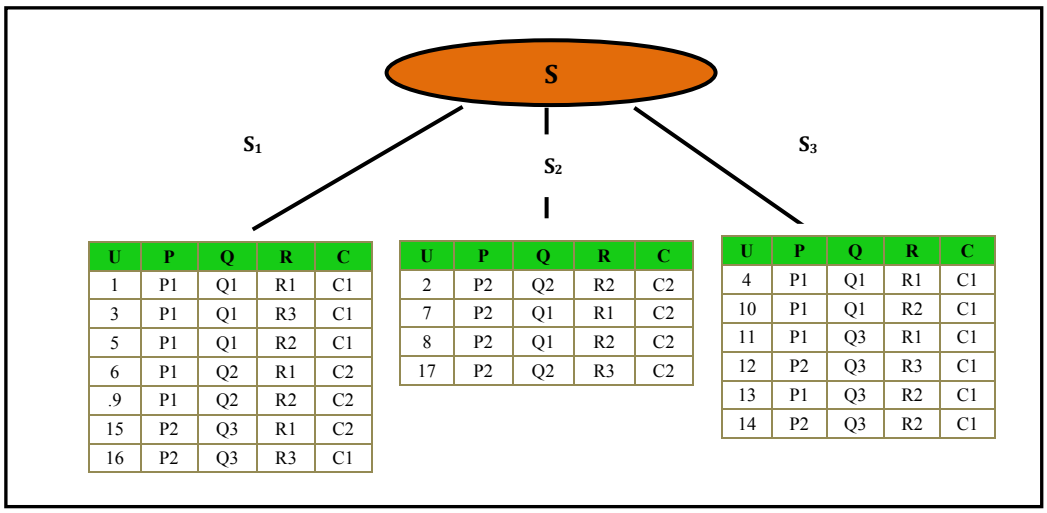



Fig. 1. Partial Decision Tree constructed by RST based Univariate Decision Tree Induction algorithm

U/C = { $T_1$, $T_2$}
Where $T_1$={1,3,4,5,10,11,12,13,14,16} and
$T_2$= {2,6,7,8,9,15,17}

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

Now, the Q-Positive region of C can be calculated using (1) and is obtained as,

$POS_Q(C) = \underline{Q}T_1 \cup \underline{Q}T_2$
$\underline{Q}T_1 = \Phi$ and
$\underline{Q}T_2 = \{2,6,9,17\}$ since, $\{2,6,9,17\} \subseteq T_2$
So, $|POS_Q(C)| = |\underline{Q}T_1 \cup \underline{Q}T_2| = |\{2,6,9,17\}| = 4$

Similarly calculating for P,R, and S,

$POS_P(C) = \Phi$, $POS_R(C) = \Phi$.
$POS_S(C) = \{2,7,8,17,4,10,11,12,13,14\}$
$|POS_P(C)| = 0$, $|POS_Q(C)| = 4$
$|POS_R(C)| = 0$. $|POS_S(C)| = 9$

The attribute with highest Positive region is the attribute S and hence select the attribute S as the splitting attribute and is the root node. The known possible values of attribute S are $\{S_1,S_2,S_3\}$. So, at root node for each value of S a separate branch will exist i.e., all 17 objects will be partitioned into 3 samples. The tree generated at root level is obtained as shown in Fig. 1.

The complete decision tree induced by the RST based Univariate decision tree algorithm is obtained as shown in Fig.2. The induced tree is consists of 3 internal nodes and total 7 leaf nodes. So, the tree size of the induced decision tree is 10 nodes.

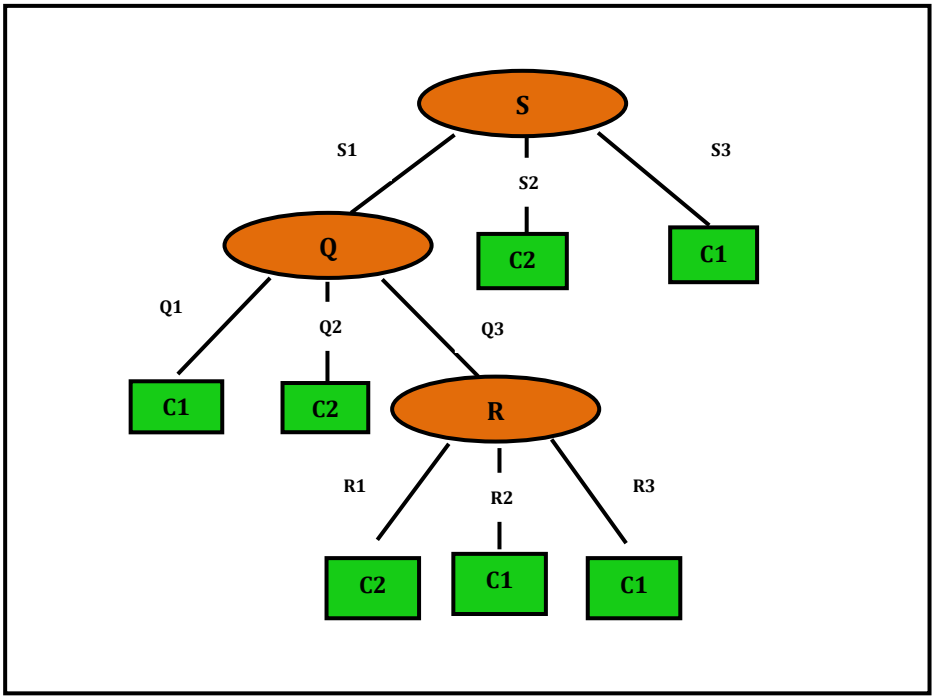



Fig. 2. RST based Univariate Decision Tree

*A. RST based Multivariate Decision Tree*

In the RST based Multivariate decision tree induction algorithm[5], the attribute selection measure always considers the attributes that are very much essential for making a decision. The concepts of Rough Sets can be applied to find out the Core[5] attributes in an Information System. As per the concepts of RS, an attribute X∈A is unnecessary if and only if there is no change in the decision upon the removal of X from the IS.

Let P and Q represents two equivalence classes generated on U, then an attribute X∈A is Q-unnecessary[5] if and only if the equation (2) is true otherwise X is Q-necessary.

$$POS_P(Q) = POS_{P-\{X\}}(Q) \quad (2)$$

a given IS, $CORE_Q(P)$ represents the set of all Q-necessary attributes. RST based multivariate decision tree construction considers the attributes in $CORE_Q(P)$ to make multivariate tests. The RST based multivariate tests are based on the concept of Relative Generalization[5] denoted by $GEN_Q(P)$ and is defined as,

Let,

$U/P = \{P_1, P_2, P_3, \ldots, P_n\}$ and
$U/Q = \{Q_1, Q_2, Q_3, \ldots, Q_m\}$
Then,

$$GEN_Q(P) = \{G_1, G_2, G_3, \ldots, G_m, G_{m+1}\} \quad (3)$$

Where,
$\forall$ i=1,2,3,…,m and j=1,2,…,n

$$G_i = \bigcup_{Pj \in U/Q} Pj \in Qi \quad (4)$$

$$G_{m+1} = \bigcup_{Pj \in U/Q} Pj \in Qj \ , \ \forall i \quad (5)$$

At each level in the process of constructing a multivariate decision tree using RST approach, samples at that level will be partitioned as per the relative generalization on the set of conditional attributes with respect to the decision attribute D. This process continues in a top-down manner till all the samples get trained.

Illustration:
Take the IS given in TABLE I as an example and compute the $CORE_A(C)$ using (2) as follows,

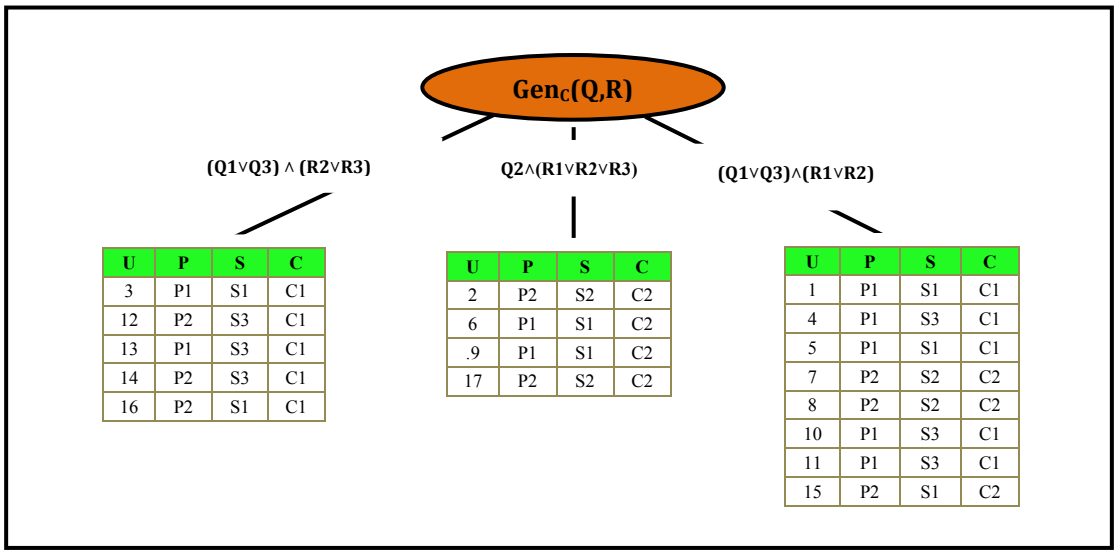
U/(A-{P}) ={{1}{2}{3}{4}{5}{6}{7}{8}{9}{10}{11}
{12}{13,14}{15}{16}{17} }
U/(A-{Q})={{1,6} {2,8} {3} {4,11} {5,9} {7} {10,13}
{12} {14} {15}{16}{17}}
U/(A-{R})={{ 1,3,5} {2,17} {4,10} {6,9} {7,8}
{11,13} {12,14} {15,16} }
U/(A-{S})={{ 1,4} {2}{3}{5,10}{6}{7}{8}{9}{11}
{12,16}{13}{14}{15} {17}}

So,
$|POS_A(C)| = 17$ , $|POS_{A-\{P\}}(C)| = 17$, $|POS_{A-\{Q\}}(C)| = 13$
$|POS_{A-\{R\}}(C)| = 15$, and $|POS_{A-\{S\}}(C)| = 17$

Therefore,
$CORE_A(C) = \{Q,R\}$
U/ $CORE_A(C)$ = {{1,4,7} {2,9} {3}{5,8,10}{6}
{11,15}{12,16}{13,14}{17} }

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

Now, compute $GEN_C(\{Q,R\})$ using (3),(4) &(5) as,




| U | P | S | C |
|---|---|---|---|
| 3 | P1 | S1 | C1 |
| 12 | P2 | S3 | C1 |
| 13 | P1 | S3 | C1 |
| 14 | P2 | S3 | C1 |
| 16 | P2 | S1 | C1 |

| U | P | S | C |
|---|---|---|---|
| 2 | P2 | S2 | C2 |
| 6 | P1 | S1 | C2 |
| .9 | P1 | S1 | C2 |
| 17 | P2 | S2 | C2 |

| U | P | S | C |
|---|---|---|---|
| 1 | P1 | S1 | C1 |
| 4 | P1 | S3 | C1 |
| 5 | P1 | S1 | C1 |
| 7 | P2 | S2 | C2 |
| 8 | P2 | S2 | C2 |
| 10 | P1 | S3 | C1 |
| 11 | P1 | S3 | C1 |
| 15 | P2 | S1 | C2 |

Fig. 3. Partial Decision Tree constructed by RST based Multivariate Decision Tree Induction algorithm

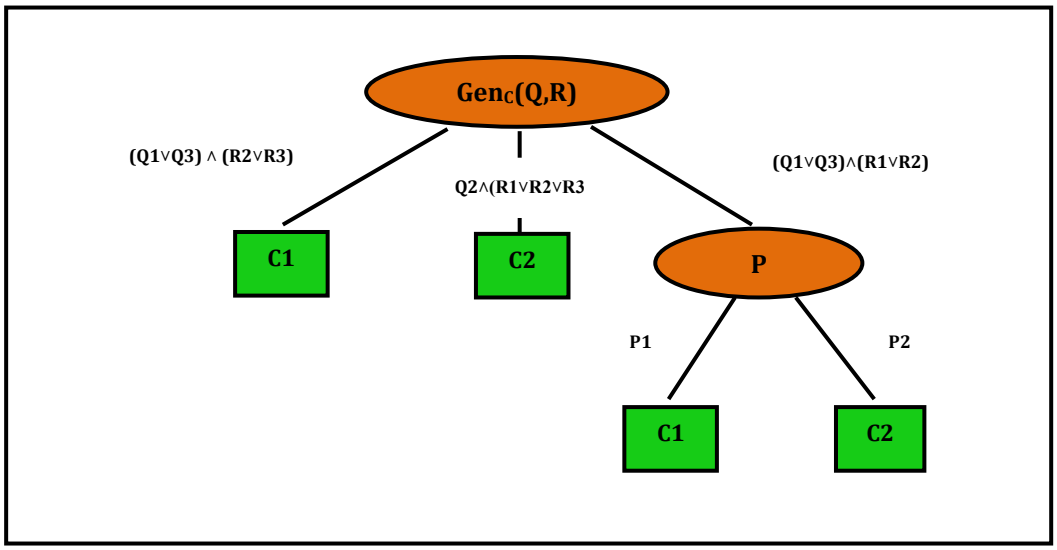



Fig. 4. RST based Multivariate Decision Tree

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

$GEN_C(\{Q,R\})$ is the root node and all the samples at root node will be partitioned into three samples based on the outcome of the $GEN_C(\{Q,R\})$.

The RST based Multivariate decision tree generated at root level is shown in Fig. 3 and the complete multivariate decision tree induced by RST approach is shown in Fig.4. The induced tree is consisting of 2 internal nodes and 4 leaf nodes. So, the total size of the induced decision tree is 6 nodes.

## III. Experimental Observations

In this section, the performance of Univariate and Multivariate decision trees generated using RST approach is compared by conducting a series of experiments on the 5 datasets taken from UCI ML repository.

As decision tree classification technique is a white box model it is mostly applied in medical diagnosis[15] and hence experiments were conducted on the UCI ML medical datasets. The description of the medical datasets is given in TABLE II.

TABLE II. Description of Datasets

| DataSet | Number of Instances | Number of Attributes | Number of Classes |
|---|---|---|---|
| BreastTissue | 100 | 9 | 4 |
| Liver Disorder | 2000 | 26 | 3 |
| Thoraric Surgery | 400 | 16 | 2 |
| Diabetes | 600 | 8 | 2 |
| PrimaryTumor | 300 | 16 | 20 |

The performance of a classification technique is measured in terms of its classification accuracy[16,17]. The prediction accuracy of a classification technique can be calculated as,

Accuracy= [(Number of correctly classified records ) / (Total number of records) ]×100

The classification accuracies of RST based Univariate and Multivariate decision trees is given in TABLE III

TABLE III. Classification Accuracies of Univariate and Multivariate Decision Trees

| DataSet | Classification Accuracy(%) | |
|---|---|---|
| | Univariate Decision Tree | Multivariate Decision Tree |
| Breast Tissue | 44.67 | 50.95 |
| Liver Disorder | 91 | 93.2 |
| Thoraric Surgery | 73.26 | 78 |
| Diabetes | 65.71 | 63.2 |
| Primary Tumor | 40.12 | 38 |

Observing the classification abilities of the Univariate and Multivariate decision tree techniques on the above mentioned datasets, the classification ability of the multivariate decision tree is better for Breast Tissue, Liver Disorder, and Thoraric Surgery datasets. Whereas, for the other two Diabetes and PrimaryTumor datasets Univariate decision tree classification technique generated good accuracy rates.

The complexity of a decision tree classification technique is measured as the number of internal and leaf nodes. As the number of nodes increases, the number of conditions to be verified while classifying an unknown record will also increase. Hence, the computational complexity of the tree is directly proportional to its size.

The tree sizes of the RST based Univariate and Multivariate decision trees is shown in Fig.5.

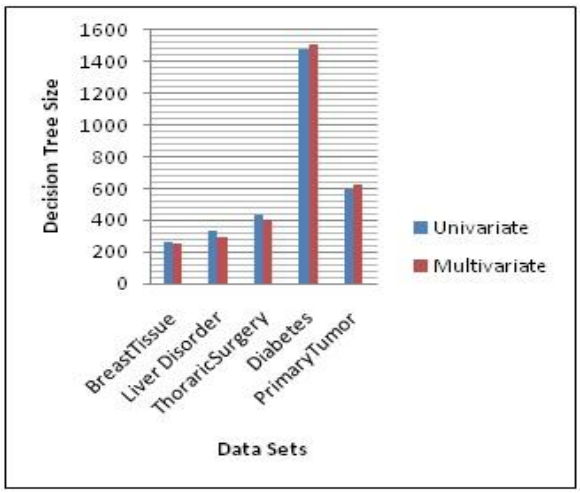



Fig. 5. Comparison of the Tree Sizes generated by RST based Univariate and Multivariate Decision Trees.

A decision tree classification technique is said to be efficient if and only if it is less complex and more accurate.

## IV. CONCLUSION

This paper clearly illustrates the RST based Univariate and Multivariate decision tree construction by taking a sample Information System. A Univariate test always performs on a single attribute whereas a Multivariate test considers more than one attribute to split the data samples and hence multivariate tree generation algorithms are generating more generalized and less complex trees than Univariate tree generation algorithms.

## REFERENCES

[1]. L. Rockah and O. Maimon, "Data mining with decision trees theory and applications", Machine Perception Artificial Intelligence, 2014,Vol. 69.

[2]. A.S.Bhatt, "Comparative analysis of attribute selection measures used for attribute selection in decision tree induction", in Proc. IEEE International Conference on Radar, Communication and Computing (ICRCC) 2012, Dec. 21-22, 2012, pp. 230-234.

[3]. J. R. Quinlan, "Introduction of decision trees ", Machine Learning,, 1986, pp. 81-106.

[4]. C.E.Broadley and P.E.Utgoff, "Multivariate versus univariate decision trees", COINS Technical Report, Jan. 1992.

[5]. X.Liu, H.Huang, and W.Xu, "A contribution to decision tree construction based on rough set theory", in Proc. 4th International Conference, RSCTC 2004, pp. 637-642, June 1-4, 2004.

[6]. Surekha. S, "An RST based efficient preprocessing technique for handling inconsistent data", in Proc. IEEE ICCIC 2016, pp.298-305, Dec. 15-16, 2016.

[7]. Z. Pawlak, "Rough Set approach to knowledge based Decision-Support", European Journal of Operational Research, Vol.99, No.1, 1997, pp.48-57.

[8]. Z. Pawlak, "Rough Sets", International Journal of

Computer and Information Sciences, 1982, Vol.11 No. 5, pp. 341-356.

[9]. J.G.Bazan, H.S. Nguyen, S.H.Nguyen, P.Synak, and J.Wroblewski, "Rough set algorithms in classification problem," Springer Rough set Methods and Applications, Physica-Verlag2, 2000, pp.49-88.

[10]. J. M. Wei," Rough Set based approach to selection of node", International Journal of Computational Cognition, 2003, Vol. 1, No. 2, pp. 25–40.

[11]. UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml/]. Irvine, A University of California, Center for Machine Learning and Intelligent Systems.

[12]. T. M. Mitchell, "Machine Learning", Singapore, McGraw-Hill, 1997.

[13]. Surekha. S, "A survey of various tree based classification techniques", International Journal of Advanced Research in Computer Science and Software Engineering, March 2017, Vol. 7, No. 3, pp.59-65.

[14]. 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, Vol.20, pp.695-702.

[15]. P. Pattaraintakorn and N. Cercone, "Integrating rough set theory and medical applications", Applied Mathematics Letters, April 2008, Vol.21, No.4, pp.400-403.

[16]. Q. Wei, L.Ronald, and Jr. Dunbrack, " The role of balanced training and testing data sets for binary classifiers in bioinformatics", PloS one, July 2013, Vol.8, No.7, e67863.

[17]. 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, Vol.3, No.2, pp.52-64.