ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

Predicting lung disease severity evaluation and comparison of hybird decision tree algorithm

K.Rohini[1], Dr.G.Suseendran[2]

Ph.D. Research Scholar[1], Assistant Professor[2]
Department of Information Technology, School of Computing Sciences, Vels University, Chennai, India
rrohini16@gmail.com, suseendar_1234@yahoo.co.in

Abstract:

Objective: To focus on classification algorithms to arrive better prediction model for Lung Disease Severity.
Methods/Statistical analysis: In therapeutic analyses, the part of information mining methodologies is being expanded. Especially Classification calculations are exceptionally useful in arranging the information, which is critical for basic leadership prepare for therapeutic experts. In this paper the analysis is done in the WEKA apparatus on the spiro informational index.
Findings: The paper embarks to make relative assessment of classifiers, for example, J48, Random forest and proposed Hybird Decision Tree(HDT) Algorithm with regards to Spiro dataset to amplify genuine positive rate and limit false positive rate of defaulters as opposed to accomplishing just higher grouping exactness utilizing WEKA instrument. The tests comes about appeared in this paper are about grouping exactness, affectability and specificity.
Application/Improvements: The outcomes created on this dataset likewise demonstrate that the productivity and exactness of J48 is superior to anything other choice tree classifiers. J48 develops purge branches, it is the most urgent stride for govern era in J48. In more often than not this approach over fits the preparation cases with boisterous information. The proposed Hybird Decision Tree (HDT) Algorithm demonstrates great exactness in less time.
Keywords: Decision Tree, Pulmonary Function Test Means, Spirometry Data, Hybird Decision Tree Algorithm, J48 Algorithm

1. Introduction

Data mining is an interdisciplinary subfield of computer science which involves the extraction of hidden predictive information from large databases. It is widely used in various applications like analysis of organic compounds, medical diagnosis, product design, targeted marketing, credit card fraud detection, financial forecasting, automatic abstraction, predicting shares of television audiences etc. Data mining can be applied to any kind of information repository.A number of significant techniques such as pre processing, severity classification, clustering is performed in data mining using WEKA tool. In the previous paper k-means clustering and decision tree algorithm were used in spiromertry data for prediction of lung disease. We concluded with implementation with DBScan clustering algorithm for future work [1]. But it is preferably used for spatial data. So we proceed with classification techniques, which are widely used for medical databases.

Clustering involves identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. In this paper we focus upon classification to enhance prediction process and get better result. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

The Novel Risk Factors of Chronic Obstructive Pulmonary Disease on behalf of the Environmental and Occupational Health Assembly Committee on Non smoking COPD, Done their work on the parameters, such as genetics, air pollution, smoking and occupation. They added that the attributable fractions were comparatively very higher in industrialized countries compared with developing countries [2].

The technology for building knowledge-based systems by inductive inference from examples has been demonstrated successfully in several practical applications. This paper summarizes an approach to synthesizing decision trees that has been used in a variety of systems, and it describes one such system,D3, in detail. Results from recent studies show ways in which the methodology can be modified to deal with information that is noisy and/or incomplete. A reported shortcoming of the basic algorithm is discussed and two means of overcoming it are compared. The paper concludes with illustrations of current research directions [3].

They collected data for eight activities using a single tri axial accelerometer worn near the pelvic region. In this study, the performance of base-level classifiers and meta-level classifiers was compared. They found that combining classifiers using Plurality Voting provided the best overall results. Plurality Voting chooses the class that has been predicted by a majority of the base-level classifiers as the final predicted class. Of the base-level classifiers, the decision tree C4.5 performed the best[4].

Online activity recognition algorithms could be run on their Sensor But on miniaturised wireless sensor platform. They examined seven office worker activities such as drinking water and using a mouse via accelerometer and light sensor data. They used k-NN and J48/C4.5 classifiers with both providing 98% accuracy during off line evaluation. Accuracy dropped during online implementation due to floating point bit[5].

They Surveyed the execution of calculations in distinguishing twenty exercises under semi-naturalistic, recreated certifiable conditions utilizing five biaxial accelerometers. Choice table, occurrence based learning, choice tree (C4.5) and innocent Bayes classifiers were utilized with C4.5 giving the best execution perceiving ordinary exercises with a general precision of 84%[6].

2. Selection of decision tree classification

A decision tree can be utilized to outwardly and expressly speak to choices and basic leadership. Easy to comprehend and translate. Individuals can comprehend choice tree models after a concise clarification. Trees can likewise be shown graphically in a way that is simple for non-specialists to translate. It is ready to deal with both numerical and nominal information. Different systems are typically represented considerable authority in investigating datasets that have just a single sort of factor. It requires little information readiness. Different strategies frequently require information standardization. Since trees can deal with subjective indicators, there is no compelling reason to make sham variables

It Mirrors human basic leadership more nearly than different methodologies. A decision tree can be utilized to outwardly and expressly speak to choices and basic leadership. Easy to comprehend and decipher. Individuals can comprehend choice tree models after a concise clarification. Trees can likewise be shown graphically in a way that is simple for non-specialists to decipher. Decision trees are capable arrangement calculations that are ending up plainly progressively more well-known with the development of information mining in the field of data frameworks. Well known choice tree calculations incorporate. This procedure recursively isolates perceptions in branches to build a tree with the end goal of enhancing the expectation precision. ID a variable and comparing edge for the variable that parts the information perception into at least two subgroups. This progression is rehashed at each leaf hub until the entire tree is developed. The goal of the part calculation is to locate a variable-edge combine that amplifies the homogeneity (request) of the subsequent at least two subgroups of tests. A choice tree is a stream graph like structure, where each inside (non-leaf) hub means a test on a trait, each branch speaks to the result of a test, and each leaf (or terminal) hub holds a class name. The highest hub in a tree is the root hub.

3. PǊƶLJƶăŜŘ YŜűKƶŘƶfƶƎè

3.1. Experimental study and analysis

We utilize WEKA an open source information digging apparatus for our investigation. WEKA is created by the University of Waikato in New Zealand that actualizes information mining calculations utilizing the JAVA dialect. WEKA is a best in class device for creating machine learning (ML) methods and their application to certifiable information mining issues. It is an accumulation of machine learning calculations for information mining undertakings. The calculations are connected straightforwardly to a dataset. WEKA actualizes calculations for information preparing, include lessening, order, relapse, grouping, and affiliation rules. It additionally incorporates representation devices. The new machine learning calculations can be utilized with it and existing calculations can likewise be stretched out with this device.

3.2. Dataset description

We performed computer simulation on an UCI Machine Learning Repository dataset heart-h.arff and on our available Spiro dataset. The feature Spiro dataset describes different factors which influences the obstructive and restrictive lung diseases and its severity. The Spiro dataset is got from processed data of three files related to 2000 raw data elements got from Asthma and Allergy Resource Centre, Tamil Nadu, India. This dataset contains 1003 instances and 15 attributes. The resultant key attribute values of three data files are taken for consideration for final classification purpose.

The data set consists

- ✓ Patient identification number(PID)
- ✓ Report attribute that contains the result of spirometry data(PFT Report)
- ✓ Environmental data details(Env attribute)
- ▪ Living environment
- ▪ Working environment
- ▪ Smoking
- ✓ Parental history(genetics) –Boolean. The possible values of the resultant key attributes and the sample dataset are given in Table1 and Table 2 respectively.

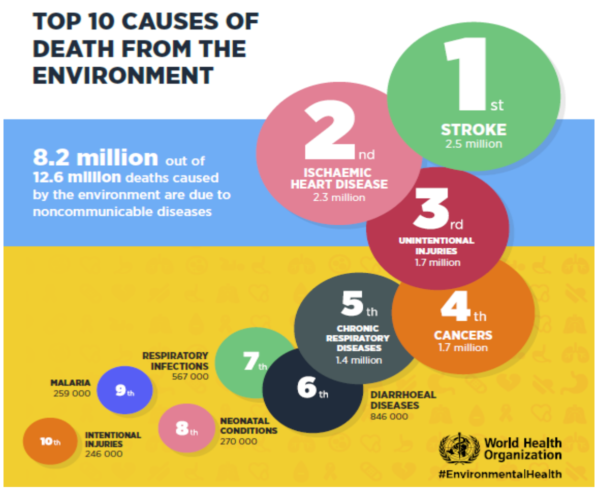
Table 1.Possible Values of the attributes

Env	Paren	Report
GOOD	Yes	RE1
BAD	N0	ob1
MOD		Nor
MGOOD		RE1ob1
MBAD		Re2
		ob2
		Re2ob2
		RE1ob2
		Re2ob1

Table 2.Sample dataset

Env	Paren	Report
MOD	Yes	RE1
MBAD	No	ob1
GOOD	No	Nor
MOD	No	RE1ob1
MBAD	No	RE1
MOD	Yes	RE1
GOOD	Yes	RE1
BAD	No	Re2ob2
MOD	Yes	RE1
MBAD	Yes	RE1ob2
MGOOD	No	Nor
BAD	Yes	Re2
BAD	No	RE1ob2
MGOOD	Yes	Ob2

Figure 1.PHE health topics



The three key parameters play a big role in the severity of lung diseases. The strength of evidence on “smoking is a cause of COPD” has been growing for more than 40 years and has been extensively reviewed in three U.S. Surgeon General’s Reports. The1984 Surgeon General Report concluded that 80–90% of COPD in United States is attributable to smoking the estimated fraction of COPD mortality attributable to smoking was around 53% for men and were around 20% for women. The attributable fractions were comparatively very higher in industrialized countries compared with developing countries. The genetics, air pollution, smoking, occupation are attributes selected by researcher, as these are the factors that greatly influences the lung diseases [7].

In Novel Risk Factors of Chronic Obstructive Pulmonary Disease the Environmental and Occupational Health Assembly Committee on Non smoking COPD, estimated 12.6 million deaths each year are attributable to unhealthy environments - nearly one in four of total global deaths. Environmental risk factors, such as air, water and soil pollution, chemical exposures, climate change and ultraviolet radiation, contribute to more than 100 diseases and injuries[9]. An estimation of WHO on the topic “Burden of disease from Household Air Pollution for 2012 “ stated that Globally, 4.3 million deaths were attributable to household air pollution (HAP) in 2012, almost all in low and middle income (LMI) countries. The South East Asian and Western Pacific regions bear most of the burden with 1.69 and 1.62 million deaths, respectively. Some 3 million deaths a year are linked to exposure to outdoor air pollution. Indoor air pollution can be just as deadly. In 2012, an estimated 6.5 million deaths (11.6% of all global deaths) were associated with indoor and outdoor air pollution together, as per the WHO health topics of (PHE) Public Health, Environmental and Social Determinants of Health. The info graphic image in Figure 1 depicts that, out of 12.6 million deaths 8.2 million deaths are caused due to environment factors and also states that the 5th largest cause for this fatal situation is chronic respiratory diseases [8].

3.3A. Decision tree algorithm J48:

J48 classifier is a basic C4.5 choice tree for grouping. At every hub of the tree, J48 picks the quality of the information that most adequately parts its arrangement of tests into subsets improved in one class or the other. The part standard is the standardized data pick up (distinction in entropy). The property with the most astounding standardized data pick up is settled on the choice. The J48 calculation then repeats on the littler sub records.

This calculation has a couple base conditions. On the off chance that All the examples in the rundown have a place with a similar class, then it just makes a leaf hub for the choice tree saying to pick that class. In the event that None of the elements give any data increase, Then J48 makes a choice hub higher up the tree utilizing the normal estimation of the class. In the event that Instance of beforehand concealed class experienced, Then J48 makes a choice hub higher up the tree utilizing the normal esteem[10].

In pseudo code, the general algorithm for building decision trees is:

```
1. Check for the above base conditions.
2. For each attribute A find the normalized information gain ratio from splitting on a.
3. Let A_best be the attribute with the highest normalized information gain.
4. Create a decision node that splits on A_best.
5. Recur on the sublists obtained by splitting on A_best, and add those nodes as children of node.
```

It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple.

Algorithm enhanced J48:

```
INPUT:
D //Training data
OUTPUT
T //Decision tree
DTBUILD (*D)
{
T=ϕ;
T= Create root node and label with splitting attribute;
T= Add arc to root node for each split predicate and
label;
For each arc do
D= Database created by applying splitting
predicate to D;
If stopping point reached for this path, then
T'= create leaf node and label with
appropriate class;
Else
T'= DTBUILD(D);
T= add T' to arc;
}
```

J48 is an open source java implementation of the C4.5 algorithm in the Weka data mining tool.J48 implements both C4.5's confidence-based post- pruning (default) and sub-tree rising. Decision trees are more likely to face problem of Data over-fitting. This problem arises when the algorithm splits the data until it make pure sets. This Problem of Data over-fitting is fixed in its extension that is J48 by using Pruning. Pruning trees after creation –J48 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

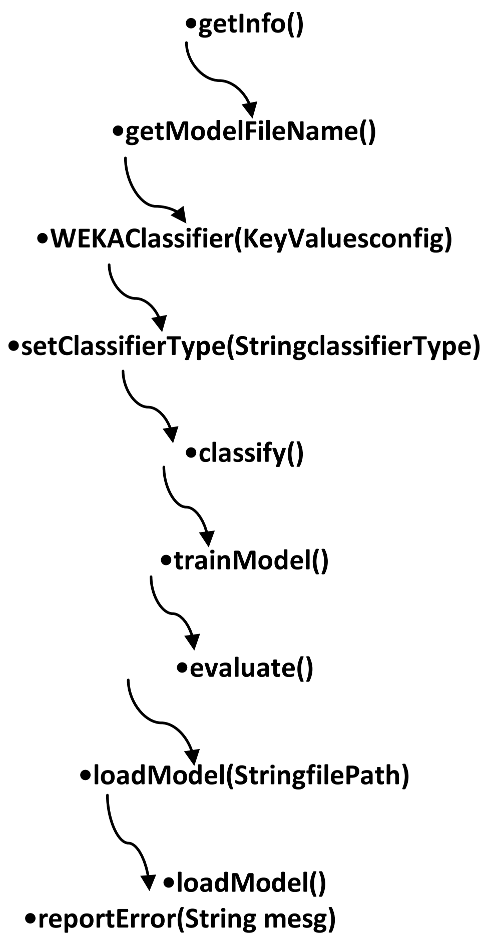
Draw Backs: J48 develops exhaust branches, it is the most critical stride for govern era in J48. We have discovered numerous hubs with zero values or near zero qualities. These qualities neither add to produce rules nor help to develop any class for characterization errand. Or maybe it makes the tree greater and more unpredictable. Over fitting happens when calculation demonstrate grabs information with exceptional qualities. For the most part J48 calculation develops trees and develops it branches 'sufficiently profound to impeccably order the preparation cases'. This technique performs well with commotion free information. Yet, more often than not this approach over fits the preparation cases with loud information. As of now there are two methodologies are generally utilizing to sidestep this over fitting in choice tree learning [11].

The proposed enhanced hybrid decision tree algorithm uses improved split code to find the best attribute to split and proceed further until pure sets formation [12]. The split code is responsible for finding the best split in node so as to get the pure set easily. Here in this paper the classification accuracy of Naïve bayes,J48 and proposed enhanced-J48 algorithms are comparatively analyzed.The splitting criteria for J48 algorithm is gain ratio. K-fold Cross validation is used for Validating the Model.

Pseudo code for Proposed Hybird Decision Tree with modified Model Selection code(classifier split model)

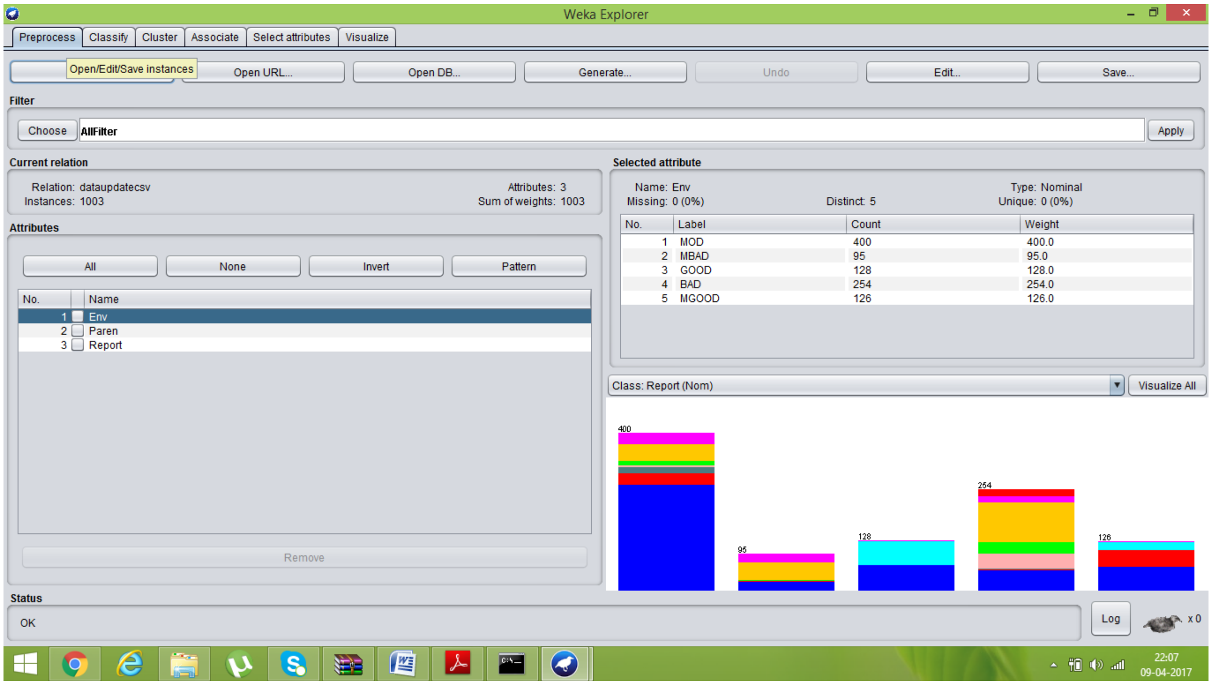
Step1: Abstract class for classification models that can be used recursively to split the data.
Step2: Builds the classifier split model(hybrid split model) for the given set of instances.
Step3: Checks if generated hybrid split model is valid.
Step4: Prints left side of condition satisfied by instances.
Step5: Prints left side of condition satisfied by instances in subset index
Step6: Returns weights if instance is assigned to more than one subset
Step7: Returns index of subset instance is assigned

Outline : architecture flow of proposed HDT(Hybird Decision Tree) Algorithm



- getInfo()
- getModelFileName()
- WEKAClassifier(KeyValuesconfig)
- setClassifierType(StringclassifierType)
- classify()
- trainModel()
- evaluate()
- loadModel(StringfilePath)
- loadModel()
- reportError(String mesg)

Figure 2. The Attributes of Spiro Dataset.



Filter: AllFilter

Current relation

Relation: dataupdatecsv
Instances: 1003
Attributes: 3
Sum of weights: 1003

Attributes

No.	Name
1	Env
2	Paren
3	Report

Selected attribute

Name: Env
Missing: 0 (0%)
Distinct: 5
Type: Nominal
Unique: 0 (0%)

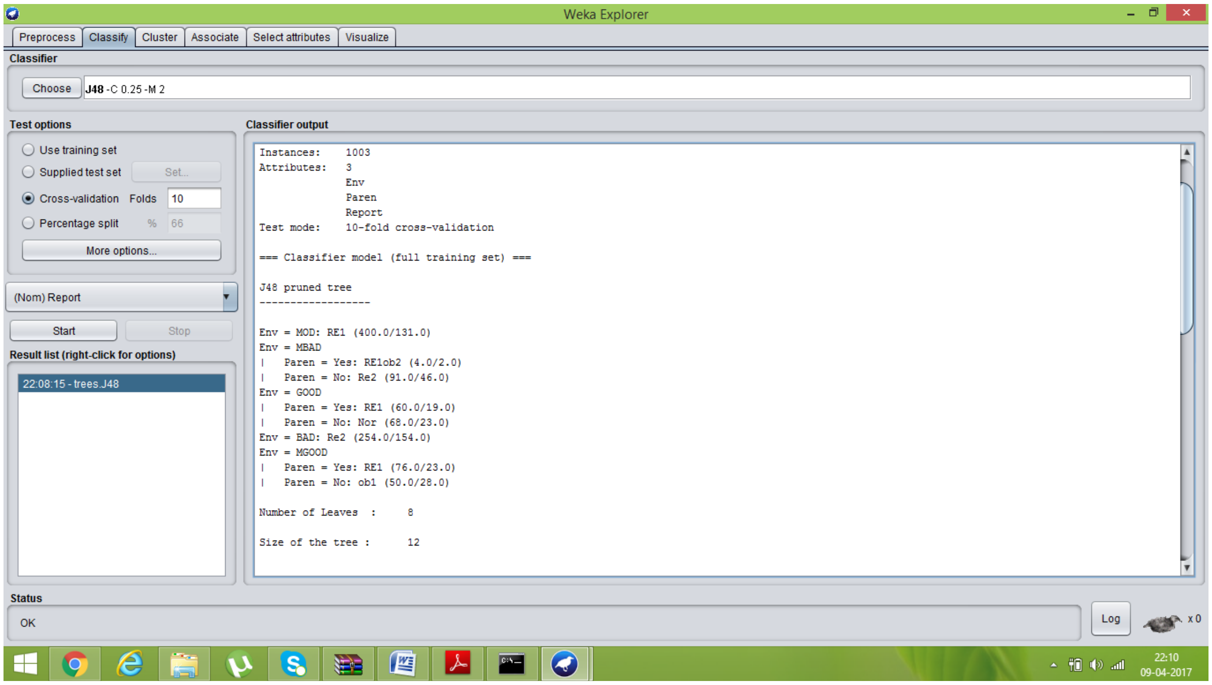
No.	Label	Count	Weight
1	MOD	400	400.0
2	MBAD	95	95.0
3	GOOD	128	128.0
4	BAD	254	254.0
5	MGOOD	126	126.0

Class: Report (Nom)

Status

OK

Figure 3. J48 Pruned Tree Classification Description



Classifier: J48 -C 0.25 -M 2

● Cross-validation Folds 10

(Nom) Report

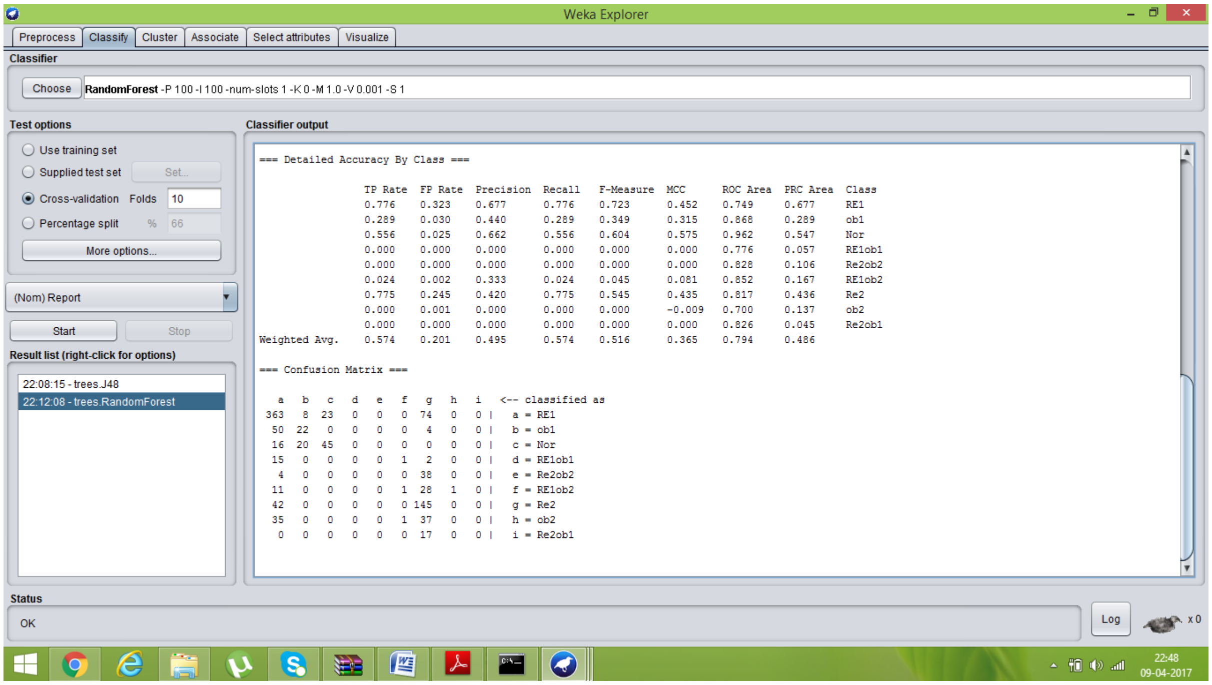
Result list (right-click for options): 22:08:15 - trees.J48

Status: OK

Classifier output

```
Instances:    1003
Attributes:   3
              Env
              Paren
              Report
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Env = MOD: RE1 (400.0/131.0)
Env = MBAD
|   Paren = Yes: RE1ob2 (4.0/2.0)
|   Paren = No: Re2 (91.0/46.0)
Env = GOOD
|   Paren = Yes: RE1 (60.0/19.0)
|   Paren = No: Nor (68.0/23.0)
Env = BAD: Re2 (254.0/154.0)
Env = MGOOD
|   Paren = Yes: RE1 (76.0/23.0)
|   Paren = No: ob1 (50.0/28.0)

Number of Leaves  :     8

Size of the tree :      12
```

Figure 4.Random Forest Classification



Classifier: RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

● Cross-validation Folds 10

(Nom) Report

Result list (right-click for options): 22:08:15 - trees.J48; 22:12:08 - trees.RandomForest

Status: OK

Classifier output

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.776    0.323    0.677      0.776    0.723      0.452    0.749     0.677     RE1
                 0.289    0.030    0.440      0.289    0.349      0.315    0.868     0.289     ob1
                 0.556    0.025    0.662      0.556    0.604      0.575    0.962     0.547     Nor
                 0.000    0.000    0.000      0.000    0.000      0.000    0.776     0.057     RE1ob1
                 0.000    0.000    0.000      0.000    0.000      0.000    0.828     0.106     Re2ob2
                 0.024    0.002    0.333      0.024    0.045      0.081    0.852     0.167     RE1ob2
                 0.775    0.245    0.420      0.775    0.545      0.435    0.817     0.436     Re2
                 0.000    0.001    0.000      0.000    0.000      -0.009   0.700     0.137     ob2
                 0.000    0.000    0.000      0.000    0.000      0.000    0.826     0.045     Re2ob1
Weighted Avg.    0.574    0.201    0.495      0.574    0.516      0.365    0.794     0.486

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 363   8  23   0   0   0  74   0   0 |   a = RE1
  50  22   0   0   0   0   4   0   0 |   b = ob1
  16  20  45   0   0   0   0   0   0 |   c = Nor
  15   0   0   0   0   1   2   0   0 |   d = RE1ob1
   4   0   0   0   0   0  38   0   0 |   e = Re2ob2
  11   0   0   0   0   1  28   1   0 |   f = RE1ob2
  42   0   0   0   0   0 145   0   0 |   g = Re2
  35   0   0   0   0   1  37   0   0 |   h = ob2
   0   0   0   0   0   0  17   0   0 |   i = Re2ob1
```

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

Figure 5.Decision Tree for J48 Algorithm

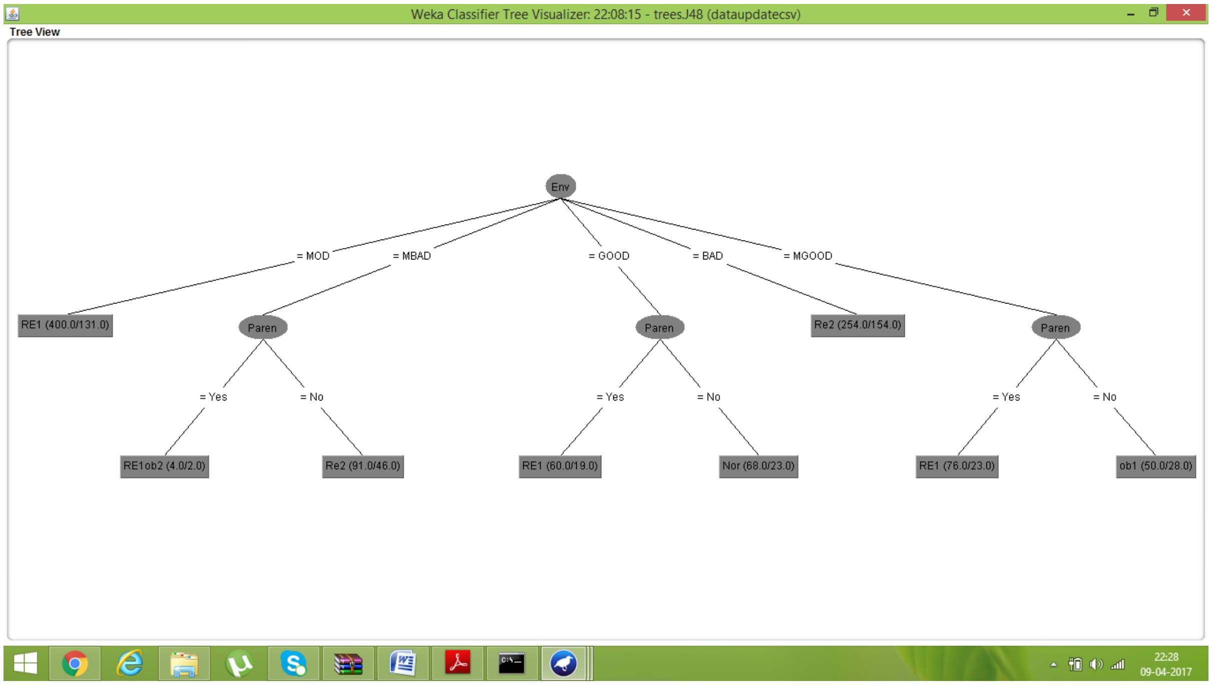
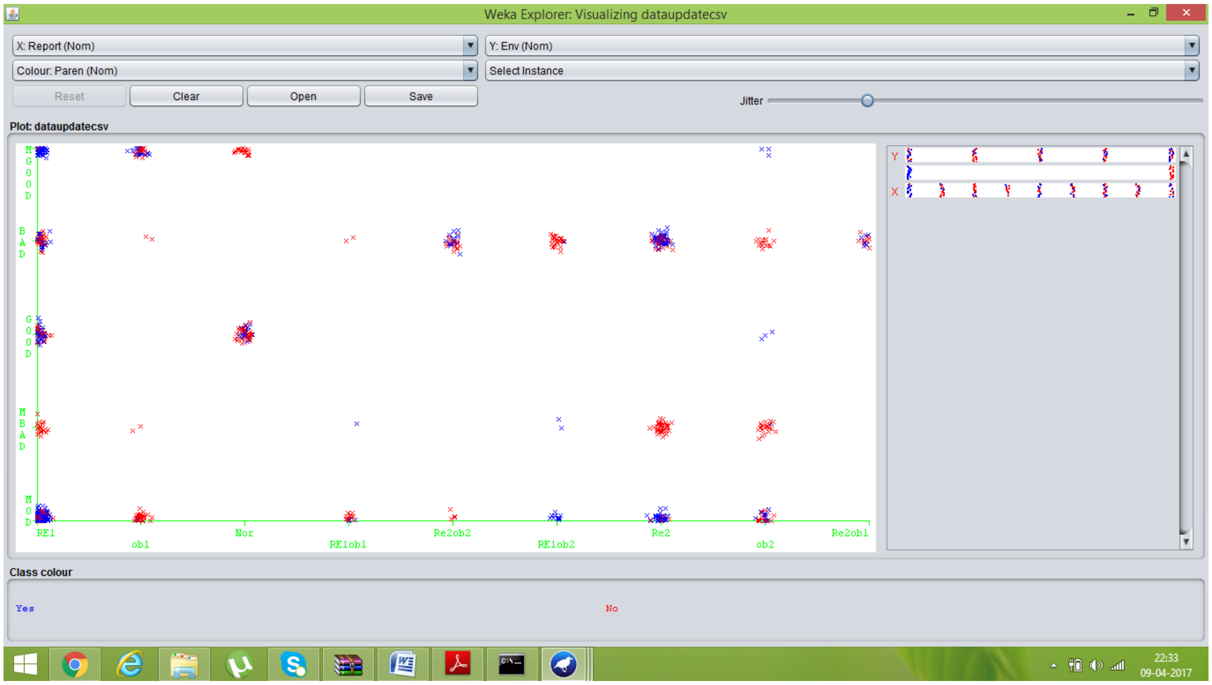


Figure 6. Plot Matrix



Weka Explorer: Visualizing dataupdatecsv

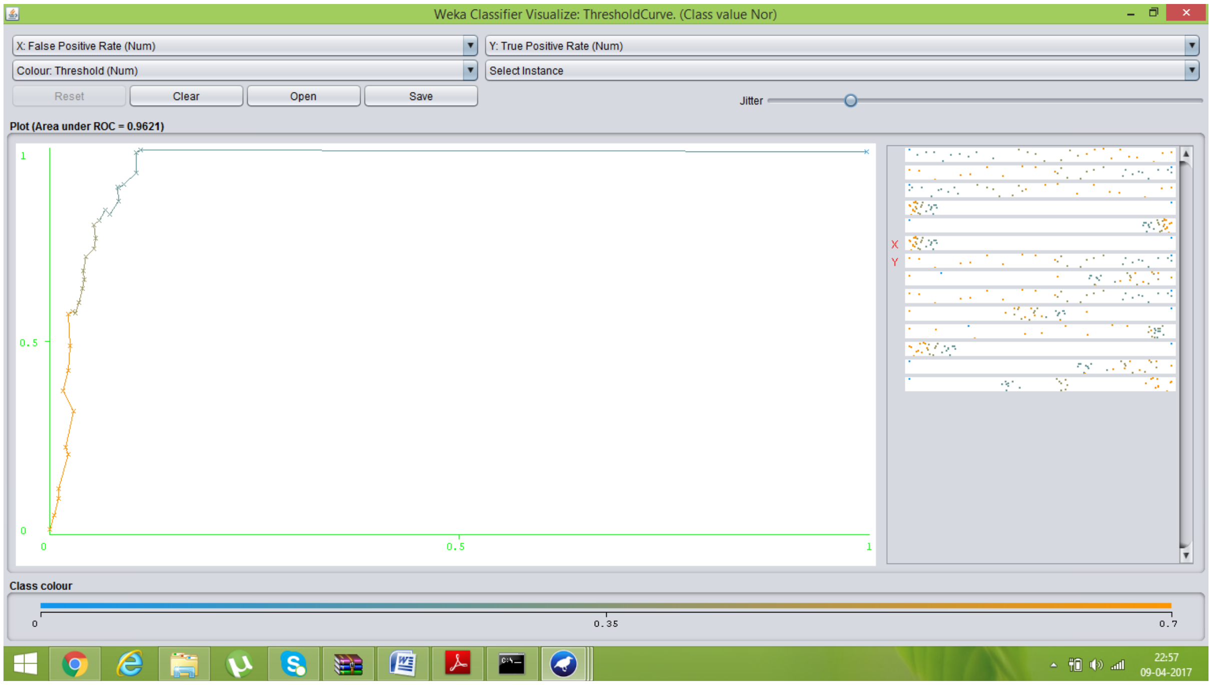
X: Report (Nom)

Y: Env (Nom)

Colour: Paren (Nom)

Select Instance

Figure 7. Threshold Curve



Weka Classifier Visualize: ThresholdCurve. (Class value Nor)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.9621)

After the implementation of the spiro dataset in J48 algorithm and Random Tree in Weka tool, we compared the results along with the HDT algorithm. Figure 2 Shows The Attributes of Spiro Dataset.Figure3 Shows the Attributes of Spiro Dataset while theJ48 Pruned Tree Classification Description is shown in Figure4. Figure5 depicts the Random Forest Classification and Figure6 explains the Decision Tree For J48 Algorithm. The plot matrix is as shown in the Figure7. At last the Threshold Curve shows the false positive rate ,true positive rate and Roc curve area.

We focus on the decision tree classification especially for enhanced J48 algorithm were the classification accuracy and the performance measure of the classifier algorithms based upon TP rate and FP rate, Metrics such as

- Information gain is based on the concept of entropy from information theory.
- Confidence Factor — The confidence factor used for pruning (smaller values incur
- more pruning).
- minNumObj — The minimum number of instances per leaf.
- The results of enhanced -J48 is Compared along with other classifying algorithms like Random Tree in Weka tool [13].

4. Conclusion

In this paper we have concentrated the different essential properties of the choice tree calculations which give as a superior comprehension of these calculations. Classifiers, for example, C4.5, J48 and AD tree are generally utilized as a part of the medicinal informational collection. In this paper the analysis is completed in the WEKA apparatus on the spiro informational index. The examinations comes about appeared in this paper are about characterization exactness, affectability and specificity. The results in the paper on this dataset also show that the efficiency and accuracy of J48 is better than other decision tree classifiers, but except for the time taken to build the model, construction of empty branches, over fitting of the training examples with noisy data. The proposed Hybird Decision Tree(HDT) Algorithm shows good accuracy in less time. We can apply them on different types of data sets having different types of values and properties and can attain a best result.

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

5. References

1. K.Rohini, G. Suseendran. Aggregated k means clustering and decision tree. *Indian Journal of Science and Technology*.2016; 9(44), 1-6.
2. Gareth James, Witten Daniela, Hastie Trevor, Tibshirani Robert. An Introduction to Statistical Learning. *New York: Springer*. 2015, 315.
3. J. Quinlan. Induction of decision trees. *Machine Learning*. 1986; 1, 81—106.
4. J. Quinlan. C4.5: programs for machine learning. *San Mateo, CA: Morgan Kaufmann*; 1993.
5. L. Breiman, J. H. Friedman, R.A. Olshen, C.J. Stone. Classification and regression trees. *Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software*. 1984.
6. M.D. Eisner, Nicholas Anthonisen, David Coultas, Nino Kuenzli, Rogelio Perez-Padilla, Dirkje Postma, Isabelle Romieu, E.K. Silverman, J.R. Balmes. Novel risk factors and the global burden of chronic obstructive pulmonary disease(on behalf of the environmental and occupational health assembly committee on nonsmoking COPD) .*American Journal Of Respiratory And Critical Care Medicine*. 2010; 182, 1-26.
7. M.H. Danham, S. Sridhar. Data mining, Introductory and Advanced Topics. *Pearson education*. 1st edition. 2006.
8. WHO - PHE health topics-2016. http:// www.who.int/mediacentre/news/releases/2016/air-pollution-estimates/en/. Date Accessed: 27/09/2016.
9. Wenke Lee, S.J. Stolfo, K.W. Mok. A data mining framework for building intrusion detection models. Security and Privacy.1999, Proceedings of the 1999 IEEE Symposium on. 1999.
10. Aman Kumar Sharma, Suruchi Sahni. A comparative study of classification algorithms for spam email data analysis. IJCSE. 2011; 3(5), 1890-1895.
11. J.R. QUINLAN (munnari! nswitgould.oz! quinlan@ seismo.css.gov) on induction of decision trees, centre for advanced computing sciences. *New South Wales Institute of Technology*, *Sydney*. 2007.
12. Nishkam Ravi, Nikhil Dadekar, Preetham Mysore, M.L. Littman. Activity recognition from accelerometer data. American Association for Artificial Intelligence, 2005.
13. Clemens Lombriser, N.B. Bharatula, Daniel Roggen, Gerhard Tr¨oster. On-body activity recognition in a dynamic sensor network. In BodyNets ’07: *Proceedings of the ICST 2nd international conference on Body area networks*. *ICST, Brussels, Belgium*. 2007, 1-6.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:
K.Rohini, Dr.G.Suseendran. Predicting lung disease severity evaluation and comparison of hybird decision tree algorithm. *Indian Journal of Innovations and Developments*. 2017; 6(1), January.