International Journal of Current Pharmaceutical Research

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

Original Article

DECISION TREE CLASSIFIERS FOR CLASSIFICATION OF BREAST CANCER

P. HAMSAGAYATHRI, P. SAMPATH

Department of ECE, Bannari Amman Institute of Technology, Sathyamangalam, Erode
Email: hamsagayathri.co14@bitsathy.ac.in

Received: 25 Oct 2016, Revised and Accepted: 29 Dec 2016

ABSTRACT

Objective: 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. Survey: 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 united states. Also, 246,660 new cases of women with cancer are estimated for the year 2016.

Methods: Early diagnosis of breast cancer is a key factor for long-term survival of cancer patients. Classification is one of the vital techniques used by researchers to analyze and classify the medical data.

Results: This paper analyzes the different decision tree classifier algorithms for seer breast cancer 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.

Conclusion: The simulation results shows REPTree classifier classifies the data with 93.63% accuracy and minimum RMSE of 0.1628 REPTree algorithm consumes less time to build the model with 0.929 ROC and 0.959 PRC values. By comparing classification results, we confirm that a REPTree algorithm is better than other classification algorithms for SEER dataset.

Keywords: Classification, J48, REP Tree, Random Forest, Random Tree, Accuracy, RMSE, Confusion matrix

DOI: http://dx.doi.org/10.22159/ijcpr.2017v9i1.17377

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 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% [5].

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.

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 SEER dataset.

Research scope

The scope of the research is to apply the classifier algorithms such as J48, REP Tree, Random Forest and Random Tree on SEER Breast cancer dataset that contains 762691 instances with 134 attributes. Data cleaning and reduction are performed, and 7 Key attributes are finalised 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 categorized 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 about the simulation results that are obtained for various decision tree algorithms.

MATERIALS AND METHODS

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.

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. There are 762691 instances with 134 attributes in Breast cancer dataset, but only 7 attributes are considered for classification using feature selection technique.

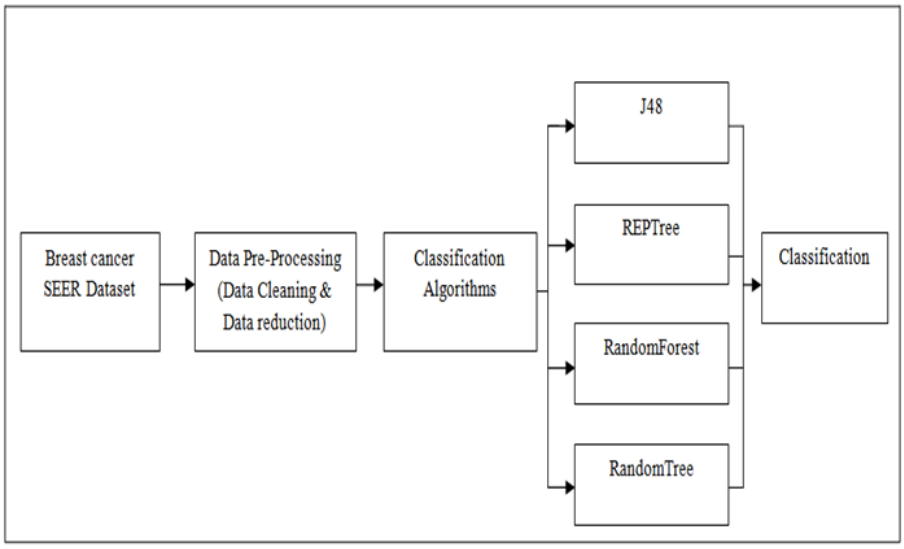

Fig. 1: Methodology for data classification

Table 1: Dataset attributes for classification

Attribute	Length	Description
Age at diagnosis	3	Age of the patient at diagnosis
Grade	1	Specify T-cell, B-Cell involvement in lymphoma and leukaemia
CS Tumor Size	3	Information on Tumour Size
CS Extension	3	Information on extension of Tumour
CS Lymph nodes	3	Information on the involvement of lymph nodes.
CS Mets at DX	2	Information on distant metastasis.
Behavior code ICD-O-3	1	Describes on the nature of tumour as begin, in situ or malignant

Classification algorithms

There have been various algorithms used for classification of Breast cancer. This paper provides the detailed description on decision tree algorithms (J48 and REPTree) and evaluates based on the performance measures like accuracy, sensitivity, specificity, entropy, ROC, PR area 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. Decision tree is 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 minimizing 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 maintains the accuracy in estimation. It can handle 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

RESULTS AND DISCUSSION

For this research work, decision tree classifier algorithms are applied to the breast cancer dataset from Surveillance, Epidemiology, and End Results (SEER) repository. The breast cancer dataset has 769261 numbers of instances and each instance consists of 134 attributes including the class attribute. The class attribute has four values like Benign (0), uncertain benign or malignant (1), Carcinoma in situ (2) and Malignant (3). All the attributes of the data *set al.* ong with their range of values are available in seer data dictionary [6]. The classification algorithms are applied for the input parameters mentioned in the table 1. The classifiers with 10 fold cross validation are analyzed 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 for all the 7 attributes 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:

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

The different classifier algorithms are imposed on the pre-processed data.

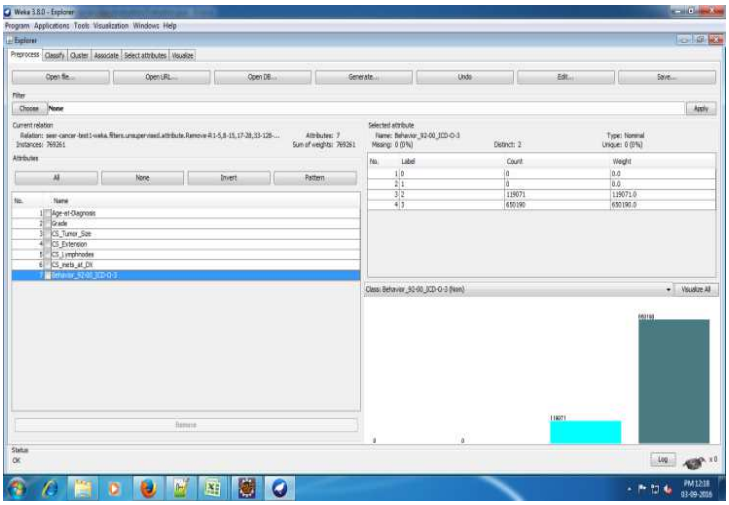

Fig. 2: Data pre-processing of selected attributes

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.

Results of J48

Table 2: Performance parameters of J48

Parameters	Class (0)	Class(1)	Class(2)	Class(3)
TP Rate	0	0	0.68	0.983
FP Rate	0	0	0.017	0.32
Precision	0	0	0.881	0.944
Recall	0	0	0.68	0.983
F-Measure	0	0	0.767	0.963
ROC	0	0	0.891	0.891
PRC	0	0	0.782	0.972

Results on REPTree

Table 3: Performance parameters of REPTree

Parameters	Class (0)	Class(1)	Class(2)	Class(3)
TP Rate	0	0	0.678	0.983
FP Rate	0	0	0.017	0.322
Precision	0	0	0.883	0.943
Recall	0	0	0.678	0.983
F-Measure	0	0	0.767	0.963
ROC	0	0	0.929	0.929
PRC	0	0	0.826	0.984

Results on random forest

Table 4: Performance parameters of random forest

Parameters	Class (0)	Class(1)	Class(2)	Class(3)
TP Rate	0	0	0.677	0.983
FP Rate	0	0	0.017	0.323
Precision	0	0	0.880	0.943
Recall	0	0	0.677	0.983
F-Measure	0	0	0.766	0.963
ROC	0	0	0.932	0.932
PRC	0	0	0.832	0.985

Results on RandomTree

Table 5: Performance parameters of RandomTree

Parameters	Class (0)	Class(1)	Class(2)	Class(3)
TP Rate	0	0	0.679	0.983
FP Rate	0	0	0.017	0.321
Precision	0	0	0.878	0.943
Recall	0	0	0.679	0.983
F-Measure	0	0	0.765	0.963
ROC	0	0	0.931	0.931
PRC	0	0	0.827	0.984

The simulation results of the four decision tree classifiers for all possible values of a class attribute is summarised in the table 2, 3, 4 and 5. When compared to other classifiers, REPTree Classifier has high TP Rate of 0.936 and low FP Rate of 0.275. REPTree holds good Receiver Operating Curve (ROC) of 0.929 with 0.959 Precision-Recall areas and it is shown in fig. 3. The comparison of information score of decision tree classifier is clearly shown in table 7 and REPTree has 280488.65 bits per instance.

Error results

Table 6: Error comparison of IBK and Naïve Bayes

Error	J48	REPTree	Random forest	Random tree
Kappa statistic	0.7311	0.7308	0.7293	0.7288
Mean absolute error	0.0546	0.0529	0.0524	0.0524
Root mean squared error	0.1653	0.1628	0.1626	0.1631
Relative absolute error	41.73%	40.46%	40.03%	40.02%
Root relative squared error	64.62%	63.65%	63.59%	63.78%

Random Forest Classifier classifies the data with minimized Mean Squared Error value of 0.1626 and various error statistics are captured in table 6.

Entropy of the algorithms

Table 7: Entropy comparison of IBK and naïve bayes

Entropy	J48	REPTree	Random forest	Random tree
Information Score in bits	280133.9	280488.65	276837.28	276852.8

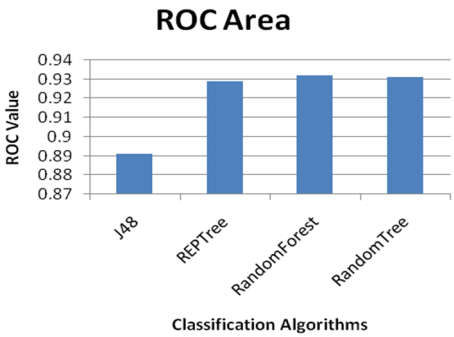

Fig. 3: Comparisons of ROC values

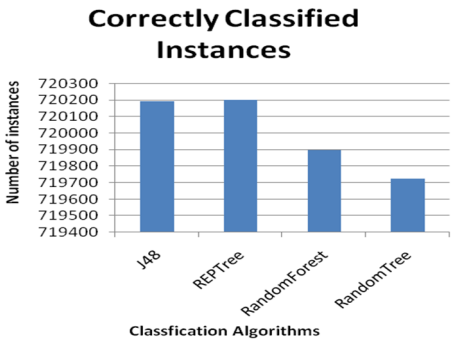

Fig. 4: Comparison of correctly classified instances

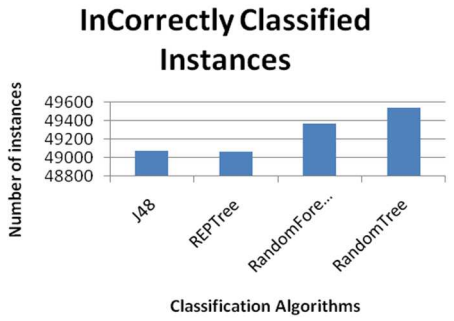

Fig. 5: Comparison of incorrectly classified instances

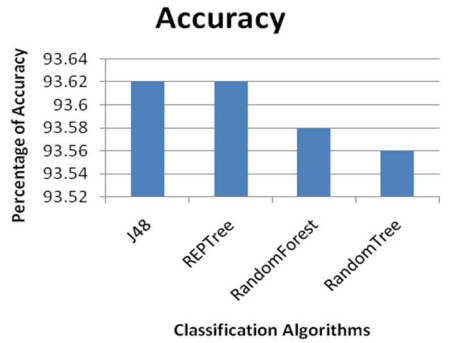

Fig. 6: Comparison of performance accuracy

REPTree Classifier with performance accuracy of 93.63%, 720198 instances are correctly classified and only 49063 instances are classified as incorrect. The charts are captured in fig. 4, 5 and 6 respectively.

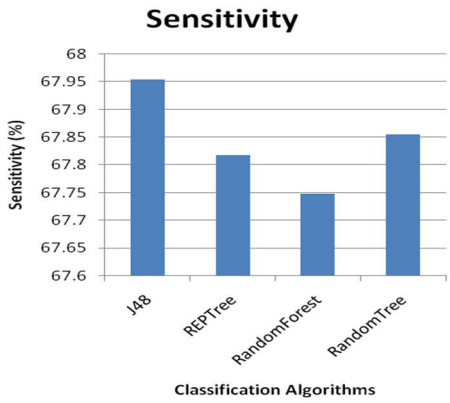

Fig. 7: Comparison of sensitivity

REPTree has a specificity of 98.35% and sensitivity of 67.83% and thus it has better capability to classify the data and graphs are depicted in the fig. 7 and fig. 8.

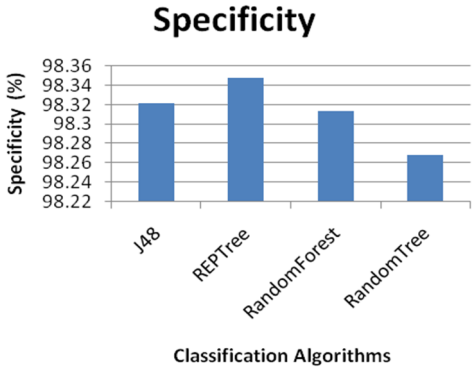

Fig. 8: Comparison of specificity

CONCLUSION

In this research work, we analyzed the performance of the four different decision tree algorithms for Breast cancer classification. The simulation results shows REPTree classifier classifies the data with 93.63% accuracy and minimum RMSE of 0.1628 REPTree algorithm consumes less time to build the model with 0.929 ROC and 0.959 PRC values. By comparing classification results, we confirm that an REPTree algorithm is better than other classification algorithms for SEER dataset.

ACKNOWLEDGMENT

I would like to express my gratitude to SEER Database for providing 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. Aruna S, Rajagopalan SP, Nandakishore LV. Knowledge-based analysis of various statistical tools in detecting breast cancer. Computer Science and Information Technology. 2011;2:37–45.
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. RW Brause. Medical analysis and diagnosis by neural networks. Lecture Notes Comput Sci 2001;2199:1-13.
6. http://seer.cancer.gov/popdata/popdic.html-SEER dictionary. [Last accessed on 20 Sep 2016]
7. TM Cover. Geometrical and statistical properties of systems of linear with applications in pattern recognition. IEEE Transactions on Electronic Computers EC-14; 1965. p. 326-34.
8. Ramnath Takiar. Projections of a number of cancer cases in India (2010-2020) by Cancer Groups. Asian Pacific J Cancer Prevention 2010;11:1045-9.
9. Evanthia E Tripoliti. Automated diagnosis of diseases based on classification: dynamic determination of the number of trees in random forests algorithm. IEEE Transactions On Information Technology In Biomedicine; 2012. p. 16.

How to cite this article

- P Hamsagayathri, P Sampath. Decision tree classifiers for classification of breast cancer. Int J Curr Pharm Res 2017;9(2):31-36.