A Novel Approach for the Assessment of Decision Stump & Upgraded Rf Classification Algorithms

M.Jayakameswaraiah,
Assistant Professor
School of Computer Science and Applications
REVA University, Bangalore, Karnataka India.
drjayakameswar@gmail.com

Pinakapani R
Assistant Professor
School of Computer Science and Applications
REVA University, Bangalore, Karnataka India.
rppani.mca@gmail.com

Ravi Dandu
Assistant Professor
School of Computer Science and Applications
REVA University, Bangalore, Karnataka India.
ravi_d@reva.edu.in

S.Ramakrishna
Professor, Department of Computer Science
Sri Venkateswara University,
Tirupati, Andhra Pradesh, drsramakrishna@yahoo.com

ABSTRACT

The classification models in data mining consists of decision tree, neural network, genetic algorithm, rough set, statistical model, etc. In this research, we have proposed and deliberated a new algorithm called Upgraded Random Forest, which is applied for the classification of sensor discrimination dataset. Here we considered for classification of multisource Sensor Discrimination data. The Upgraded RF approach becomes extreme attention for multi-source classification. The methodology which we are developed is not only a nonparametric but it also applies for the assessment and significance of the specific variables in the classification.

Keywords: **Data Mining, Classification, Decision Stump, Random Forest and Upgraded RF.**

I. INTRODUCTION

In data mining field we have various classification algorithms that include Decision Tree, Neural Networks, K-Nearest Neighbor and many more. In the development way of the KDD, the process will starts from the collection of significant data; which is going to be processed and transform processed data into useful information. Then the mining process will carry out from the hidden pattern [4,8]. The knowledge discovery in database process can be categorized as:

- Data Selection
- Pre-Processing
- Transformation
- Data Mining
- Interpretation/Evaluation

II. LITERATURE REVIEW

A. Classification

Classification is one of the most powerful technique to implement for supervised learning approaches. This is most widely used in data mining to classify the data from raw data. In the process of classification, a significant value is assigned to each and every item in a set of data to set the data in a form of a class. We have different types of classifier models in data mining, those simulation models are mathematical techniques which are used to classify the data in an effective manner[6,9]. The most useful classification techniques in data mining are decision tree, Bayesian classification, SVM, neural networks and the association-based classification. We studied and analyzed the performance of several algorithms for data classification in data mining such as Random Forest and Decision Stump classifiers[2,5]. For classification of data, the structural model shown in Figure: 1 can be designed and applied algorithms to process in effective way.

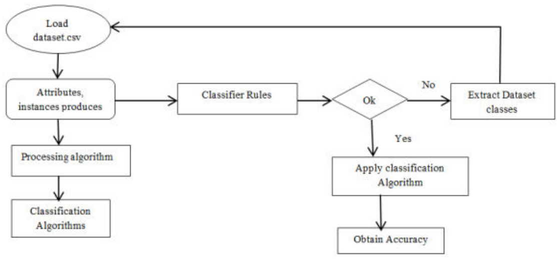


Fig 1: Procedure for the Implementation of an Algorithm

a) Random Forest Classification Algorithm

This algorithm is one of the most powerful in the field of data mining. This classification algorithm produces various CART-like trees. The outcome of this classifier is determined by a majority votes of the trees[1,10]. In the process of training the data, the Random Forest algorithm searches and determines only through a widely selected subsets of the input variables to regulate all the splits. In this process the total number of variables are completely user-defined, but this classification algorithm is not complex to it. The value by default is one of the set to the square root of the total number of inputs given. By selecting the total number of variables which are used for all the splits, the complexity of the algorithm becomes small, and the correlation among all the trees are also decreased. Then finally, the trees are not pruned [3,7].

b) Decision Stump

Let x = (x1, x2, …, xn)
Decision Stump hi,t
If xi ≥ t then class =1
else class = −1
Given data of the form x = (x1, x2, …, xn), one run of the training process defines the best hi,t.
Algorithm:

```
For each xi in the training set:
Step 1: Sort values then eliminate replicas.
Step 2: Build candidate thresholds t below min value, above max value and midway between successive features
Step 3: For every hi,t, calculate error on a training set.
Step 4: Return hi,t that maximizes . |1/2-error(h_(i,t) )|
Run Adaboost for T iterations, with L being the decision-stump learning algorithm is defined.
Step 5: Decision stump ht is learned using training data designated from current dissemination at time t.
Coefficient αt is considered by running ht on all training data.
Step 6: If a decision stump using feature xi is preferred on iteration, eliminate it from a group of features for the following iteration.
Step 7: Finally, next to T iterations, run ensemble classifier H on experimental data.
```

III. PROPOSED METHODOLOGY

We proposed an Upgraded RF classification algorithm with the future selection method, which will gives better performance when compared to other most popular algorithms. The algorithm is shown below.

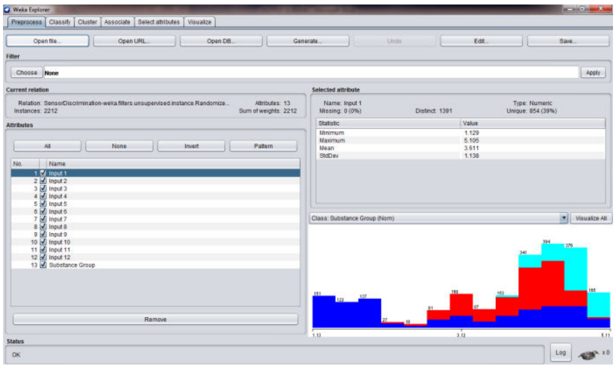
A. Upgraded RF Classification Algorithm

```
Input:
D: Give training data set,
A: Select feature space {A1, A2,...,AM},
Y: Select feature space {y1, y2,...,yq},
K: Total number of formed trees,
M: specific size of all subspaces.

Output: An Upgraded Random Forest μ

Technique:
Step 1: for i=1 to K do
Step 2: To appeal a bootstrap activity as a sample in-of-bag data subset IOBi and the out-of-bag data subset OOBi from all the training dataset (D);
Step 3:Therefore hi(IOBi) = createTree(IOBi);
Step 4: create a new node η using createTree( );
Step 5: if ending criteria is met then
Step 6: then return η as a leaf node in a tree;
Step 7: else for j=1 to j=M do
Step 8: Then calculate measure corr(Aj,Y);
Step 9: end for
Step 10: To compute feature weights {w1, w2,...,wM};
Step 11: To use the feature weighting technique in random wise and select m features;
Step 12: Then create the better split for the node to be segregated;
Step 13: Then call createTree() for each and every split;
Step 14: endif
Step 15: return η;
Step 16: Then it will use out-of-bag data subset OOBi for to analyze the out-of-bag exactness OOBAcci of the respective tree classifier hi(IOBi) by Equation;
Step 17: end for
Step 18: Then sort all the K tree classifiers in their OOBAcc in the form of descending order;
Step 19: Finally it will select the top most 80% trees with unlimited OOBAcc values and then it will combine the 80% tree classifiers into an upgraded RF as μ;
```

The algorithm performs with the input parameters which are selected for training data set. The actual performance of the algorithm depends on all the feature spaces, the respective class feature and the total number of trees in the random forest. And also it depends on the actual size of all the subspaces. The classification algorithm development is described with steps below. The Steps from 1 to 5 are the loop for building K decision trees. That means in the loop, from Step 2 to step 9 focuses on samples of training data with the bootstrap method to produce an in-of-bag data subset to construct a tree classifier, and generate an out-of-bag data subset for testing the tree classifier on out-of-bag accuracy. The Steps from 10 to 15 determines to call the function createTree( ) to build a new tree classifier. The Steps from 16 to 17 uses out-of-bag data subset to calculate the accuracy of out-of-bag for the tree classifier. The step 18 sorts all built tree with the help of classifiers in their out-of-bag accuracies in the level of descending order. The step 19 will selects top 80% of trees with high out-of-bag accuracy values. Then it will combines the 80% of tree classifiers into an Upgraded RF model. In this experiment the proposed method gives more than 80% of accuracy when compared to other algorithms.

IV. RESULTS

In this experimental research, we used the data mining tool for the classification of data with different parameters, to analyze the data and to determine the prediction accuracy of different classification algorithms. The classification with accuracy of various data mining algorithms has been analyzed and then we proposed a new algorithm called Upgraded Random Forest, it is one of the best fit algorithm to get better classification performance with good accuracy. In this research work, we proposed a technique, which is best suited to predict the classes given in Sensor Discrimination dataset. The trials has been performed with Sensor Discrimination dataset available on the UCI Machine Learning Repository. The selected dataset is used for analysis purpose, which is having a total number of 2212 instances and 12 attributes with one class attribute. The experimental result with the dataset is as follows.



Filter: None

Relation: SensorDiscrimination-weka.filters.unsupervised.instance.Randomize...
Instances: 2212
Attributes: 13
Sum of weights: 2212

No.		Name
1	☑	Input 1
2	☑	Input 2
3	☑	Input 3
4	☑	Input 4
5	☑	Input 5
6	☑	Input 6
7	☑	Input 7
8	☑	Input 8
9	☑	Input 9
10	☑	Input 10
11	☑	Input 11
12	☑	Input 12
13	☑	Substance Group

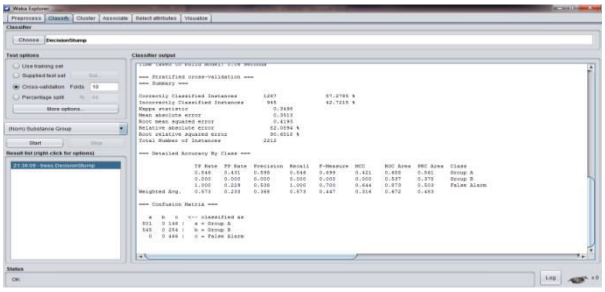
Name: Input 1
Missing: 0 (0%)
Distinct: 1391
Type: Numeric
Unique: 854 (39%)

Statistic	Value
Minimum	1.129
Maximum	5.105
Mean	3.611
StdDev	1.138

Class: Substance Group (Nom)

OK

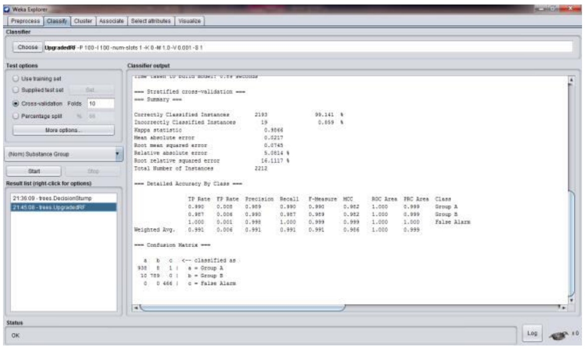
Fig 2: Pre-Processing of data



```
Classifier
DecisionStump

Test options
○ Use training set
○ Supplied test set
● Cross-validation  Folds 10
○ Percentage split  % 66
(Nom) Substance Group

Result list (right-click for options)
21:36:09 - trees.DecisionStump

Status
OK
```

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1267               57.2785 %
Incorrectly Classified Instances       945               42.7215 %
Kappa statistic                          0.3498
Mean absolute error                      0.3513
Root mean squared error                  0.4193
Relative absolute error                 82.0894 %
Root relative squared error             90.6518 %
Total Number of Instances             2212

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.846    0.431    0.595      0.846    0.699      0.421    0.688     0.561     Group A
                 0.000    0.000    0.000      0.000    0.000      0.000    0.537     0.378     Group B
                 1.000    0.228    0.538      1.000    0.700      0.644    0.873     0.503     False Alarm
Weighted Avg.    0.573    0.233    0.368      0.573    0.447      0.316    0.672     0.483

=== Confusion Matrix ===

   a   b   c   <-- classified as
 801   0 146 |   a = Group A
 545   0 254 |   b = Group B
   0   0 466 |   c = False Alarm
```

Fig 3: Result of the Decision Stump Algorithm



```
Classifier
UpgradedRF -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options
○ Use training set
○ Supplied test set
● Cross-validation  Folds 10
○ Percentage split  % 66
(Nom) Substance Group

Result list (right-click for options)
21:36:09 - trees.DecisionStump
21:45:08 - trees.UpgradedRF

Status
OK
```

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2193               99.141  %
Incorrectly Classified Instances        19                0.859  %
Kappa statistic                          0.9866
Mean absolute error                      0.0217
Root mean squared error                  0.0745
Relative absolute error                  5.0814 %
Root relative squared error             16.1117 %
Total Number of Instances             2212

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.990    0.008    0.989      0.990    0.990      0.982    1.000     0.999     Group A
                 0.987    0.006    0.990      0.987    0.989      0.982    1.000     0.999     Group B
                 1.000    0.001    0.998      1.000    0.999      0.999    1.000     1.000     False Alarm
Weighted Avg.    0.991    0.006    0.991      0.991    0.991      0.986    1.000     0.999

=== Confusion Matrix ===

   a   b   c   <-- classified as
 938   8   1 |   a = Group A
  10 789   0 |   b = Group B
   0   0 466 |   c = False Alarm
```

Fig 4: Result of the Upgraded Random Forest Algorithm

Table 1: Classification Performance of the Algorithms

Classification Algorithm	Kappa Statistics	MAE(Mean Absolute Error)	Percentage of In-Correctly Classified · Instances	Percentage of Correctly Classified Instances
Decision Stump	0.3498	0.3513	42.7215 %	57.2785 %
Upgraded RF	0.9866	0.0217	0.859 %	99.141 %

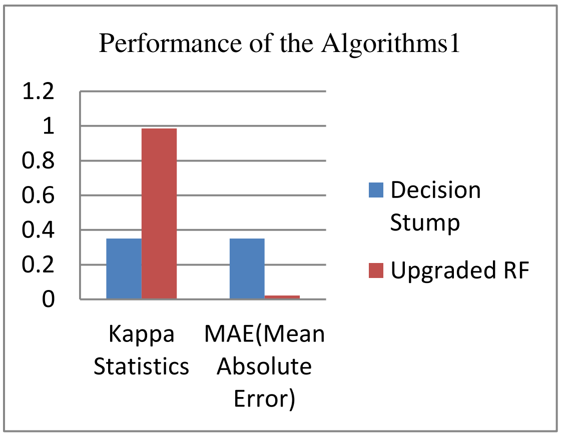


Fig 5: Graphical representation of Classification Performance of the Algorithms

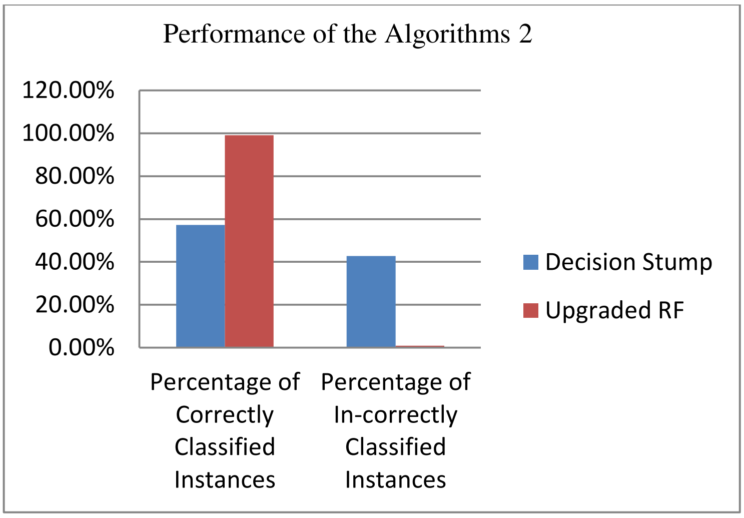


Fig 6: Graphical representation of Classification Performance of the Algorithms

V. CONCLUSION

In this paper, the new results proved that our Upgraded Random Forests classification algorithm is proposed a new method for to reduce the simplification error and to increase the test accuracy and performance of the the classification. Also, the time of computation is recorded to bring out the efficiency of the classifier. The results shows that the accuracy of Decision Stump classifier and performance is 57.2785 %, this is very less. But our Proposed Upgraded RF algorithm is applied on sensor discrimination dataset and evaluated using 10 folds cross-validation. Then the results shows 99.141% accuracy in classification with the Upgraded RF algorithm. Which means our proposed classification algorithm gives better classification accuracy with the lowest level of computational complexity. In future work we are going to use the real world datasets for the development of the performance of classification algorithms in data mining.

REFERENCES

[1]. Dr.K.Suresh Kumar Reddy, Dr.M.Jayakameswaraiah, Prof.S.Ramakrishna, Prof.M.Padmavathamma," Development Of Data Mining System To Compute The Performance Of Improved Random Tree And J48 Classification Tree Learning Algorithms", International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS), Volume.3, Special Issue.1, March.2017, Page 128-132, ISSN: 2454-356X

[2]. Dr.M.Jayakameswariah,Dr.K.Saritha,Prof.S.Ramakrishna,Prof.S.Jyothi,"Development of Data Mining System to Evaluate Performance Accuracy of J48 and Enhanced Naïve Bayes Classifiers using Car Dataset", International Journal Computational

Science, Mathematics and Engineering,SCSMB-16-March-2016,PP- 167-170,E-ISSN: 2349-8439.

[3]. G. Subbalakshami et al., “Decision Support in Heart Disease System using Naïve Bayes”, IJCSE, Vol. 2 No. 2, pp. 170-176, 2011, ISSN : 0976-5166

[4]. L. Breiman, “Random Forests,” Machine Learning, Vol. 40. No. 1. 2001.

[5]. R. Duda, P. Hart and D. Stork. Pattern Classification, 2nd edition. John Wiley, New York, 2001.

[6]. S. Ramya , Dr. N. Radha, “Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms”, International Journal of Innovative Research in Computer and Communication Engineering, pp. 812-820 Vol 4, issue 1, ISSN: 2320-9798, 2016.

[7]. S. Liu, R. Gao, D. John, J. Staudenmayer, and P. Freedson, “Multi-sensor data fusion for physical activity assessment,” IEEE Transactions on Biomedical Engineering, vol. 59, no. 3, pp. 687-696, March 2012.

[8]. Sunil Joshi and R. C. Jain., “A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database”, In proc of Second International Conference on Communication Software and Networks IEEE., p498-501. ISBN: 978-1-4244-5727-4, 2010.

[9]. T. Garg and S.S Khurana, “Comparison of classification techniques for intrusion detection dataset using WEKA,” In IEEE Recent Advances and Innovations in Engineering (ICRAIE), pp. 1-5, 2014.

[10]. V.Karthikeyani,”Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction” International Journal of Computer Applications (0975 – 8887) Volume 60– No.12, December 2012.