An Optimization Rough Set Boundary Region based Random Forest Classifier

Prerna Diwakar
DAVV, Indore

Anand More
Asst. Professor
DAVV, Indore

ABSTRACT
Machine learning is a concerned with the design and development of algorithms. Machine learning is a programming approach to computers to achieve optimization .Classification is the prediction approach in data mining techniques. Decision tree algorithm is the most common classifier to build tree because of it is easier to implement and understand. Attribute selection is a concept by which we want select attributes that are more significant in the given datasets. We proposed a novel hybrid approach combination of Rough Set with Boundary Region and Random Forest algorithm called Rough Set Boundary Region based Random Forest Classifier (RSBRRF Classifier) which is use to deal with uncertainties, vagueness and ambiguity associated with datasets. In this approach, we select significant attributes based on rough set theory with boundary region as an input to random forest classifier for constructing the decision tree is more efficient and scalable approach for classification of various datasets.

Keywords
Rough Set, Boundary Region, Decision Tree, Random Forest.

1. INTRODUCTION
Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology. In machine, learning and data mining fields, such databases establish with a considerable number of attributes are often detect. It is actually similar to find some of these attributes are irrelevant or redundant, which not only occupy extensive computing resources, but also seriously contact the decision making process. For these reasons, it becomes natural to defeat the irrelevant or redundant information and make the data set compact. Attribute reduction, also called feature selection, which is perform to process an information system, has been extensively research.Rough set theory, first designed by Pawlak (1982), can serve to deal with data classification problems by adopting the view of equivalence classes [6]. It provides a powerful tool for achieve reducts of information system. Such reducts do not contain redundant data, but uphold the same classification ability as the original information system. The role of rough set is to process an information system through removing redundant attributes. Many attribute reduction methods in rough set theory have been propose for achieving reduct and have wide applications in many fields [3, 15,7].

The reducts of information systems usually may be not one and only and all the reducts can be capture with the top-down attribute selection algorithm However, it has been prove that gathering all the possible reducts. Fortunately, in most of real-world applications, it is dispensable to find all the reducts. Generally, the reduct with the least attributes is elect as the optimal one in no show of other sources of information A growing number of attribute reduction methods are developing to acquire only one reduct. They usually implemented through a secure measure to evaluate the significance of attributes.

They induce with a candidate for the reduct composed of an empty attribute set and then focus on selecting attributes according to a sure selection criterion until the reduct achieves the same discrimination ability as the full decision table does. The attribute significance measures can be dependence and consistency [2].However, such approaches probably converge to a local optimum, and the acquired reducts may be not the minimum ones, but fake optimal reducts. In case a fake optimal reduct is treat as the optimal reduct, the complexity of convinced rules may increase and false decisions may be make. Therefore, the issue about how to acquire the optimal reduct of a decision table needs to be inspecting. According to the above analysis, this paper aims at proposing rough set boundary region based random forest classifier in rough set to modify an enhancement heuristic searching strategy, and finding the optimal reduct more effectively. Rough set boundary region based random for improving the attribute significance based heuristic attribute reduction methods is devise in this paper, providing a means of effectively achieving the optimal reduct for the dependence based heuristic reduction algorithm. By incorporating, the enhancement into representative one is algorithm with random forest classifier their improved versions are constructed. Numerical experiments demonstrate that of the improved algorithm can effectively achieve the optimal reduct without a huge increase in time complexity.

The precedent representative heuristic attribute reduction algorithms have a full variety of applications in machine learning and in data mining fields. However, they may be short due to the calculation of the significance measures. The problem start when more than one attribute has the equal greatest contribution, that is, the significances of different attributes are given as equally largest in some domical especially for the small data sets. In such cases, any one of such attributes is in any case select .In order to check whether the obtained reduct is the optimal one, the top-down attribute selection algorithm is using in foremost to acquire each reduct.

2. CONCEPTS

2.1Rough set theory
Rough set theory is an addition of conventional set theory that supports approximations in decision making [1,4,5] .It acquires many features in undistinguished. The lower approximation is an explication of the domain objects, which are know with certainty to belong to the subset of affection. The upper approximation is an explication of the objects, which not impossibly belong to the subset and the boundary region is the set of objects that cannot impossibly, but not certainly [8, 9]. A rough set is itself the approximation of an

ambiguous concept by limited concepts, called lower, upper and boundary approximations, which are a classification of the domain of affection into disjoint categories [12].It works by analyzing the granularity structure of the data only. In fact, by using only the given report, the theory estimates that the data is a true and accurate absorption of the real world. The numerical and other contextual expression of the data are ignored which may seem to be a significant exclusion, but keeps model assumptions to a minimal.

2.2 Information and Decision Systems

An information system represents data as a table of data, with rows (objects in the table) and columns (attributes). In medical datasets, for example, patients represented as objects and their analysis such as blood pressure, form attributes. The attributes values for a distinct patient are their specific reading for that their analysis. An information system may be expanding by the insertion of decision attributes. The terms attribute feature and variable are use mutually. For example, the medical information system discussed previously could be extending to insert patient classification information, such as whether a patient is fit or not. A more example of a decision system can be seeing in table 1. Here, the table consists of four conditional attributes (a; b; c; d), a decision attribute (e) with eight objects. A decision information system is consistent if for every set of objects. Moreover, whose attribute values are the compatible; the corresponding decision attributes are same difference. More formally, Q= (U, A) is an information system, where U is a non-empty set of finite objects and A is a non-empty finite set of attributes. Such a way that a: U $\rightarrow$ Va for every a $\in$ A. Va is the set of values that attribute a may take.For decision systems, A = {C $\cup$ D} where C is the set of input features and D is the set of class indices. Here, a class ordering d $\in$ D is itself a variable d : U$\rightarrow$ {0,1} such that for a $\in$ U, d(a) = 1 if a has class d and d(a) = 0 otherwise.

2.3 Indiscernibility

With any $\boldsymbol{P} \subseteq \boldsymbol{A}$ there is an associated evenness relation IND (p):

$$\text{IND (P)} = \{(i,j) \in U^2 \mid \forall a \in P, a(x) = a(j)\} \quad (1)$$

Note that this correlate to the evenness relation for which two objects are same difference if we have same vectors of attribute values for the attributes in P. The allotment of U, determined by IND (P) is express by U/IND (P), which is simply the set of equivalence classes reproduce by IND (P):

$$\text{U/IND (P)} = \bigotimes\{U/IND\ (\{a\}) \mid a \in P\} \quad (2)$$

Where,

$$A \bigotimes B = \{X \cap Y \mid \forall X \in A, \forall Y \in B, X \cap Y \neq \phi\} \quad (3)$$

If (i; j) $\in$ IND (P), then j and i are obscure by attributes from P. The equivalence classes of the indiscernibility relation with consideration to P are denoted [i]p, i$\in$ U. For the delineative example, if P = {b,c} then objects 1, 6 and 7 are indiscernible; as are objects 0 and 4. IND (P) creates the following allotment of U :

$$BNDp(I) = \bigcup_{i \in U/I} \overline{P}X - \bigcup_{i \in U/I} \underline{P}X \quad (8)$$

$$\text{U/IND (P)} = U/IND\ (b) \otimes U/IND(c)$$
$$= \{\{0,2,4\},\{1,3,6,7\},\{5\}\}$$
$$= \bigotimes\{\{2,3,5\},\{1,6,7\},\{0,4\}\}$$
$$= \{\{2\},\{0,4\},\{3\},\{1,6,7\},\{5\}\}$$

Table 1 An example dataset

i$\in$U	a	b	C	d	e
0	S	R	T	T	R
1	R	S	S	S	T
2	T	R	R	S	S
3	S	S	R	T	T
4	S	R	T	R	S
5	T	T	R	S	S
6	T	S	S	S	T
7	R	S	S	R	S

$$\text{U/IND (P)} = U/IND\ (b) \otimes U/IND(c)$$
$$= \{\{0,2,4\},\{1,3,6,7\},\{5\}\}$$
$$= \bigotimes\{\{2,3,5\},\{1,6,7\},\{0,4\}\}$$
$$= \{\{2\},\{0,4\},\{3\},\{1,6,7\},\{5\}\}$$

2.4 Lower and Upper Approximations

Let X$\subseteq$U,X can be approximate using only the information accommodate within P by constitute the P-lower and P-upper approximations of the classical crisp set X:

$$\underline{P}X = \{i \mid [i]p \subseteq X\} \quad (4)$$

$$\overline{P}X = \{i \mid [i]p \cap X \neq \phi \quad (5)$$

It is such a collection of rows and columns $\{\underline{P}X, \overline{P}X\}$ that is termed a rough set. Consider the approximation of abstraction X in figure 1. Each square in the diagram represents an evenness class; induce by indiscernibility enclosed by object values. Using the features in set B, via these evenness classes, the lower and upper approximations of X can be constructing. Evenness classes contained within X exist to the lower approximation. Objects equivocating within this region can be say to exist definitely to concept X. Evenness classes within X and onward its boundary form the upper approximation. Those objects in this region can only be saying to possibly exist to the concept.

2.5 Positive,Negative and BoundaryRegions

Let P and I be equivalence relations over U, then the positive (lower approximation), negative (upper approximation) and boundary regions are define as:

$$POSp(I) = \bigcup_{i \in U/I} \underline{P}X \quad (6)$$

$$NEGp(I) = U - \bigcup_{i \in \frac{U}{I}} \overline{P}X \quad (7)$$

The lower approximation constitutes all objects of U that can be classifying to classes of U/I using the information contained within attribute P. The upper approximation, NEG_p (I), is the set of objects that cannot be classifying to classes of

U/I. The boundary approximation, $BND_p(I)$, is the set of objects that can possibly, but not surly, be classified in this way. For example, let (P = {b,c}) & I = {e}, then

POSp (I) = $U\{\phi,\{2,5\},\{3\}\}=\{2,3,5\}$

NEGp (I) =U–$U\{\{0,4\},\{2,0,4,1,6,7,5\},\{3,1,6,7\}\}$
$= \phi$

BNDp (I) = $U\{\{0,4\},\{2,0,4,1,6,7,5\},\{3,1,6,7\}\}$
$-\{2,3,5\}$
$=\{0,1,4,6,7\}$

This means that objects 2, 3 and 5 can surly be classifying as belonging to a class in attribute e, when considering attributes b and c. The rest of the objects cannot be classifying, as the information that would make them appreciable is vanished.

2.6 Feature Dependency and Significance

A critical in data analysis is determining dependencies enclosed by attributes. Intuitively, a set of attributes I depends totally on a set of attributes P, express by P⇒ I, if all attribute values from I are antithetically determine by values of attributes from P. If there, existing a functional dependency between values of I and P. Then I depend totally on P. In rough set theory, dependency is express in the following way:

For P,I ⊂A, said that I depends on P in a degree q ($0 \leq q \leq 1$), denoted P⇒q I,if

$$q = \gamma_p(I) = \frac{|BNDp(I)|}{|U|} \qquad (9)$$

Where |S| stands for the cardinality of set S. If q = 1, I depends totally on I, if 0 < q < 1, I depends partially (in a degree q) on P, and if q = 0 then I does not depend on P:

$$\gamma_{\{b,c\}}(\{e\}) = \frac{|BND_{\{b,c\}}(\{e\})|}{|U|} = \frac{|\{0,1,4,6,7\}|}{|\{0,1,2,3,4,5,6,7\}|} = \frac{5}{8}$$

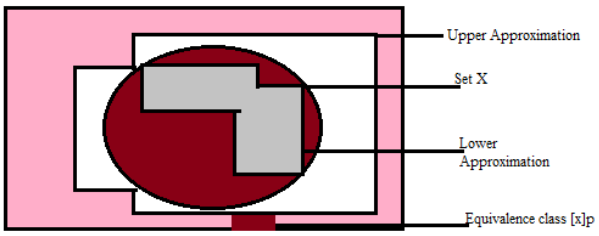

Figure 1: Rough set

In the example, the degree of dependency of attribute {e}from the attributes{b,c} is By computing the change in dependency when a feature is removed from the set of considered possible features, an estimate feature of the given significance of that feature can be obtained. The higher the change in dependency, the more significant the feature is. If the significance is 0, then the feature is dispensible. More formally, given P,I and a feature i ∈ P, the significance of feature i upon I is express by

$$\sigma p\ (I,a) = |\gamma_p\ (I)| - |\gamma_p\ (I) - \{a\}^{(I)}| \qquad (10)$$

For example, if we have P = {a,b,c} & I = e, then the following results will come

$$\gamma_{\{a,b,c\}}(\{e\}) = |\{0,1,4,7\}|/\ 8 = 4/8$$

$$\gamma_{(a,b)}(\{e\}) = |\{0,1,4,7\}|/\ 8 = 4/8$$

$$\gamma_{(b,c)}(\{e\}) = |\{0,1,4,6,7\}|/8 = 5/8$$

$$\gamma_{(a,c)}(\{e\}) = |\{0,1,4,7\}|/8 = 4/8$$

In addition, computing the significance of the three attributes gives:

$$\sigma p\ (I,a) = |\gamma_{\{a,b,c\}}(\{e\})| - |\gamma_{(b,c)}(\{e\})| = 1/8$$

$$\sigma p\ (I,b) = |\gamma_{\{a,b,c\}}(\{e\})| - |\gamma_{(a,c)}(\{e\})| = 0$$

$$\sigma p\ (I,c) = |\gamma_{\{a,b,c\}}(\{e\})| - |\gamma_{(a,b)}(\{e\})| = 0$$

From the of follows that attribute is an indispensable, but attributes b and c can be allocate with when considering the dependency between the given individual conditions attribute the decision attributes e.

2.7 Reducts

For many application problems, it is often basic to maintain a compact form of the information system [10,11,13]. One process to implement this is to search for a minimal representation of the original dataset. For this, the conception of a reduct is way out and defined as a minimal subset R of the initiatory feature set C such that for a given set of features D, $\gamma_R(D) = \gamma_C(D)$.

3. The Proposed System

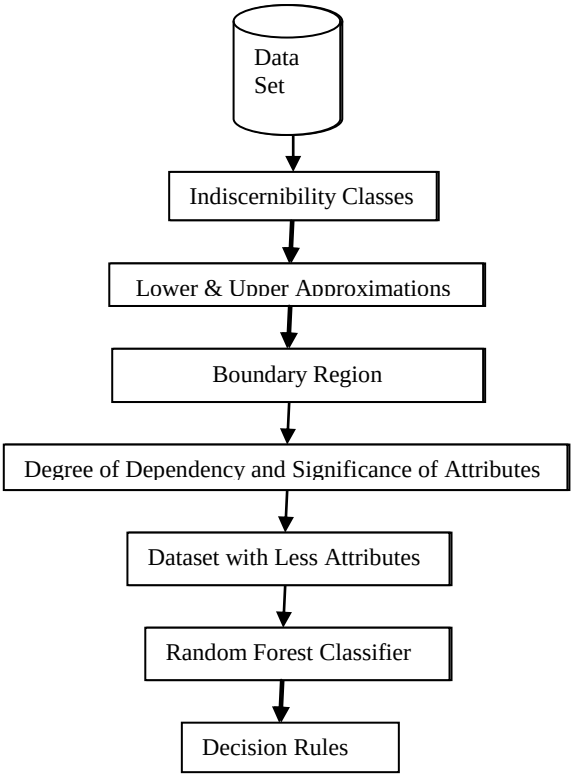

Figure 2: The Proposed System

In order to find reduct data set .we have proposed and algorithm based on random forest. Initially be having taken four UCI dataset. This dataset have uncertainties, vagueness and ambiguity associated In order to remove this problem be have introduced new approach a novel hybrid approach combination of Rough Set with Boundary Region and Random Forest algorithm called Rough Set Boundary Region based Random Forest Classifier (RSBRRF Classifier) which is use to deal with uncertainties, vagueness and ambiguity associated with datasets. In this approach, we select significant attributes based on rough set theory with boundary region as an input to random forest classifier for constructing

the decision tree is more efficient and scalable approach for classification of various datasets.

Step 1 Initially four UCI dataset has been taken.
Step 2 Secondly, we have purposed preprocessor based on rough set boundary region in matlab.
Step 3 We applied four UCI data set on rough set boundary region.
Step 4 Reduct dataset have been obtained.
Step 5 This Reduct data set using interface applies on random forest classifier in WEKA.
Step 6 Result occurs in term of accuracy and Time consumption.
Step 7 Compared result with J48 classifier and high accuracy as compare to j48.

3.1 Consistency based attribute reduction algorithm

Input: An information system $IS = (U, C \cup D, V, f)$
Output: One reduct R of the information system IS
Step 1: Compute the consistency $\delta_C(D)$ based on Rough Set boundary region.

Step 2: If $R \neq Core(C)$, remove redundant attribute if exists

Step 3: Output R.

3.2 RSBRRF Classifier

Now we propose our algorithm to generate a decision tree in the following way:
Input: An information system $IS = (U, C \cup D, V, f)$

Output: A decision tree *T*.

Step 1: All labeled samples initially assigned to root node, which is available in, reduct R of dataset.

Step2: N ← root node

Step3: With node N do

- Find the feature F among a random subset of features + threshold value T...

• ... that split the samples assigned to N into 2 subsets S_{left} and S_{right}...

• ... so as to maximize the label purity within these subsets

- Assign (F, T) to N
- If S_{left} and S_{right} too small to be splitted

• Attach child leaf nodes L_{left} and L_{right} to N

• Tag the leaves with the present label in S_{left} and S_{right}, respectively.

- Else

• Attach child nodes N_{left} and N_{right} to N

• Assign S_{left} and S_{right} to them, resp.

• Repeat procedure for N = N_{left} and N = N_{right}

Step4: Random subset of features

• Random sketch repeated at each node

• For D-dimensional samples, usual subset size = round (sqrt (D)) (also round (log2(x)))

• → Increases variety among the rCARTs + reduces computational load

Step 5: Output the decision tree T.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The implementation of the proposed Rough Set with Boundary Region based Random Forest Classifier is provided. Therefore, first, the required tools and techniques are discussed then after the code implementation and development of the system is provided. The following software and hardware require to implementation of the proposed system
Hardware Requirement: 2.0 GHz Processor required (Pentium 4 and above), Minimum 2 GB Random Access Memory 40 GB hard disk space. Software Requirement: Operating System (Windows 7 and above), MATLAB R2015b, Weka 3.7.2JDK1.

4.1 The Datasets

There are four data set taken here is:[14]
1. Lung Cancer
2. Hepatitis Data Set
3. Banknote Authentication Dataset
4. Lymphography Dataset

4.2 Results Analysis

4.2.1 Accuracy

Accuracy of proposed classification algorithm is a measurement of total accurate identified instances over the given samples. The accuracy of the classification can be evaluated on following datasets [15].

Table 2: Accuracy Comparisons between J48 and Rough Set Boundary Region based Random Forest Classifier

Datasets	**Instances**	**Attributes**	**J48 Accuracy (%)**	**RSBRRF Accuracy (%)**
Lung-Cancer	32	56	87.5%	99.90%
Hepatitis	155	19	92.25%	99.91%
Banknote Authentication	1372	5	90.26%	99.34%
Lymphography	148	18	93.24%	99.32%

The comparative accuracy of two algorithms are given using Table 2 shows the better performance of RSBRRF Classifier than J48 Classifier. According to the evaluated results the performance of the proposed algorithm is much better as compared to other algorithm.

4.2.2 Time Consumption

The amount of time consumption required to developing data model using proposed algorithm is as on following datasets. Time consumption means time complexity of the algorithm on various datasets.

The comparative time complexity of algorithms is giving using Table 3 shows the better performance of RSBRRF Classifier than J48 Classifier.

5. CONCLUSION AND FUTURE WORKS

This chapter draws the conclusion of entire study about the decision tree algorithms and their methods of performance enhancement. Therefore, a number of approaches are develop in recent years by which the classifiers are claim to provide much efficient classification accuracy in less complexity to overcome these computationally expensive in our proposed approach. In this presented work, feature selection is done by using Rough Set with Boundary Region and decision tree is constructed by Random Forest Classifier. The proposed algorithm is enhancing classification accuracy of datasets, reducing the size of tree and minimizing the redundancy in data.

Table 3: Time Consumption of J48 and Rough Set Boundary Region based Random Forest Classifier

Datasets	**Instances**	**Attributes**	**J48 Time Consumption(In Seconds)**	**RSBRRF Time Consumption(In Seconds)**
Lung-Cancer	32	56	0.03	0.02
Hepatitis	155	19	0.06	0.06
Banknote Authentication	1372	5	0.07	0.06
Lymphography	148	18	0.03	0.02

The proposed model is implemented using WEKA 3.7.2 and MATLAB R2015b and the comparative study is performed with respect to the J48 Classifier and RSBRRF Classifier. The comparison among these algorithms is performed in case of accuracy and time complexity.

The proposed classifier, RSBRRF produces high accuracy, low error rate and consumes less time as compared with J48 classifier. The proposed classifier is efficient and accurate which provides effective results as compared to the traditional algorithms. In future, we will optimize the performance of classification in terms of memory consumption and training time. In future, we will parallel this classifier for analysis of big data.

6. REFERENCES

[1] Z. Pawlak. Rough Sets.International Journal of Computer and Information Sciences, vol. 11, no. 5, pp. 341–356, (1982).

[2] Dash, M., & Liu, H. Consistency-based search in feature selection. Artificial Intelligence, vol.151, no.1-2, pp. 155–176,(2003).

[3] Dai, J. H. Set approach to incomplete data. Information Sciences, vol.241,pp. 43,no.572002,(2013).

[4] I. D¨untsch, G. Gediga. Rough Set Data Analysis.In: A. Kent & J. G.Williams (Eds.Encyclopedia of Computer Science and Technology, vol.43, no. 28, pp. 281–301, (2000).

[5] H. Sever. The status of research on rough sets for knowledge discovery in databases. In: Proceedings of the Second International Conference on Nonlinear Problems in Aviation and Aerospace vol.2, no.98 pp.673–680,(1998).

[6] Ahmad, A., & Dey, L. A feature selection technique for classificatory analysis. Pattern Recognition Letters, vol. 26,no.1, pp.43–56,(2005).

[7] Chai, J. Y., & Liu, J. N. C. (2014). A novel believable rough set approach for supplier selection. Expert Systems with Applications, vol.41,no.1,pp. 92–104,(2014).

[8] A. Skowron, Z. Pawlak, J. Komorowski, L. Polkowski. A rough set perspective on data and knowledge. Handbook of data mining and knowledge discovery, , Oxford University Press, pp. 134–149 (2002).

[9] A. Skowron, S. K. Pal. Special issue: Rough sets, pattern recognition and data mining. Pattern Recognition Letters, vol. 24, no. 6, pp. 829–933,(2003).

[10] Hu, Q. H., Zhao, H., Xie, Z. X., & Yu, D. R. Consistency based attribute reduction. In Z.-H. Zhou, H. Li, & Q. Yang (Eds.), PAKDDLNCS (LNAI) Vol.4426,(2007).

[11] Deng, T. Q., Yang, C. D., & Wang, X. F. A reduct derived from feature selection. Pattern Recognition Letters, vol,33, pp.1628–1646,(2012).

[12] Qian, X. F. Application research of rough set theory in transformer faultdiagnosis (Master Thesis). Nanjing University of Science and Technology. Hindawi Publishing Corporation Journal of Applied Mathematics Vol.2013, Article ID 263905.(2005).

[13] Teng, S. H., Wu, J. W., Sun, J. X., Zhou, S. L., & Liu, G. Q. An efficient attribute reduction algorithm. In The Proceedings of 2nd international conference on advanced computer control (ICACC 2010) pp.471–475. China(2012).

[14] UCI Machine Learning Repository: Data Sets https://archive.ics.uci.edu/ml/datasets.html.

[15] Pawlak, Z. Rough set approach to knowledge-based decision support. European Journal of Operational Research, vol. 99, pp.48–57,(1997).