Random Forest Classifier based on Variable Precision Rough Set Theory

Subi Jain
M.Tech Scholar
Samrat Ashok Technological Institute, Vidisha M.P. India

Gagan Vishwakarma
Assistant Professor
Samrat Ashok Technological Institute, Vidisha M.P. India

Yogendra Kumar Jain
HOD, Computer Science
Samrat Ashok Technological Institute, Vidisha M.P. India

ABSTRACT

Decision-making process is supported by Machine learning-based classification techniques in many areas of health care. Classification performance of decision system can be improved using the attribute reduction mainly in the situation of high data dimensionality dilemma .This paper proposes, Random forest Classifier (RFC) approach which is based on the Variable Precision Rough Set (VPRS) theory. The first phase of proposed approach focus at attribute reduction of available dataset using VPRS .Directing from dimensionality reduction to predictive model construction, and in next phase, the obtained abridged dataset is provided as the input of RFC to build a more accurate classification model. The performance is evaluated in terms of classification accuracy and time complexity. The experimental results show that the enhanced RFC has higher accuracy and correctly classified instances as compared with the existing algorithms.

Keywords

Decision tree classification, Attribute reduction, Variable Precision Rough Set **(**VPRS), Random forest classifier (RFC).

1. INTRODUCTION

Today's rapidly evolving torrent of data represents huge opportunity for researchers immersed in the area of data mining, information discovery, intelligence control knowledge extraction, and data pattern discovery. A tremendous amount of data is available around and volume is swelling at cyclonic speed but the potential that subsists in massive streams of structured and unstructured data cannot be take in for decision making or obtaining conclusions until this data is transformed into useful information.

Rough Set Theory [1] was brought into existence by Pawlak (1982) is an advanced form of traditional set theory offers distinctive approach to deal with the imprecise and incomplete data. Thus became a prevalent approach in the area of Artificial Intelligence, decision analysis and machine learning. Purpose of any machine learning algorithm is to learn only most suitable attributes for building their decisions. In data mining, decision tables are structured with an enormous number of attributes. Few of these are insignificant, which tends to increase the workload on resources, influencing the Rule Extraction Process and ultimately declining the accuracy of result. Because of the adverse effect of unimportant attributes, it is necessary to precede learning process with attribute reduction phase which eradicate the futile or irrelevant information. This procedure of removing redundant attributes from the decision table is also termed as feature selection, and accomplished to refine an information system. However in these algorithms there have been additionally some disadvantages like the create**d** tree was too complicated and they lacked the flexibility to tolerate possible noises in real world data sets. Due to this reason, several researchers used VPRS to induce decision tree.

Variable Precision Rough Set Theory (VPRST) is an extension to RST which shows quite robustness to misclassification and noise in data. VPRS is better than RST as it relaxes the strict inclusion in approximations of RST to partial inclusion by taking into account a parameter as an inclusion degree. VPRS allows objects to be classified with some admissible error, denoted by β (beta) where $0\leq \beta<0.5$, that is less than a particular predefined level [2]. Decision tree classification is a widely used technique in the field of data mining. Decision tree classifiers have been played an important role in many supervised-learning tasks. A Decision tree is a group of nodes and edges organized in a tree-like structure. Splits take place on internal nodes, while class labels get stored in the terminal nodes recognized as leaves. Decision trees are more effective and expeditious as compare with the other data mining techniques. ID3, C4.5, CARPT, CHAID, PUBLIC, SLIQ AND SPRLN are some of the well-known decision tree algorithm. Numeric and Nominal both kinds of attributes can be handled by these algorithms. The potentiality of dealing with the datasets that may have errors or missing values makes the decision trees more efficient [3].Problem concerned with single decision trees is that they are probable to suffer from high Variance. Another trouble associated with individual decision trees is that they are likely to over-fit and generalize poorly. Ensembles of decision trees, such as random forests, remove the difficulty of over-fitting by introducing a factor of randomness while constructing the individual trees and generating an assembly of such randomized trees [2, 3].

Random Forest is an influential assemble prediction technique that utilizes the strength of multiple number of decision trees and cautious randomization for generate accurate predictive models. This methodology was originally proposed by Ho [4], Amit and Geman [5] and later on by Breiman [6]. In the task of regression and classification obviously RFC have been found to be more accurate than individual decision tree. A number of trees are developed independently and parallel.

Proposed Algorithm applies variable Precision rough set theory to eliminate the redundant attributes in the decision system. Reducing dimensionality of data by erasing unsuitable attributes increases the performance of learning algorithm, converging attention on most relevant attributes. And then exploits the outcome of VPRS as input to RFC to generate the more accurate predictive classification model.

2. RELATED WORK

Various improvements were presented to various decision tree classifiers including RFC. Some of them are discussed here.

In 2009, Kemal Polat and gune proposed A hybrid intelligent method based on C4.5 decision tree classifier and one against all approach for multiclass classification problems including lympography, dermatology, and image segmentation. This methodology demonstrates the accuracy of 87.95%, 96.71%, and 95.81% on above datasets respectively [7].

In 2011, Galian and Ghimir explored the performance of the RF classifier for land cover classification. Mapping accuracy, sensitivity to data set size and Noise these three parameters were taken as valuation criteria. Results depicts that the RFC approach yields accurate land cover classifications, with 92% overall accuracy. RF is robust to training data reduction and noise because significant. In addition, variables which RFC recognized as suitable for classifying land cover coincided with expectations [8].

In 2012, Alexander Hapfelmeierthis presents wide investigations of Random Forests for the analysis of data with missing values. Important aspects like predictive accuracy, variable importance and variable selection are also examined [9].

In 2013, Ahmad and Hanna contrived a random forest classifier (RFC) approach to diagnose lymph diseases. Initial stage of the this system focus at developing diverse feature selection using the algorithms such as genetic algorithm (GA), Relief-F, Principal Component Analysis (PCA), Fisher, Sequential Backward Floating Search (SBFS)Sequential Forward Floating Search (SFFS) for minimizing the dimension of lymph diseases dataset. In the second stage this reduced dataset is fed into RFC for proficient classification. This approach demonstrates that GA-RFC achieved the highest classification accuracy of around 92.2%. The size of input feature space is reduced from eighteen to six features by using GA [10].

In 2014, Mojtaba and Tasdizen presented a supervised classification method, called disjunctive normal random forest (DNRF). A DNRF is a collection of randomly trained disjunctive normal decision trees (DNDT).For constructing a DNDT, each decision tree in the random forest is taken as a disjunction of rules, which are actually conjunctions of Boolean functions. DNRFs were proved to be better to learn complex decision boundaries and attaining the low generalization error. The both DNRF and DNDT show better classification performance than conventional decision trees and random forest [11].

In this paper more refined approach of RFC have been proposed using VPRS at initial stage for dimensionality reduction, ultimately in improvement in the classification model.

3. RANDOM FOREST CLASSIFIER (RFC) AND VARIABLE PRECISION ROUGH SET THEORY (VPRST): PRELIMINARIES

This section provides a concise explanation of the basic structure of Random Forest Classifier and Variable Precision Rough Set Theory (VPRST) along with some of the key definitions.

3.1 Rough Set Theory

Let $\hat{I}_s = (\hat{U}, \hat{A})$ is denoting an Information System, where Û and Â are representing the finite and non-empty set of objects.Û Is the universal set and Âis set of features comprises of conditional and decisional attributes [15].

Definition1. An indiscernibility relation associated with non-empty finite attribute subset S⊆ Â can be expressed as

$$IND(\mathrm{S}) =\{(\mathrm{x}, \mathrm{x}') \in \hat{U}^2 | \forall a \in \mathrm{S}, f_a(x) = f_a(x')(1)$$

Here equation (1)$IND(\mathrm{S})$ is called the S-indiscernibility relation .Universe of Discourse Û is divided into finite subsets by $IND(\mathrm{S})$ termed equivalence classes and group of equivalence classes made by $IND(\mathrm{S})$ is symbolized by $\hat{U}/IND(\mathrm{S})$ or$[x]_S$. If $(x, x') \in \mathrm{S}$ then $xandx'$ are indistinguishable by attributes from S [3, 14].

Definition2. Let X $\subseteq \hat{U} andS \subseteq \hat{A}$ then lower and upper approximation of set X can be expressed as follows

$$\mathrm{S}_{lower}(\mathrm{X}) = \{x \in \hat{U} \mid [x]_\mathrm{S} \subseteq \mathrm{X}\} \qquad (2)$$

s

$$S^{upper}(\mathrm{X}) = \{x \in \hat{U} \mid [x]_\mathrm{S} \cap \mathrm{X} \neq \phi\} \qquad (3)$$

Rough set relies on the principle of upper approximations and lower approximations. Lower approximations are the group of objects that assuredly belong to subclass of interest, while an upper approximation is the group of objects that possibly belong to subclass. Collection of objects that cannot be classified with certainty to be neither inside the subset nor outside is come under the boundary region [2].

Definition3. Let P and Q are equivalence relation on Ûand P, Q ⊆ Âthen positive regions negative region and boundary region can be defined as

$$POSR_P(Q) = \bigcup_{\mathrm{X} \in \mathbb{U}/Q} \mathrm{S}_{lower}(\mathrm{X})$$

$$NEGR_P(Q) = \hat{U} - \bigcup_{\mathrm{X} \in \mathbb{U}/Q} S^{upper}(\mathrm{X})$$

$$BNDR_P(Q) = \bigcup_{\mathrm{X} \in \mathbb{U}/Q} \dot{\mathrm{S}}(\mathrm{X}) - \bigcup_{\mathrm{X} \in \mathbb{U}/Q} \d{\mathrm{S}}(\mathrm{X})$$

*Definition4.*Let a decision table T = $(\hat{U}, \check{C}, \check{D})$ then dependency of decisional attribute Ď on conditional attribute $\check{C}$ can be denoted by following equation

$$\gamma_{\check{C}}(\check{D}) = \frac{|POSR_{\check{C}}(\check{D})|}{|\hat{U}|}$$

Let we have known the conditional attribute M, and $M \subseteq \check{C}$ the significance of attribute $a\epsilon \subseteq \check{C} - M$ for $\check{D}$

$$\sigma(a, \mathrm{M}, \check{D}) = \gamma_{M \cup \{a\}}(\check{D}) - \gamma_M(\check{D})$$

Definition5. Let Decision table T = $(\hat{U}, \check{C})$.$\check{C}$ Is said to be independent, if for all c $\in \check{C}$ is indispensable. If R⊆ $\check{C}$, $IND(R) = IND(\check{C})$ and R is independent then R is the reduction of $\check{C}$ denoted as $RED(\check{C})$. After getting reduct core can be defined as Core (A) = RED (A) [3].

Reduct and Core are two major concepts of rough set. Reduct is a reduced subset of genuine conditional attribute set which does not contain any redundant or extraneous attribute along with holding the exactness of original set [2].Computation cost for rule extraction is also minimized by calculating Reduct. Core is the most obligatory and crucial attribute from the set of Reduct.

3.2 Variable Precision Rough Set Theory (VPRST)

Successful operation of RST dependent on the discrete data is sometime seen as major shortcoming of this approach. Indeed, this demand of RST implies objectivity within the data that is simply not present. As an illustration, during a medical data set for a Headache attribute values like Yes' or 'No' may not be considered as an objective because simply it cannot determine whether an individual includes a headache or not to a high degree of accuracy.

In rough set literature, many extensions are developed that conceive to handle higher the uncertainty present in real world data. Particularly, variable precision rough sets could be a generalized model of rough sets, permitting a controlled degree of misclassification by relaxing the subset operator. VPRS is a feature selection measure based on the approximate accuracy. It may be used to solve the issues brought by attribute explicit in some extent, and the computational complexity is under the attribute measure method of information gain. Experiments proved that this methodology has higher classification accuracy.The main theme of VPRS is to permit objects to be classified with an error smaller than a certain predefined level [15]. Let X, Y ⊆ Û the relative classification error can be written as

$$C_{error}(\mathrm{X},\mathrm{Y}) = 1 - \frac{|\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X}|}$$

Here note that
$C_{error}(\mathrm{X},\mathrm{Y}) = 0$ if only if X⊆ Y. In classification a degree of inclusion can be acquired by permitting a defined level of error i.e. β. X$\subseteq_{\beta}$ Y if and only if$C_{error}(\mathrm{X},\mathrm{Y}) \leq \beta$, 0≤ $\beta \leq$ 0.5.This range 0≤ $\beta \leq 0.5$ is called β Positive region.If we use $\subseteq_{\beta}$instead of⊆ then

β upper approximation can be defined as

$\dot{R}_{\beta}(\mathrm{X}) = \cup \{[x_R] \in \hat{\mathrm{U}}/R \mid [x_R] \subseteq_{\beta} \mathrm{X}\}$

β lower approximation can be defined as

$$\underset{\cdot}{R}_{\beta}(\mathrm{X}) = \cup \{[x_R] \in \hat{\mathrm{U}}/R \mid C_{error}[x_R], \mathrm{X} < 1 - \beta\}$$

$\dot{R}_{\beta}(\mathrm{X}\ ) = \underset{\cdot}{R}_{\beta}(\mathrm{X})$for β=0, this acts as conventional pawlak's RST.

3.3 Random Forest Classifier

Random forest is a method of aggregating multiple decision trees, which are trained on different parts of same training set, thus overcoming the trouble of individual tree [6].

High variance is considered to be the major disadvantage of tree classifiers. They are also suffering from over fitting problem and ignorance of a variable in case of small sample size. In practice it's not rare for little change within the training dataset to end in a totally different tree. A decision forest methodology has been invented with the motive to make the tree classification more stable. A random decision forest is an ensemble of random decision trees [10].

The random forest method contains mainly two ideas .First one is "bagging" invented by Breiman and second is "random selection features" devised by Ho. **Bagging** is short form for "bootstrap aggregation", is a kind of ensemble learning , so as to boost the accuracy of a weak classifier by making a group of classifiers. A process of forming a number of decision tree predictor, which can be used in aggregation to form a decision by consent, is called bagging [4, 5].

Bootstrap replicates are used for decision tree construction and it simply implies that each tree is build by all the training samples sampled uniformly with replacement. If the number of instances in a dataset is N, nearly 2/3 of the original size is at random elected through bootstrapping manner for N times. The remaining instances are used as an out-of-bag collection to be calculated. The group of out-of-bag is those instances that don't seem to be accustomed build the sub-trees and then used for calculating the error prediction.at each node, a random feature selection is raised for constructing a decision node. The decision forest is trained in such a way to optimize the parameters at every node of all tree In training a decision tree, every node of tree have access to merely a randomly selected subset of the whole set of features. If m is the number of features, then size of feature selected at every split is around $\sqrt{m}$or $^{m}/_{2}$.All sub-trees are largest trees because no pruning is done. A classifier is created by learning scheme from the sample and aggregate all the classifiers generated from the various trial to shape the final classifier. To categorize or classify an instance, each classifier accounts a vote for the class to which it belongs and the instance is categorized as a member of the class with the majority votes. If two more class equally gains the highest votes then winner is chosen at random [12].

In ensemble, each tree is developed independently. Observations which are not incorporated in this tree are "out-of-bag" for this. By calculating predictions intended for every tree on its out-of-bag observations, the prediction error of the bagged assemble can be estimated. Bagging performs dropping the variation of an unbiased base learner, like decision tree. Since random selection of attributes or features lower the correlation between different trees in the ensemble this lean to improve the predictive power of the ensemble. Random forests were proved to be quite robust to the effect of noise as well as outliers.

4. PROPOSED METHODOLOGY

RFC can be used with high dimensionality data but classification accuracy can be improved if irrelevant features are removed from dataset. This also leads towards the pruned tress, ultimately increasing the performance and classification accuracy .In proposed work , mainly two phases are there First is attribute reduction also known as feature selection and second is classification phase. In former one, VPRS has been applied to gain the optimal feature set from original dataset. Second phase is classification phase in which random forest classifier is used for improving the prediction accuracy and reducing the variance power.

Algorithm 1: VPRS

Input: Original Dataset with redundant attributes.

Output: Dataset after Attribute reduction.

Begin

Step1. Calculate the equivalence classes by using indiscernibility relation on each feature.

Step2. Compute the classification error of attribute a_i with respect to all attributesa_j, where i ≠ j.

Step3. Find the β_lower and β_upper approximations of feature a_i with respect to all a_j, where i ≠ j.

Step4. Determine the Significance of attribute a_i with reference to all a_j.

Step5. Select the features with maximum significance.

End

The Fig1 is showing the working of proposed algorithm in two phases. In first phase dataset is fed to attribute reduction process i.e. VPRS. In second phase output of first phase is fed to random forest classifier for classification of instances. Knowledge discovery is acting as boundary line for these two phases.

Algorithm2: VPRS-RFC

Input: Optimized N training samples after applying VPRS, each denoted by feature length of d.
Output: Random forest classification model.
Step1.Begin
Step2.For every tree in the forest
Step3.DoSample N data training sample with replacement
Step4.for all internal node in the decision tree do randomly sample m attributes (m $=\sqrt{d}$).
Step5.Opt the feature from randomly selected set which is having the most Information gain.
Step6.Split the input data points by use of the selected feature, generating left and right children nodes.
Step7**.** Return completely grown-up decision tree.
Output: Random Forest.

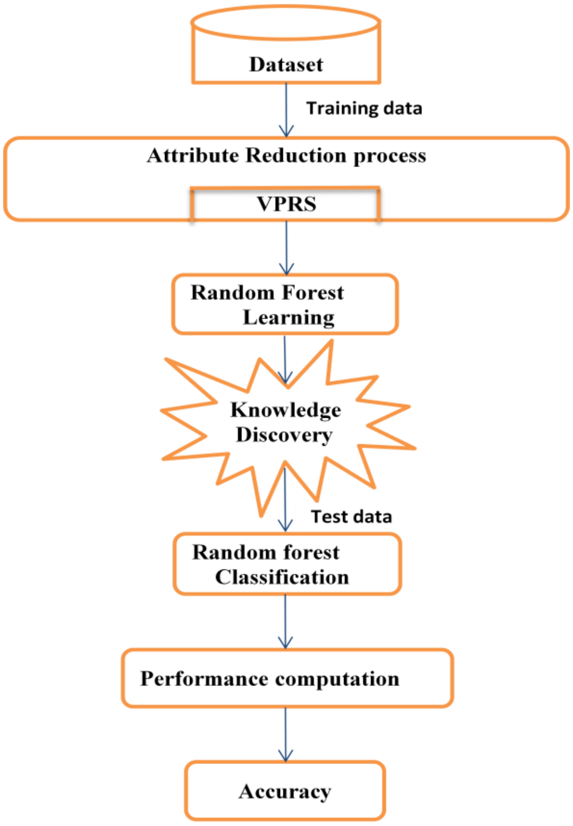


Figure 1. Model of Proposed Approach VPRS-RFC

5. RESULTS AND ANALYSIS

In this section the performance of VPRS- RFC is presented. All the experiments are performed by using Intel(R) Core(TM) i3 CPU @ 1.90GHz, 4.00 GB RAM personal computer and Microsoft Windows 8 64 bit operating system. Proposed algorithm is implemented using MATLAB tool.

5.1 Datasets

We have performed he implementation of this algorithm on lymphography dataset which is obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia .There are total 148 instances and 18 numeric valued attributes with no missing attributes. Data set is having four classes, namely normal, metastases, malign lymph [12].

5.2 Performance Analysis

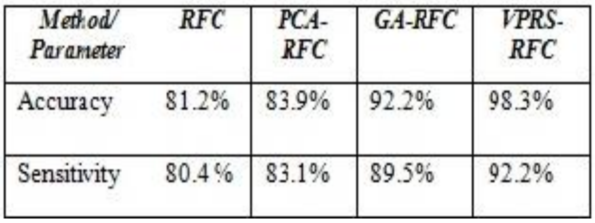
The performance of VPRS RFC is estimated by using performance parameters like classification accuracy and sensitivity, Calculation formula for which are given as below :

Classification Accuracy=$\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$*100%

Sensitivity=$\frac{TP}{TP+FN}$*100%

Where TP=True Positive or These are those positive tuples which were correctly labeled by the model.TN= True Negative. These represent negative tuples or cases which are correctly labeled by the classifier model. FP=False Positive. These are the negative tuples which are incorrectly marked as positive. FN=False Negative. These are positive tuples that were mismarked as negative.

Table 1
Comparison between VPRS-RFC and other approach



Method/ Parameter	*RFC*	*PCA-RFC*	*GA-RFC*	*VPRS-RFC*
Accuracy	81.2%	83.9%	92.2%	98.3%
Sensitivity	80.4 %	83.1%	89.5%	92.2%

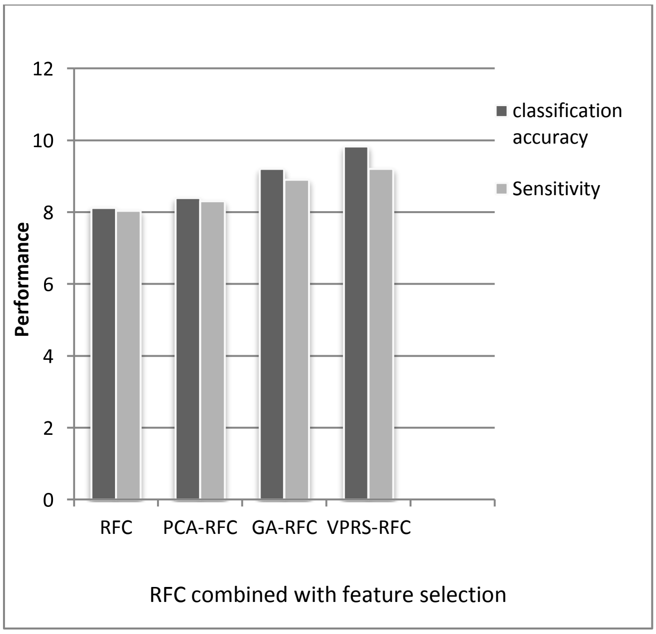


Figure 2. Comparison between VPRS-RFC and other approach

These four terms are also known as confusion matrix. Here we have compared our VPRS-RFC approach with RFC without any feature selection method, RFC with feature selection method named PCA and GA (Genetic Algorithm). Feature selection method plays a noticeable role in betterment of performance of RFC. Results which are depicted in Table 1 also show the same. For getting the reliable estimates for accuracy each experiment is executed using 10-fold cross-validation.

As shown in table 1 RFC without any attribute reduction shows the classification accuracy of 81 % and sensitivity80% for lympography dataset. PCA-RFC and GA-RFC demonstrates the accuracy of 84% and 92 % respectively. Methodology we projected Random Forest Classifier based on Variable Precision Rough Set Theory that is VPRS-RFC evidently illustrate the accuracy 98% which is obviously greater than other approach. Sensitivity for RFC, PCA-RFC, GA-RFC and VPRS-RFC is 80%, 83%, 89 % respectively. Sensitivity for our approach is 92%. A graph in fig 2 also gives a comparative picture of different methods on two parameters [10].

6. CONCLUSION AND FUTURE WORK

Accurate classification of multiclass dataset like lymphography is an imperative topic of concern in data mining. The concept of applying data mining technique to medical data can help in better prognosis and diagnosis of disease by dig out the hidden knowledge. The idea of attribute reduction performs a major task of discovering significant features. In this paper, a hybrid method based on VPRS and RFC is proposed with application to lymphography dataset. VPRS is employed for lessening the dimension of lymph dataset and RFC is utilized for intelligent and quick classification. The objective of this proposed work is to utilize the relevant and important features of RFC which tends to enhanced performance, swift learning speed, easier and less time consuming. Random forests classifier (RFC) is one amongst the foremost productive ensemble learning techniques and performs the classification by averaging the multiple decision trees. RFC proved to be more accurate, and fast as comparison to any other single classifier because of removing the difficulty of over fitting and high variance [13].

The proposed VRRS-RFC model performance is compared with other feature selection approaches united with RFC such as PCA, GA and alone RFC. VPRS- RFC achieved 98.3% classification accuracy and 92.2% sensitivity which is more than other approaches. Thus, these results validate effectiveness of VPRS-RFC strategy. Experimental results confirmed that proposed system worked appreciably well in diagnosis of lymph disease. The study also demonstrated that this approach can be used to attain proficient automatic diagnostic reports for other diseases. In future research we can apply this scheme for diagnosis of other disease. RFC can be hybrid with some other feature selection method for better results.

7. REFERENCES

[1] Pawlak, “Rough sets”, International Journal of Computing. Information Sciences, vol.11, pp.341-345, 1982.

[2] QiangShen and Richard Jensen, “Rough Sets, their Extensions and Applications,” International Journal of Automation and Computing, 04(1), pp. 100-106, 2007.

[3] J .R .Quinlan.” Induction of Decision Trees.” Machine Learning. 1, 81-106, 1 (Mar. 1986).

[4] T. Ho, “Random decision forest, in: 3rd International Conf. on Document Analysis and Recognition”, pp. 278–282, 1995.

[5] Y. Amit, D. Geman , “Shape quantization and recognition with randomized trees”, Neural Computing. 9, 1997.

[6] L. Breiman, “Random forests”, Mach. Learn. 45, pp 5–32, 2001.

[7] K. Polat, S. Gunes, “A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems”, Expert Syst. App. .: Int. J. 36 , 2009.

[8] V.F. Rodriguez-Galiano, B. Ghimire , J. Rogan , M. Chica-Olmo , J.P. Rigol-Sanchez,” An assessment of the effectiveness of a random forest classifier for land-cover classification”, Elsevier J. of ISPRS Journal of Photogrammetry and Remote Sensing 67, pp 93–104,2012.

[9] Alexander Hapfelmeier , “Analysis of Missing Data with Random Forests”,2012.

[10] Ahmad TaherAzar, Hanaa Ismail Elshazly, Aboul Ella Hassanien,Abeer Mohamed Elkorany”,A random forest classifier for lymph diseases”,Elsevier J. of computer methods and programs in biomedicine 113,pp 465–473,2014.

[11] Mojtaba Seyedhosseini, Tolga Tasdizen, “Disjunctive normal random forests”, Pattern Recognition, vol 48, pp 976–983, 2015.

[12] UCI. Machine Learning Repository. http://archive.ics.uci.edu/ml/index.html

[13] Faraz Akram, Seung Moo Han, Tae-Seong Kim, “An efficient word typing P300-BCI system using a modified T9interface and random forest classifier”, Computers in Biology and Medicine, vol. 56 ,pp 30–36,2014.

[14] CHEN Jiajun, HUANG Yuanyuan, “Decision Tree Construction Algorithm for Incomplete Information System”, IEEE fourth International Conference on Computational and Information Sciences, pp 404-407, 2012.

[15] In-Kyoo Park a , Gyoo-Seok Choi ,” A variable-precision information-entropy rough set approach for job searching” , Elsevier J. of Information Systems 48 ,279–288, 2015.