Novel Hybrid Approach with Combination of Rough Set and Random Forest Algorithm

Gourav Goyal
Department of information technology
SATI
VIdisha (M.P) India

Rashmi Nigoti
Department of information technology
SATI
VIdisha (M.P) India

ABSTRACT

Machine learning is a concerned with the design and development of algorithms. Machine learning is a programming approach 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 to select more significant attributes in the given datasets. This Paper proposed a novel hybrid approach with a combination of rough set and Random Forest algorithm called Rough Set based Random Forest Classifier (RSRF Classifier) which is used to deal with uncertainties, vagueness, and ambiguity associated with datasets. In this approach, the selection of significant attributes based on rough set theory as an input to Random Forest classifier for constructing the decision tree which is more efficient and scalable approach as compare to related work for lymph disease diagnosis studies.

Keywords

Machine learning, Rough Set, Decision Tree, Random Forest Classifier, Lymph disease

1. INTRODUCTION

The Concept of data mining is growing in popularity in the realm of Commerce business activities. Generally but it's kind of a misconceived or misunderstood topic. Data mining technique is a regression, classification, clustering and association rules. Data mining is applicable across industry sectors generally to extract trends and patterns. In the business world let's apply it to advertising marketing effectiveness. And apply e-commerce initiatives also apply it to that health care processes, Supply chain processes. There is a just a number businesses that are can be the mind with these techniques simply put any organization work that has data processes can be analysed with data mining and the results are extracting information. The data mining topic this whole an idea the concept is growing popularity wide because data continues to grow to think about social networking now LinkedIn, Twitter, and Facebook, what is it? it's more data it stated that described people what they do, what they like or they all are when you are out buying or doing whatever I was far as using services just conducting your daily lives more there is data gathering, data capturing and it's just the way it is the information economy the way to extract strategic information from that collected data. the data resources are with a data mining.

2. RANDOM FOREST ALGORITHM

In Random Forest Algorithm, the rationale term is a key point to understand it. The whole concept is the Combination of learning models increases the classification accuracy. When using different learning models, it can increase the accuracy of classification, Which is the main idea of a technique that is called bagging[1].

Bagging: The main idea is to average noisy and unbiased models in order to create another Model with a lower variance is called Bagging[2]. Random Forest, a very brief definition Random forest algorithm works as a large collection Of the correlated decision trees [3]. The name forest is because we use a lot of decision trees. The algorithm of random forest creates a lot of decision trees and use them to make a classification, that is why it is a technique based on the bagging technique.

$$S = \begin{pmatrix} f_{A1} & f_{B2} & f_{C1}\ldots\ldots\ldots & C_1 \\ f_{A2} & f_{B2} & f_{C2}\ldots\ldots\ldots & C_2 \\ \cdot & \cdot & \cdot & \cdot \\ f_{AN} & f_{BN} & f_{CN}\ldots\ldots\ldots & C_N \end{pmatrix}$$

With help of example, elaborated this concept suppose a matrix S, This matrix S is a matrix of training samples that will submitted to the algorithm to create a classification model. In this case fA1, fB1, fC1. These are a lot of features, for example the fA1 is the feature A of the first sample And Continue this samples up to N. The fBN is the feature B of the Nth sample and also have in the last column here the C1 and CN, which means that have lots of features and also have a training class So the aim is to create a random forest to classify this sample set and from this sample set create a lot of subsets with random values. Which means that, for example, in the first subset we used the line number 12, the line number 15 and also the line number 35 and also some other random elements here So get a subset From that previously shown sample set and from this element, create a decision tree name as Number1 Then make another random subset with different values, take the Sample number 2, Sample number 6 and the sample number 20 and many others With these values, create a Decision tree number 2. And from this M sample, create the M number of a decision tree. So, that is why call it a forest because it have a lot of decision tree and after the created all of these decision trees. After it going to repeat that used different samples from each of the sample set here, The created a subset of random samples. Then with all these decision trees. We have different variations of the main classification then use all of this decision trees to create a ranking of classifiers So, in the example here it show how can make the class prediction. suppose 4 decision trees is create. Here only to see how the class prediction will work. In this case we have 4 decision trees, the forest is composed of 4 trees, If there is a new element to classify then it going to ask for the first tree what is the prediction Suppose the first tree said that the classification of this sample is class 1, and then we called to the second decision tree and it says that is class number 3 and then the other says that it is class number 1 and the other says is class number 2 So we have 4 Decision trees, Independent decision trees here which were created using subsamples of the entire sample set and now count the number of votes for each class. It is obvious that it have 1 vote here and another vote here for class 1, so we have 2 votes for class 1 and a single vote for class 2 and a single vote for class 3 So the result is going to be class 1.This random forest classified all the elements and class 1 was the selected class of this classification. So it is a very simple process, the difficult here is to create a lot of decision trees. But the creation of decision tree belongs to the algorithm of decision trees. So, this is the idea behind the random forest algorithm.

3. PROPOSED METHODOLOGY

RSRF Classifier(Rough Set Random forest classifier) :

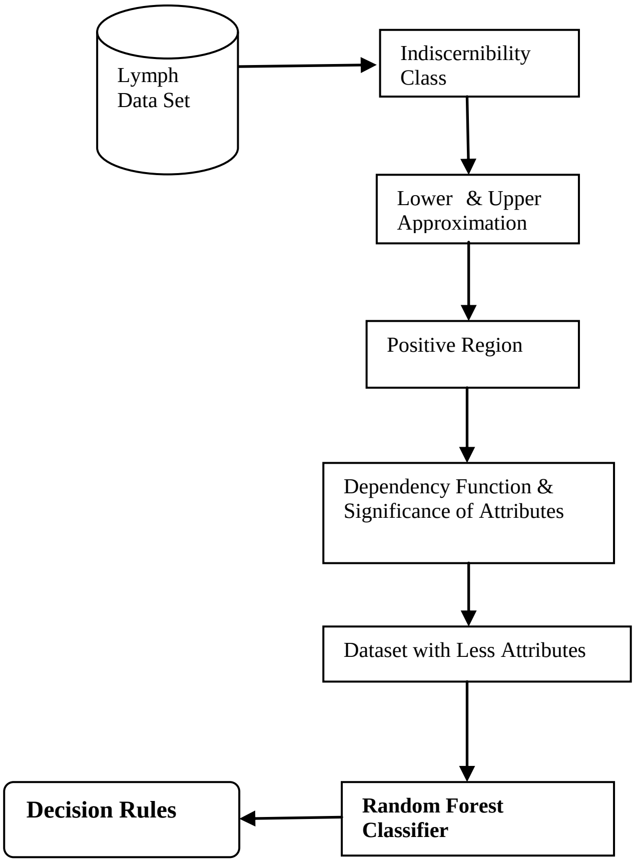

Lymph
Data Set
Indiscernibility Class
Lower & Upper Approximation
Positive Region
Dependency Function & Significance of Attributes
Dataset with Less Attributes
Random Forest Classifier
Decision Rules

Fig1 : The Proposed System

Now we propose our algorithm to generate a decision tree in the following way.

Input: An Information System IS= (U, A)

Output: Decision tree T.

Step 1: All labeled samples initially assigned to root node which is available in reduct 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 Sleft and Sright so as to maximize the label purity within these subsets

Assign (F, T) to N

If Sleft and Sright too small to be splitted

- Attach child leaf nodes L_{left} and L_{right} to N
- Tag the leaves with the most present label in S_{left} and S_{right}, respectively.

Else

- Attach child nodes Nleft and Nright to N
- Assign Sleft and Sright to them, resp.
- Repeat procedure for N = Nleft and N = Nright

Step4: Random subset of features

- Random drawing repeated at each node
- For D-dimensional samples, typical subset size = round (sqrt (D)) (also round (log2(x)))

Increases diversity among the rCARTs + reduces computational load

Step 5: Output the decision tree T.

4. EXPERIMENTAL RESULT

In this paper, The implementation of the proposed Rough Set based Random Forest algorithm is provided. Therefore first the required tools and techniques are discussed then after the code implementation and development of the system is provided.

Experimental Setup:

The following software and hardware require to an 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 3.7.2
- JDK 1.7

MATLAB R2015b

MATLAB (Matrix Laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language. It is developed by Math Works, MATLAB allows matrix manipulations, a creation of user interfaces, plotting of functions and data, implementation of algorithms and interfacing with programs written in other languages, including C, C++, Python, Java, and Fortran.

MATLAB R2015b is an 8.6 version. This version supports new MATLAB execution engine (a.k.a. LXE); graph and digraph classes to work with graphs and networks; MinGW-w64 as supported compiler on Windows.

WEKA 3.7.2

Weka is an effective tool for data mining. It is a popular suite of machine learning software written in Java programming language, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

Weka 3.7.2 is a work bench that contains a collection of algorithms and visualization tools for data analysis and predictive modelling, together with graphical user interfaces for easy access to these functions.

JDK 1.7

JDK stands for Java Development Kit. It physically exists. It contains JRE and development tools. The Java SE Development Kit (JDK) 7 Update 10 or JDK 8 is required. JVM support for dynamic languages, with the new, invoke dynamic byte code under JSR-292, various prototyping work currently done on the multi-language virtual machine. The Java SE Development Kit (JDK) 7 has several new features, enhancements, developments in server-side and core java.

The Datasets

The lymphography database was obtained from the University Medical Centre, Institute of Oncology, and Ljubljana, Yugoslavia [2]. There are 148 instances in total and there are no missing attributes. There are 18 numeric-valued attributes and four classes, namely normal, metastases, malign lymph and fibrosis as shown in Table 1, described with the majority class prevalence and the entropy of classes. The latter two numbers measure the difficulty of the classification task; higher entropy and lower prevalence of majority class indicate the most difficult problems.

Implementation using Code

This section is describing the required java based libraries and reference classes, implemented classes with their designed function and methods.

Various Functions:

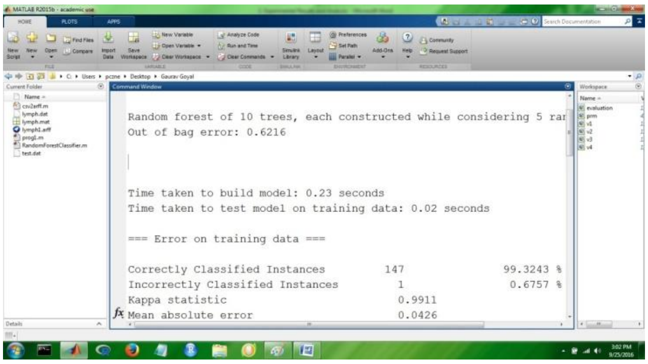
This section includes the required functions that are implemented to execute the desired task in implemented system.

Table 1: Various Functions

Functions	Description
sum(A,dim)	Sum of array elements
max(A)	Largest elements in array
diff(X)	Differences and Approximate Derivatives
repmat(A,n)	Repeat copies of array
find(X, n)	Find indices and values of nonzero elements

GUI Navigation

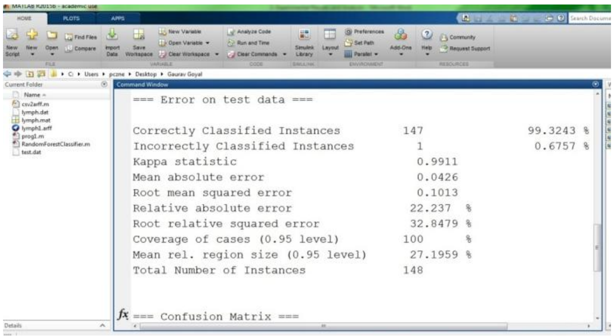
This section provides the information regarding the navigation of the current system and their implemented GUI.



```
Random forest of 10 trees, each constructed while considering 5 rar
Out of bag error: 0.6216


Time taken to build model: 0.23 seconds
Time taken to test model on training data: 0.02 seconds

=== Error on training data ===

Correctly Classified Instances         147               99.3243 %
Incorrectly Classified Instances         1                0.6757 %
Kappa statistic                          0.9911
Mean absolute error                      0.0426
```

Fig2:Rough Set based Random Forest Classifier Execution on Training Data

This Section provides the performance evaluation of the proposed algorithm. To justify the proposed solution; the comparative study is done with respect to the RSRF Classifier and related work for lymph disease diagnosis studies.



```
=== Error on test data ===

Correctly Classified Instances         147               99.3243 %
Incorrectly Classified Instances         1                0.6757 %
Kappa statistic                          0.9911
Mean absolute error                      0.0426
Root mean squared error                  0.1013
Relative absolute error                 22.237  %
Root relative squared error             32.8479 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      27.1959 %
Total Number of Instances              148


=== Confusion Matrix ===
```

Fig3: Rough Set based Random Forest Classifier Execution on Test Data

Results Analysis

Accuracy:

The Accuracy of proposed classification classifier is a measurement of total accurate identified instances over the given samples. The accuracy of the classification can be evaluated on lymph dataset.

The comparative accuracy of various algorithms are given using Table2 shows the better performance of RSRF Classifier than related algorithms for lymph disease diagnosis studies. According to the evaluated results, the performance of the proposed algorithm is much better as compared to other algorithms.

Table 2:Accuracy comparison between proposed approach and related algorithms for lymph disease diagnosis studies[1].

Author	Method	Accuracy (%)
Gutiérrez et al.	Two-stage evolutionary algorithm	85.05%
Karabulut et al.	Feature selection methods with NaïveBayes,MultilayerPerceptron (MLP), and J48 decision tree classifiers	84.46%
Abellán and Masegosa	Bagging Credal decision trees (B-CDT): C4.5/B-CDT without pruning (0% noise	79.96%/ 76.24% 79.69%/

	level) B-C4.5/B-CDT with pruning (0% noise level)	77.51%
Madden	Naïve Bayes Tree Augmented Naïve Bayes (TAN) General Bayesian network (GBN) with K2 search: GBN-K2: GBN with hill-climbing search: GBN-HC	82.16% 81.07% 77.46% 75.06%
Rodríguez et al.	ENDF:Ensembles(i.e,Randomization) of nested dichotomies using a forest method as base classifier. FND: A forest method using nested dichotomies of decision trees as base classifier.	82.57% 83.51%
Our approach	RSRF Classifier: Rough Set based Random Forest Classifier(Proposed Approach)	99.32%

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.

Table 3: Time consumption of RSRF Classifier and related algorithms for lymph disease diagnosis studies

Parameters	**Proposed RSRF Classifier**	**Related algorithms for lymph disease diagnosis studies**
Time Consumption	Low	High

The comparative time complexity of algorithms are given using Table 3 shows the better performance of RSFS Classifier than related algorithms for lymph disease diagnosis studies.

5. CONCLUSION AND FUTURE WORK

This Section draws the conclusion of entire study about the decision tree algorithms and their methods of performance enhancement. Based on the different experimentations and design aspects some essential points are observed which provided as a conclusion of research work additionally some future extension of the presented work is also provided.

Decision tree algorithm is a classical approach to supervised machine learning and data mining. There are a number of decision tree algorithms are available such as ID3, C4.5, and others. The decision tree algorithms are able to develop a transparent and reliable data model. In order to maintain the transparency and relativity between attributes decision tree algorithms are computationally expensive in terms of memory and time. Therefore a number of approaches are developed in recent years by which the classifiers are claimed to provide much efficient classification accuracy in less complexity. To overcome these computationally expensive in our proposed approach.

In this presented work rough set theory is used for feature selection and the decision tree is constructed by Random Forest Classifier. By combining this approach a new RSRF Classifier is proposed and implemented. The proposed algorithm is enhancing classification accuracy of datasets, reducing the size of a tree and minimizing the redundancy in data.

The proposed model is implemented using MATLAB R2015b and the comparative study is performed with respect to the related work for lymph disease diagnosis studies and proposed RSRF Classifier. The comparison among these algorithms is performed in a case of accuracy and time complexity.

The proposed algorithm, Rough Set based Random Forest (RSRF) Classifier produces high accuracy, low error rate and consumes less time as compared with related work for lymph disease diagnosis studies. Thus proposed algorithm provides efficient and effective results for classification of datasets.

The proposed algorithm is efficient and accurate which provides effective results as compared to related work for lymph disease diagnosis studies. In future, we will optimize the performance of classification in terms of memory consumption. In future, we will parallel this algorithm for analysis of big data. And also apply this approach to business analytical data for better productivity, for increasing the sale, for customer requirement fulfilment. This concept apply any datasets where feature selection is required.

6. REFERENCES

[1] A.Verikas , A.Gelzinis, and M.Bacauskiene "Mining data with random forests: A survey and results of new tests" ELSEVEIR, 2011

[2] Ahmad Taher Azara, Hanaa Ismail Elshazlyb, Aboul Ella Hassanienb, and Abeer Mohamed Elkorany "A random forest classifier for lymph diseases" ELSEVEIR, 2014

[3] Qiang Shen and Richard Jensen "Rough Sets, their Extensions and Applications" IJAC, 2008

[4] Xiuyi Jiaa, Lin Shangb,, Bing Zhouc, and Yiyu Yaod "Generalized attribute reduction in rough set theory" ELSEVEIR, 2015

[5] Joaquin Abellan, and Andres R. Masegosa "Bagging schemes on the presence of class noise in classification" ELSEVEIR, 2012

[6] Iffat A.Gheyas, and Leslie S.Smith "Feature sub setselection in large dimensionality domains" ELSEVEIR, 2010

[7] Mohammad Lutfi Othman, Ishak Aris, Mohammad Ridzal Othman, and Harussaleh Osman "Rough-Set-and-Genetic-Algorithm based data mining and Rule Quality Measure to hypothesize distance protective relay operation characteristics from relay event report" ELSEVEIR, 2011

[8] yiyu yao "Rough-set concept analysis: Interpreting RS-definable concepts based on ideas from formal concept analysis" ELSEVEIR, 2016

[9] Joaquin Derrac, Chris Cornelis, Salvador Garcia, and Francisco Herrera "Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection" ELSEVEIR, 2012

[10] Jianping Huaa, Waibhav D.Tembeb, Edward R.Doughertya, "Performance of feature-selection methods in the classification of high-dimension data" ELSEVEIR, 2009