INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

International Journal of Current Pharmaceutical Research

ISSN- 0975-7066 Vol 9, Issue 2, 2017

Review Article

PERFORMANCE ANALYSIS OF BREAST CANCER CLASSIFICATION USING DECISION TREE CLASSIFIERS

P. HAMSAGAYATHRI, P. SAMPATH

Department of ECE, Bannari Amman Institute of Technology, Sathyamangalam, Erode
Email: palanisamy.hamsagayathri@gmail.com

Received: 27 Nov 2016, Revised and Accepted: 23 Jan 2017

ABSTRACT

Breast cancer is one of the dangerous cancers among world's women above 35 y. The breast is made up of lobules that secrete milk and thin milk ducts to carry milk from lobules to the nipple. Breast cancer mostly occurs either in lobules or in milk ducts. The most common type of breast cancer is ductal carcinoma where it starts from ducts and spreads across the lobules and surrounding tissues. According to the medical survey, each year there are about 125.0 per 100,000 new cases of breast cancer are diagnosed and 21.5 per 100,000 women due to this disease in the United States. Also, 246,660 new cases of women with cancer are estimated for the year 2016. Early diagnosis of breast cancer is a key factor for long-term survival of cancer patients. Classification plays an important role in breast cancer detection and used by researchers to analyse and classify the medical data. In this research work, priority-based decision tree classifier algorithm has been implemented for Wisconsin Breast cancer dataset. This paper analyzes the different decision tree classifier algorithms for Wisconsin original, diagnostic and prognostic dataset using WEKA software. The performance of the classifiers are evaluated against the parameters like accuracy, Kappa statistic, Entropy, RMSE, TP Rate, FP Rate, Precision, Recall, F-Measure, ROC, Specificity, Sensitivity.

Keywords: Classification, J48, REPTree, Random Forest, Random Tree, priority, Accuracy

DOI: http://dx.doi.org/10.22159/ijcpr.2017v9i2.17383

INTRODUCTION

Breast cancer is the second leading cancer among the women worldwide. The occurrence of breast cancer is increasing every year by year, due to heredity, increase life expectancy, different lifestyles and food habits. The genuine motivation of this research is to build the classification model to classify the breast cancer and to provide the accurate diagnosis to physicians to provide effective treatment to save a life. Thus, efficient classification model increases the mortality of the women. Currently, we have different techniques like X-ray Mammogram, Ultrasound, Magnetic resonance imaging (MRI), Biopsy, Positron Emission Tomography (PET), etc to evaluate cancer in humans. Though we have different techniques; diagnosis is made by the experienced physicians. When compared to a physician, machine learning diagnosis is more correct, and it is approximated with an accuracy of 91.1% [1].

Thus, usage of machine learning classifier systems in medical diagnosis is increased. The classifier algorithms help experienced/ inexperienced physicians to diagnosis accurately by minimising possible errors. The most common classifier algorithm used to classify medical data is J48 decision tree. The main advantages of decision tree algorithms are

- Flexible
- Easy to build
- Easy to debug
- Applicable for numerical and categorical values
- Suits for classification and regression

The serious drawbacks of the decision tree algorithm are

- Overfitting
- Complexity
- Cost
- Memory
- Computation time

There are various methods such as Boosting/Bagging to ensemble various classifiers and to provide the efficient classification. Though, we have different methods to provide discriminative classification but with increased cost and complexity. In our proposed method, priority is set for various attributes in the dataset. Therefore, the priority of the attributes is also considered along with the information gain during classification.

Research objective

The objective of this research is to undergo a comparative study on various decision tree classifier algorithms and to identify the best classifier for Breast cancer classification of Wisconsin Original dataset.

Research scope

The scope of the research is to apply the classifier algorithms such as J48, REPTree, Random Forest, RandomTree and Priority based decision tree classifier on Wisconsin Breast cancer dataset. Data cleaning and reduction are performed for further classification. The comparative study on these classifiers includes classification accuracy, True Positive rate, False Positive Rate, Precision, Recall, ROC, PRC, Sensitivity, Specificity, and RMSE as performance metrics.

This paper is categorised as follows. Section 2 gives a brief description on classification algorithms that are used to classify the data and section 3 provides the detailed description on datasets and discussed the simulation results that are obtained for various decision tree algorithms.

Methodology

Classification is one of the most extensively used decision-making task in machine-based learning algorithms. The main objective of the classification is to accurately predict the target class for each instance in the data. In training phase of classification, each instance of the data has predefined target class. Whereas in testing phase unknown test instances are predicted using the model builds with the training set. Classification algorithms process a huge volume of data and classify data based on the training set. Classifications algorithms process a huge volume of data and classify data based on the training set. The analysis of classification process flow is depicted below fig. 1.

Data pre-processing precede classification to improve the quality of the data. There are several methods of pre-processing, but whereas we consider data cleaning and data reduction techniques.

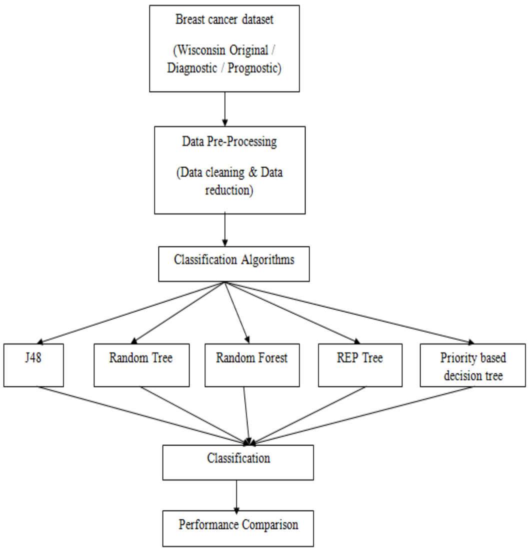

Breast cancer dataset
(Wisconsin Original / Diagnostic / Prognostic)
Data Pre-Processing
(Data cleaning & Data reduction)
Classification Algorithms
J48
Random Tree
Random Forest
REP Tree
Priority based decision tree
Classification
Performance Comparison

Fig. 1: Methodology for data classification

Data cleaning

Data Cleaning pre-processes the data to handle missing values of attributes. Missing values are replaced by the mean value for that attribute.

Data reduction

The feature selection techniques are used to reduce the dimensionality of the data. Feature selection technique removes the irrelevant and redundant attributes from the dataset that has less significance in the classification. In Priority based decision tree classifier algorithm priorities are set based on the rank of information feature selection technique.

Table 1: Dataset attributes of Wisconsin (Original)

Attribute	Values
Sample code number	1–10
Clump Thickness	1–10
Uniformity of Cell Size	1–10
Uniformity of Cell Shape	1–10
Marginal Adhesion	1–10
Single Epithelial Cell Size	1–10
Bare Nuclei	1–10
Bland Chromatin	1–10
Normal Nucleoli	1–10
Mitoses	1–10
Class	(2 for benign, 4 for malignant)

Table 2: Dataset attributes of Wisconsin (Diagnostic)

Attribute	Values
Id Number	Numeric
Diagnosis	M = malignant, B = benign
Radius (mean, standard error and worst)	Numeric
texture (mean, standard error and worst)	Numeric
perimeter (mean, standard error and worst)	Numeric
area (mean, standard error and worst)	Numeric
smoothness (mean, standard error and worst)	Numeric
compactness (mean, standard error and worst)	Numeric
concavity (mean, standard error and worst)	Numeric
concave points (mean, standard error and worst)	Numeric
symmetry (mean, standard error and worst)	Numeric
fractal dimension (mean, standard error and worst)	Numeric

Table 3: Dataset attributes of Wisconsin (Prognostic)

Attribute	Values
Id Number	Numeric
Outcome	R = recur, N = no recur
Time	recurrence time if field 2 = 'R', disease-free time if field 2= 'N'
Radius (mean, standard error and worst)	Numeric
texture (mean, standard error and worst)	Numeric
perimeter (mean, standard error and worst)	Numeric
area (mean, standard error and worst)	Numeric
smoothness (mean, standard error and worst)	Numeric
compactness (mean, standard error and worst)	Numeric
concavity (mean, standard error and worst)	Numeric
concave points (mean, standard error and worst)	Numeric
symmetry (mean, standard error and worst)	Numeric
fractal dimension (mean, standard error and worst)	Numeric

Classification algorithms

There have been various algorithms used for classification of Breast cancer. This paper provides the detailed description on decision tree algorithms and evaluates based on the performance measures like accuracy, sensitivity, specificity, entropy, ROC, PR area, complexity, the size of the decision tree, computation time and so on.

J48 algorithm

The J48 classifier is the extension of decision tree ID3 algorithm with additional features like accounting for missing values, reduced error pruning, continuous attribute value, and derivation of rules and so on. A decision tree is a supervised technique builds the classification in tree-like structure with the root node, branch node and leaf node. Decision tree breaks down the entire dataset into multiple subsets and builds the decision tree incrementally. J48 employs top-down and greedy search through all possible branches to construct a decision tree.

The algorithm

- Initially, all the training data are at root
- Input data are partitioned based on the select attributes
- Entropy and Information gain are calculated. Attribute with highest information gain are selected as decision node
- Branch with zero entropy is marked as leaf node in the decision tree
- Branch with non-zero entropy undergo further partition
- Algorithm runs recursively on non-leaf nodes until all the data is classified

Condition for stopping

- All the sample at the given node belong to the same class
- No remaining attributes for further partitioning
- No samples left

REPTree algorithm

REPTree is one of the fast decision tree classifier algorithms. It constructs the decision tree using entropy and information gain of the attribute with reduced error pruning technique. It constructs multiple trees and selects the best tree from the generated list of trees. REPTree prunes the tree using the back fitting method.

REPTree algorithm sorts all numeric fields in the dataset only once at the start, and then it utilize the sorted list to split the attributes at each tree node. It classifies the numeric attributes by minimising total variance. The non-numeric attributes classified with regular decision tree with reduced error pruning technique.

The algorithm

- Load input data
- Build multiple trees using entropy and information gain

If (numeric attributes)

Sort all numeric fields

Construct decision tree with sorted list

Else

Construct decision tree with error-pruning

- Choose the best tree from constructed list

Random forest algorithm

Random Forest is one of the most accurate machine learning algorithms. It is capable of handling thousands of attributes without any feature selection. It provides the estimates of the important attributes. It is a highly efficient algorithm for estimating the missing data, and it also maintain the accuracy in estimation. It can handle a large volume of the database. Multiple trees are constructed to choose the best tree on the split. When compared to REPTree, error pruning is not performed in Random Forest.

The algorithm

- Initialize N= Number of training cases and M = Number of variables in the classifier
- Let m = Number of input variables.
- Recursively build decision tree
- Check m<M to determine the decision node
- Choose 'n' cases with replacement from 'N' available training cases.
- Estimate the error of the tree
- Select the tree with majority vote

Random tree algorithm

Random tree classifier is one of the decision tree approaches where the 'K' attributes are chosen randomly to classify the data. It does not contains any pruning technique to minimise the error. Random tree algorithm has an option to estimate the class probabilities for classification.

The algorithm

- Load training data at the root
- Input data are partitioned based on the 'K' attributes randomly
- Construct decision tree with random split
- Algorithm runs recursively on non-leaf nodes until all the data is classified

Priority based decision tree algorithm

Though, J48 decision tree is simple, easy to construct and human readable format. It has high computational time and cost. Also, have repetitive sub trees with post pruning. The limitations of the J48 algorithm is overcome by prioritising the attributes by the user for decision tree node split.

Priority based decision tree is one of the fast decision tree classifier algorithms. It constructs the decision tree using entropy and information gain of the attribute with user-based priority if the attributes. It mainly focuses to reduce the size of the tree and number of leaf nodes of the decision tree. This classifier follows different approaches for nominal and numerical attributes and builds the decision tree. It checks for a minimum number of objects for the nominal type of attributes. Numeric attributes are in the data set are sorted only once at the start. This algorithm utilises the sorted list to split the attributes at each tree node.

The algorithm

- Load input data
- Get attributes priority list in XML format from user
- Parse XML and create the attribute priority maps using SAX Parser
- Calculate entropy and information gain for the attributes other than class attribute
- Build decision tree based on information gain and attribute priority map
- Node with high information gain and priority is selected as decision node
- Repeat step 2 and 3 until all the data are classified

RESULTS AND DISCUSSION

For this research work, decision tree classifier algorithms are applied to Wisconsin original, diagnostic and prognostic breast cancer dataset. Each instance in the breast cancer dataset consists of the class attribute. The class attribute has four values like Benign (1) and Malignant (2). The classification algorithms are applied for the input parameters mentioned in Table.1. The classifiers with 10 fold cross validation are analysed and compared using WEKA software. The configuration parameters of the classifiers are listed below.

In WEKA, Data pre-processing has been carried out as the first step, and it has been depicted in fig. 2.

The performance of the classifiers in detecting the breast cancer can be evaluated from the analysis of confusion matrix and below parameters are calculated

Accuracy is the percentage measure of correctly classified instances for all instances. It can be obtained as below

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \ldots\ldots\ldots\ldots \quad (1)$$

Precision is of correctly classified instances for those instances that are classified as positive, and it is calculated using the equation

$$Precision = \frac{TP}{TP+FP} \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

Recall is the measure of the positive instance that are correctly classified, and it can be calculated with below equation

$$Recall = \frac{TP}{TP+FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

F-Measure is the combined metric of precision and recall, i.e., it is harmonic mean of both. It shows how precise the classifier is and also how well the classifier is robust. F-measure use below equation for calculation

$$F\text{-}Measure = \frac{2*Recall*Precision}{Precision+Recall} \ldots\ldots \quad (4)$$

Sensitivity is the measure of correctly classified positive instances to a total number of positive instances.

$$Sensitivity = \frac{TP}{TP+FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

Specificity is the measure of correctly classified negative instances to a total number of negative instances.

$$Specificity = \frac{TN}{TN+FP} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

Receiver operating curve (ROC) is graphical representation of sensitivity against specificity

The precision-recall curve is the graphical representation of recall against precision.

Kappa statistic is the measure of inter-rater agreement of the instances.

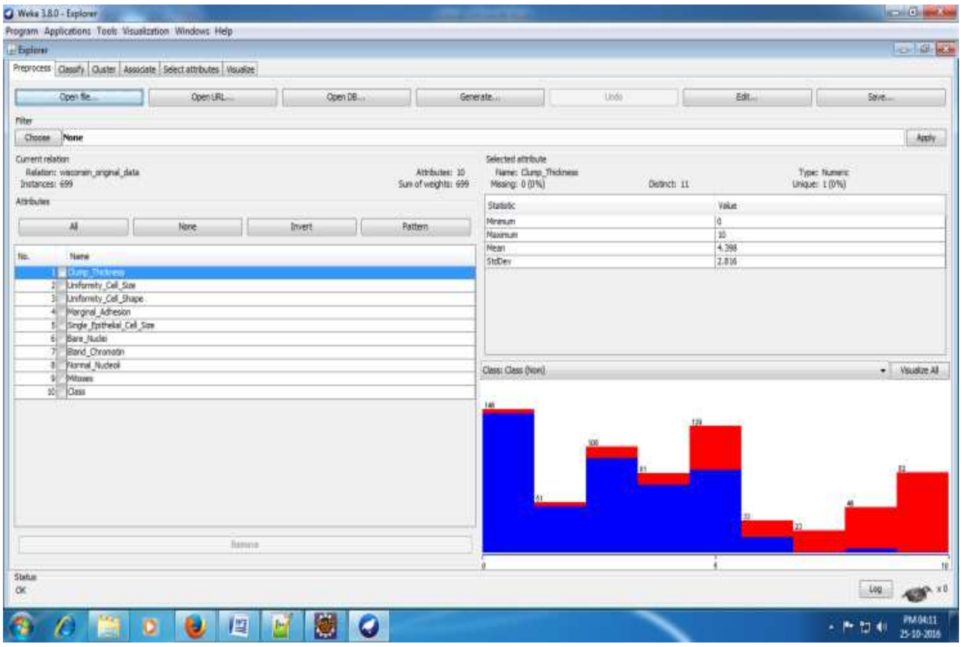
Entropy

It is a measure of uncertainty of a particular random variable. The entropy H(X) for a discrete random variable X is defined as follows

$$Entropy\ H(X) = \sum_{x=1}^{n} \ p_i \, log_b \, p_i \ldots \quad (7)$$

RMSE is the measure of the variations in predicating correct values.

Though we have more attributes as tabulated in table 1, 2 and 3, different classifier algorithms are imposed only on the pre-processed data.


Weka 3.8.0 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
All None Invert Pattern
Remove
Visualize All
Log
x 0
146
51
108
81
129
33
23
46
72
0
5
10
PM 04:11
25-10-2016

Filter
Choose None

Current relation
Relation: wisconsin_original_data
Instances: 699
Attributes: 10
Sum of weights: 699

Selected attribute
Name: Clump_Thickness
Missing: 0 (0%)
Distinct: 11
Type: Numeric
Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	10
Mean	4.398
StdDev	2.816

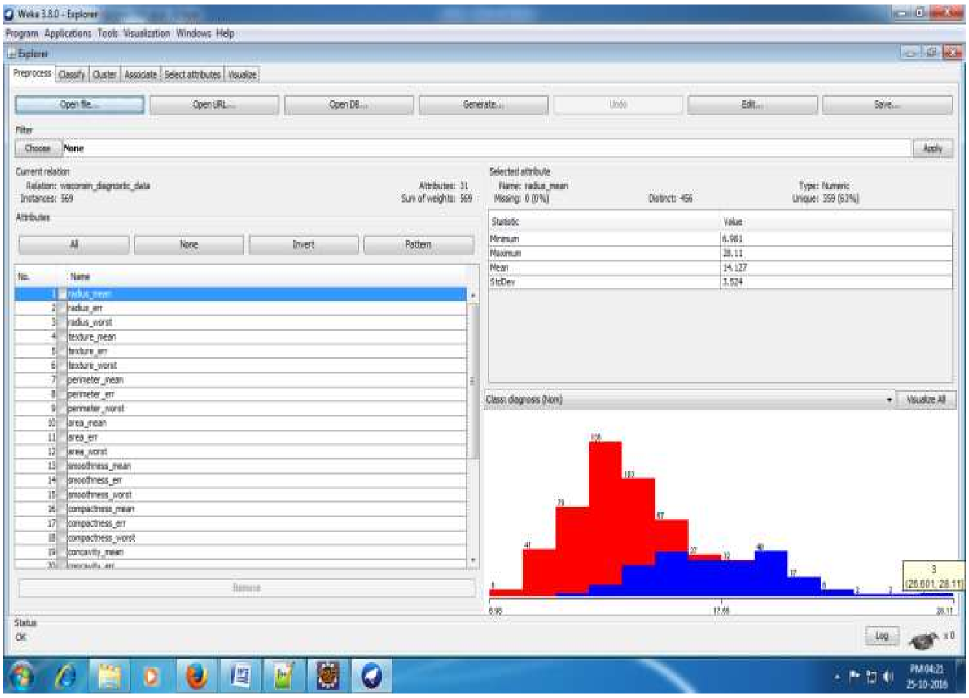
Attributes

No.	Name
1	Clump_Thickness
2	Uniformity_Cell_Size
3	Uniformity_Cell_Shape
4	Marginal_Adhesion
5	Single_Epithelial_Cell_Size
6	Bare_Nuclei
7	Bland_Chromatin
8	Normal_Nucleoli
9	Mitoses
10	Class

Class: Class (Nom)

Status
OK

Fig. 2: Data pre-processing of Wisconsin original breast cancer dataset


Weka 3.8.0 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
All None Invert Pattern
Remove
Visualize All
Log
x 0
8
41
79
158
103
87
32
32
40
17
6
2
2
3
(26.801, 28.11]
6.98
17.55
28.11
PM 04:21
25-10-2016

Filter
Choose None

Current relation
Relation: wisconsin_diagnostic_data
Instances: 569
Attributes: 31
Sum of weights: 569

Selected attribute
Name: radius_mean
Missing: 0 (0%)
Distinct: 456
Type: Numeric
Unique: 359 (63%)

Statistic	Value
Minimum	6.981
Maximum	28.11
Mean	14.127
StdDev	3.524

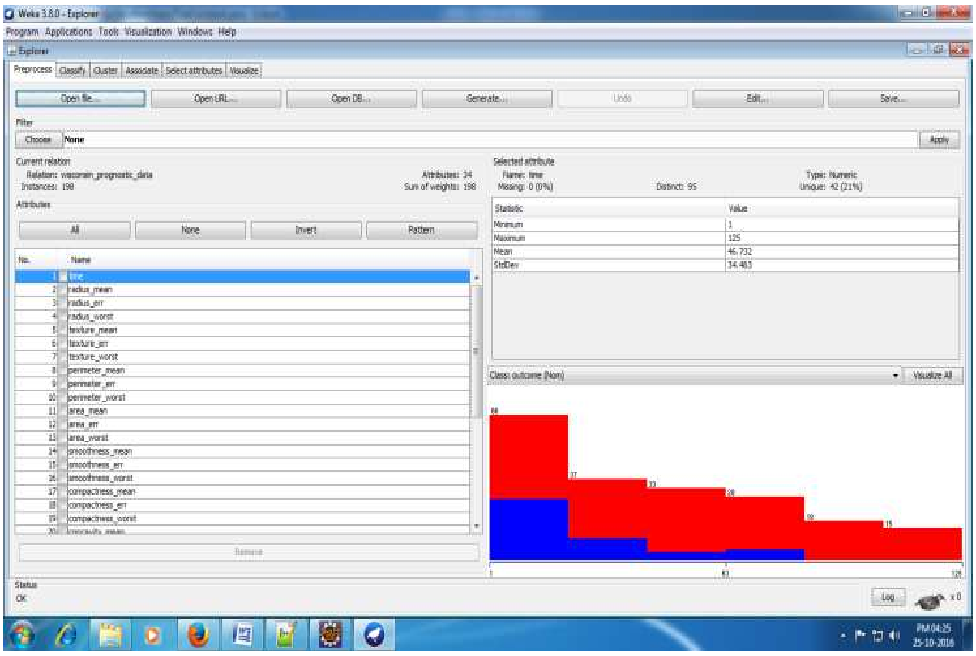
Attributes

No.	Name
1	radius_mean
2	radius_err
3	radius_worst
4	texture_mean
5	texture_err
6	texture_worst
7	perimeter_mean
8	perimeter_err
9	perimeter_worst
10	area_mean
11	area_err
12	area_worst
13	smoothness_mean
14	smoothness_err
15	smoothness_worst
16	compactness_mean
17	compactness_err
18	compactness_worst
19	concavity_mean
20	concavity_err

Class: diagnosis (Nom)

Status
OK

Fig. 3: Data pre-processing of wisconsin diagnostic breast cancer dataset


Weka 3.8.0 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose
Apply
All None Invert Pattern
Remove
Visualize All
68
37
33
28
18
15
1
63
125
Status
OK
Log
x 0
PM 04:25
25-10-2016

Filter: None

Current relation
Relation: wisconsin_prognostic_data
Instances: 198
Attributes: 34
Sum of weights: 198

Selected attribute
Name: time
Missing: 0 (0%)
Distinct: 95
Type: Numeric
Unique: 42 (21%)

Statistic	Value
Minimum	1
Maximum	125
Mean	46.732
StdDev	34.483

Class: outcome (Nom)

Attributes

No.	Name
1	time
2	radius_mean
3	radius_err
4	radius_worst
5	texture_mean
6	texture_err
7	texture_worst
8	perimeter_mean
9	perimeter_err
10	perimeter_worst
11	area_mean
12	area_err
13	area_worst
14	smoothness_mean
15	smoothness_err
16	smoothness_worst
17	compactness_mean
18	compactness_err
19	compactness_worst
20	concavity_mean

Fig. 4 Data pre-processing of wisconsin diagnostic breast cancer dataset

The simulation results of decision tree classifiers are plotted here. Confusion matrix helps us to evaluate a total number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) instance. With the help of TP, TN, FP and FN value, it is possible us to validate the various performance measures such as accuracy, precision, recall, F-measure, ROC, PRC, etc.

The performance of the classifiers is evaluated for Wisconsin original, diagnostic and prognostic breast cancer dataset. The evaluation parameters are tabulated in table 4, 5 and 6.

The important criterion of the classifier to classify the data is based on the ability of the classifier to classify the instances, sensitivity and specificity correctly. The error metrics includes Kappa, Root mean squared error, Mean absolute error, Relative mean squared error, Relative absolute error are also calculated for decision tree algorithms and tabulated in table 7, 8 and 9.

The performance parameters like True Positive Rate, False Positive Rate, Precision, Recall, F-Measure, ROC and PRC are also calculated for decision tree algorithms against Wisconsin original, diagnostic and prognostic dataset and results are tabulated in table 10, 11 and 12.

Table 4: Performance parameters of decision tree algorithms for Wisconsin original dataset

	J48	Random Forest	Random Tree	REP Tree	Priority based
Correctly Classified Instances	654	676	658	658	662
Incorrectly Classified Instances	45	23	41	41	37
Accuracy (%)	93.56	96.70	94.13	94.13	94.70
Sensitivity	0.956	0.984	0.952	0.962	0.988
Specificity	0.897	0.935	0.9194	0.902	0.879
Entropy (bits/instance)	0.769	0.809	0.800	0.764	0.746

Table 5: Performance parameters of decision tree algorithms for Wisconsin diagnostic dataset

	J48	Random forest	Random tree	REP Tree	Priority based
Correctly Classified Instances	543	550	536	537	549
Incorrectly Classified Instances	26	19	33	32	20
Accuracy (%)	95.43	96.66	94.20	94.37	96.48
Sensitivity	0.951	0.961	0.916	0.924	0.936
Specificity	0.955	0.969	0.957	0.955	0.982
Entropy (bits/instance)	0.844	0.829	0.831	0.811	0.832

Table 6: Performance parameters of decision tree algorithms for Wisconsin prognostic dataset

	J48	Random forest	Random tree	REP tree	Priority based
Correctly Classified Instances	148	165	136	152	166
Incorrectly Classified Instances	50	33	52	46	32
Accuracy (%)	74.74	83.33	68.68	76.7	83.83
Sensitivity	0.951	0.961	0.916	0.924	0.936
Specificity	0.955	0.969	0.957	0.955	0.982
Entropy (bits/instance)	0.111	0.061	0.022	0.066	0.197

Table 7: Error statistics of decision tree algorithms for Wisconsin original dataset

	J48	Random forest	Random tree	REP tree	Priority based
Kappa statistic	0.857	0.927	0.869	0.870	0.885
Mean absolute error	0.077	0.061	0.058	0.082	0.096
Root mean squared error	0.239	0.1673	0.2422	0.2311	0.2198
Relative absolute error (%)	17.20	13.658	13.029	18.338	21.378
Root relative squared error (%)	50.45	35.269	51.056	48.725	46.331

Table 8: Error statistics of decision tree algorithms for Wisconsin diagnostic dataset

	J48	Random forest	Random tree	REP tree	Priority based
Kappa statistic	0.9017	0.9284	0.8763	0.8797	0.9254
Mean absolute error	0.0537	0.0672	0.058	0.0729	0.0659
Root mean squared error	0.208	0.1576	0.2408	0.2175	0.1823
Relative absolute error (%)	11.489	14.359	12.401	15.586	14.090
Root relative squared error (%)	43.027	32.597	49.808	44.991	37.702

Table 9: Error statistics of decision tree algorithms for Wisconsin prognostic dataset

	J48	Random forest	Random tree	REP tree	Priority based
Kappa statistic	0.2704	0.4044	0.2156	0.2481	0.4486
Mean absolute error	0.2907	0.3235	0.3131	0.3258	0.2694
Root mean squared error	0.4766	0.3968	0.5596	0.4294	0.3768
Relative absolute error (%)	79.945	88.953	86.106	89.59	74.092
Root relative squared error (%)	111.99	93.2307	131.484	100.89	88.539

Table 10: Weighted average performance parameters of decision tree algorithms for Wisconsin original dataset

	J48	Random Forest	Random Tree	REP Tree	Priority based
TP Rate	0.936	0.967	0.941	0.941	0.947
FP Rate	0.074	0.031	0.075	0.065	0.038
Precision	0.936	0.968	0.941	0.942	0.951
Recall	0.936	0.967	0.941	0.941	0.947
F-Measure	0.936	0.967	0.941	0.942	0.948
ROC	0.941	0.992	0.933	0.948	0.945
PRC	0.918	0.991	0.915	0.930	0.924

Table 11: Weighted average performance parameters of decision tree algorithms for Wisconsin diagnostic dataset

	J48	Random forest	Random tree	REP tree	Priority based
TP Rate	0.954	0.967	0.942	0.944	0.965
FP Rate	0.058	0.041	0.063	0.064	0.032
Precision	0.954	0.967	0.942	0.944	0.966
Recall	0.954	0.967	0.942	0.944	0.965
F-Measure	0.954	0.967	0.942	0.944	0.965
ROC	0.952	0.996	0.939	0.955	0.958
PRC	0.938	0.996	0.917	0.942	0.941

Table 12: Weighted average performance parameters of decision tree algorithms for Wisconsin prognostic dataset

	J48	Random forest	Random tree	REP tree	Priority based
TP Rate	0.747	0.833	0.687	0.768	0.838
FP Rate	0.489	0.521	0.449	0.556	0.475
Precision	0.737	0.851	0.719	0.739	0.841
Recall	0.747	0.833	0.687	0.768	0.838
F-Measure	0.741	0.800	0.700	0.743	0.814
ROC	0.617	0.672	0.619	0.635	0.689
PRC	0.690	0.765	0.690	0.718	0.759

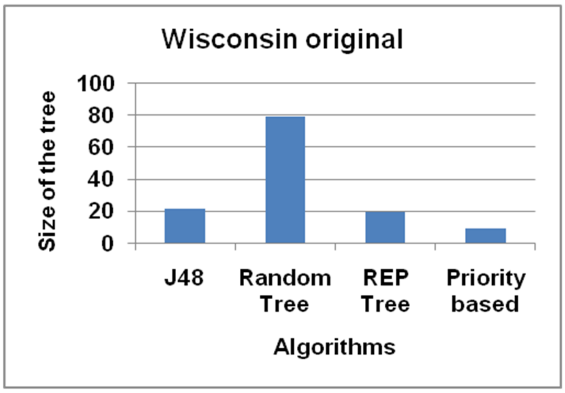


Fig. 5: Tree size of decision tree algorithms for Wisconsin original dataset

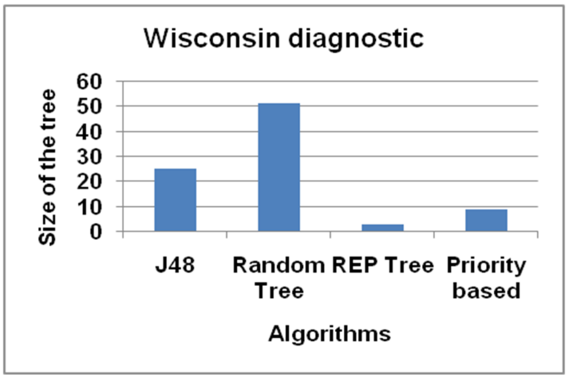


Fig. 6: Tree size of decision tree algorithms for Wisconsin diagnostic dataset

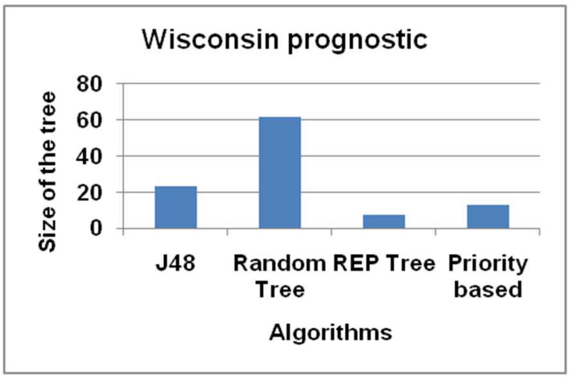


Fig. 7: Tree size of decision tree algorithms for Wisconsin prognostic dataset

The decision tree has the great impact on the computational complexity of the algorithm. When compared to other decision tree algorithm, priority-based decision tree algorithm has minimum tree size and thus it reduces the complexity of the algorithm and time consumption.

CONCLUSION

In this research work, we analysed the performance of the four different decision tree algorithms for Breast cancer classification. The simulation results show Priority based decision tree classifier classifies the data with 93.63% accuracy and minimum RMSE of 0.1628. It also consumes less time to build the model with 0.929 ROC and 0.959 PRC values. By comparing classification results, we confirm that a Priority based decision tree algorithm is better than other classification algorithms for Wisconsin original, diagnostic and prognostic breast cancer dataset.

ACKNOWLEDGMENT

I would like to express my gratitude to UCI repository for providing Wisconsin Breast cancer dataset and also thank authors, mentioned in the references which are citied below for their valuable research works which helped me to gain knowledge. I thank my mentors for their precious guidance.

CONFLICT OF INTERESTS

Declared none

REFERENCES

1. RW Brause. Medical analysis and diagnosis by neural networks. Lecture Computer Sci 2001;2199:1-13.
2. Vaidehi K, Subashini TS. Breast tissue characterization using combined K-NN classifier. Indian J Sci Technol 2015;8:23–6.
3. Williams K, Idowu PA, Balogun JA, Oluwaranti A. Breast cancer risk prediction using data mining classification techniques. Transactions Networks Communications 2015;3:1–11.
4. Xindog Wu, Vipin Kumar. Top 10 algorithms in data mining. Knowledge Information Systems 2008;14:1-37.
5. Aruna S, Rajagopalan SP, Nandakishore LV. Knowledge-based analysis of various statistical tools in detecting breast cancer. Computer Sci Inf Technol 2011;2:37–45.
6. TM Cover. Geometrical and statistical properties of systems of linear with applications in pattern recognition. IEEE Transactions Electronic Computers EC-14; 1965. p. 326-34.
7. Ramnath Takiar. Projections of a number of cancer cases in India (2010-2020) by Cancer Groups. Asian Pac J Cancer Prev 2010;11:1045-9.
8. Evanthia E Tripoliti. Automated diagnosis of diseases based on classification: dynamic determination of the number of trees in random forests algorithm. IEEE Transactions Information Technol Biomed 2012;16:615-22.

How to cite this article

- P Hamsagayathri, P Sampath. Performance analysis of breast cancer classification using decision tree classifiers. Int J Curr Pharm Res 2017;9(2):19-25.