Analysis of Decision Tree Classification algorithms for Breast Cancer detection

Prof.Bhanudas Suresh Panchbhai, R.C.Patel Arts, Commerce and Science College, Shirpur

Bharat.panchbhai@gmail.com

Abstract— *analyzing massive amounts of data is becoming a requirement. People do not have time to examine extraordinarily huge data sets such as medical, marketing, or financial data. As a result, we'll need a method for automatically analyzing data. The process of extracting valuable information from vast amounts of data and analyzing, classifying, and summarizing it into useful information from big data is known as data mining. For the diagnosis of breast cancer, a data mining classification technique is applied. In this study, we compare the performance of several classifiers on the basis of accuracy, recall, precision, F-measure, computing time, correctly classified instances, and kappa statistics, MAE, RMSE, RAE, RRSE on a breast cancer dataset. To easily assess the classifiers, we include confusion matrices from various classifiers. We looked at a variety of data mining classification methods in order to find the best ones for efficiently classifying the Breast Cancer dataset.*

Keywords- Data Mining, WEKA tool, Breast Cancer Patients dataset, Decision Tree Classification algorithm

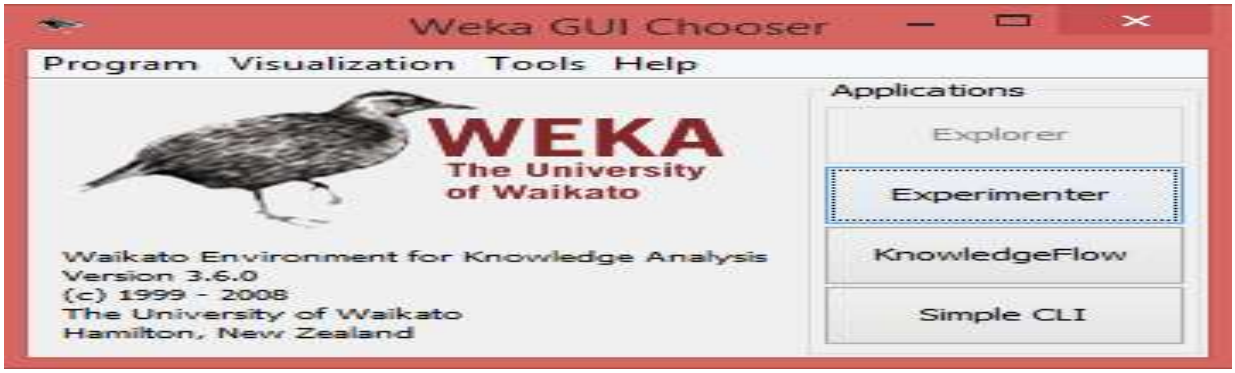
1. INTRODUCTION

Data mining is the process of converting a significant volume of data into knowledge. Exploratory data analysis, data-driven discovery, and deductive learning are other terms for it. The most often used data mining approach is classification. To aid in more accurate prediction and analysis, classification assigns categories to a collection of data. In India, breast cancer affects a huge number of women, and it is a difficult condition to diagnose. The main purpose of this study is to classify a breast cancer dataset using several decision tree classifiers in order to detect whether or not a person has a recurrence. We examine multiple classifiers to find the best one for correctly classifying the Breast cancer dataset and diagnosing the condition at a lower cost.

We employ patient characteristics such as age, menopause, tumor-size, inv-nodes, node-caps,
Deg-malig, breast, breast-quad, and irradiate to identify the disease. We classify this dataset using several decision tree classifiers in order to determine which is the most effective classifier for accurately classifying the most number of instances in the shortest amount of time.

2. **WEKA** (Waikato Environment for Knowledge Analysis)

The Waikato Environment for Knowledge Analysis (WEKA) is a prominent Java-based machine learning software suite developed at New Zealand's University of Waikato. It is GNU General Public License (GPL) licensed free software.
WEKA is a data analysis and predictive modeling workbench [1] that includes a variety of visualization tools and algorithms, as well as graphical user interfaces enabling quick access to these operations. It is mostly used to import datasets, execute algorithms, and plan and run experiments with statistically sound results that may be published.
The WEKA tool includes decision tree-based classification methods such as the J48 decision tree, rule-based classification methods such as Zero R and decision tables, and probability and regression-based classification methods such as the Nave Bye's algorithm. WEKA requires a dataset file in the ARFF format (Attribute Relation File Format), with the extension dot ARFF (.arff). WEKA can be found at www.cs.waikato.ac.nz/ml/weka on the web.



3. CLASSIFICATION

The practice of classifying data into categories for the most effective and efficient use is known as data categorization. Decision trees, logistic regression, neural networks, and other classification methods are used in data mining. For classification, we use the decision tree approach in this study. The following steps are involved in the classification process:

1. Create a data set for training.
2. Determine the attributes and classes of each class.
3. Identify classification attributes that are useful (Relevance analysis).
4. Use the training examples in the Training set to learn a model.
5. Apply the model to the unknown data samples to classify them.

3.1 DECISION TREE CLASSIFICATION METHODS

A decision tree is a network with a root node, branches, and leaf nodes. Each internal node represents an attribute test, each branch represents a test result, and each leaf node represents a class label. The root node is the highest node in the tree.

The decision tree below is for the notion buy computer, and it indicates whether or not a company's consumer is likely to buy a computer. A test on an attribute is represented by each internal node. A class is represented by each leaf node.

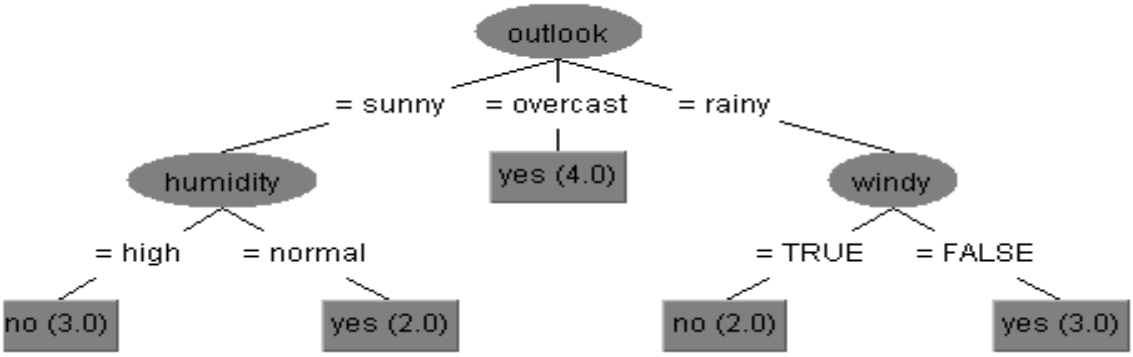


Figure I: Decision Tree

The following are some of the advantages of using a decision tree: 1. It does not require subject knowledge.

2. It is simple to understand.
3. A decision tree's learning and categorization phases are simple and quick.

Classifiers based on decision trees

3.1.1. J48: Algorithm WEKA the enhanced version of C4.5 is J48. For decision-making, the algorithm employs a greedy strategy. Different nodes, such as the root node, intermediate nodes, and leaf node, make up the structure of

the output decision tree. Each internal node in the tree represents a separate property, whereas the terminal nodes represent the dependent variable's final value.

3.1.2. Simple CART:

CART, or Classification and Regression Tree Classification Technique, was developed by Leo Brejman, Jenome Friedman, Richard Olshen, and Charles Stone in 1984. The classification tree and regression tree are used in this process.

• Classification Tree- In this case, the huge variable is categorical, and the tree is used to determine which "Class" a target variable would most likely fall into.

• Regression Tree- A regression tree is used to predict the value of a continuous variable.

3.1.3. ADTree- An alternating decision tree (ADTree) is a classification machine learning method that generalizes decision trees and is related to boosting.

An AD Tree is made up of a series of decision nodes that indicate a condition and prediction nodes that carry a single integer. An AD Tree classifies an instance by tracing all pathways that lead to it.

3.1.4 .BFTree-

The best node is the node whose split leads to the highest decrease in impurity (e.g. Gini index or information gain) among all nodes available for splitting in the Best First Tree Algorithm [7]. When fully developed, the final tree will be identical, but the sequence in which it is constructed will differ. The tree-growing approach aims to maximize within-node homogeneity. Impurity is defined as the extent to which a node does not represent a homogeneous subset of cases. A homogeneous node is one in which all cases have the same value for the dependent variable. It does not need to be split further because it is pure.

4. **DATASET**

A dataset is a grouping of data. A data set is often the contents of a single database table or statistical data matrix, where each column of the table represents a specific variable and each row represents a specific member of the data set in question.

In this paper, we use the Oncology Institute University Medical Centre's Breast Cancer Database, which is available on WEKA.

There are 286 occurrences total in the dataset, with 201 instances of one type and 85 instances of another. There are nine attributes that describe the instances, some of which are linear and some of which are nominal.

Information on the attributes:

1. Your age (10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99).
2. The Menopause (lt40, ge40, premeno).
3.Tumor–size (0-4, 5-9, 10-14, 15-19, 20-24,25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59)
4. Inv-nodes (0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39)
5. Node-caps (yes, no)
6. Deg-malig (1, 2, 3)
7. Breast (left right)
8. Breast-quad (left-up, left-low, right-up, right-low, central)
9. Irradiate (yes, no)
10. Class (no-recurrence-events, recurrence-events)

5. RESULTS AND DISCUSSION

5.1 EVALUTION MATRICS

The classification is based on the performance indicators listed below. [1]

1. Time: This is the amount of time it takes to finish training or modeling a dataset. It's measured in seconds.
2. The Kappa Statistic is a statistic that measures how much of a difference there is

Nonrandom agreement between observers or measurements of the same category variable is measured using this metric.

3. Mean Absolute Error (M.A.E.):

The average difference between expected and actual values in all test cases is the mean absolute error; it is the average prediction error.

4. The Mean Squared Error (MSE) is a measure of how accurate a calculation is.
The mean-squared error is one of the most widely used indicators of numeric prediction success. The average of the squared discrepancies between each computed value and its matching true value is used to calculate this value. The square root of the mean-squared-error is the mean-squared-error. The mean-squared error has the same dimensionality as the actual and projected values thanks to the mean-squared error.
5. Root relative squared error: The total squared error created in comparison to the error that would have occurred if the prediction had been the average of the absolute value. The square root of the relative squared error, like the root mean squared error, is used to give it the same dimensions as the anticipated value.
6. Relative Absolute Error: The total absolute error made in comparison to the error that would have occurred if the prediction had merely been the average of the actual values.

The result in Table I is achieved using these measurements.

A confusion matrix is a valuable tool for determining the accuracy of a classifier. The confusion matrix's structure is shown below.

	a	b
a	True Negative	False Positive
b	False Negative	True Positive

Figure II: confusion matrix

The result in Table II is achieved using these measurements.

Positive tuples that were successfully categorized by the classifier are referred to as True Positive (TP). Negative tuples that were correctly categorized by the classifier are referred to as True Negative (TN). Negative tuples that were mistakenly categorized by the classifier are referred to as False Positive (FP). Positive tuples that were mistakenly categorized by the classifier are referred to as False Negative (FN).

The fraction of tuples properly classified by the classifier is called accuracy.

Accuracy= (TP+TN)/ (TP+TN+FP+FN)

Recall: The proportion of examples identified as class x among all examples that genuinely have class x, i.e. how much of the class was captured, is known as recall.

Recall=TP/ (TP+FN)

Precision:

Precision is the proportion of the examples which truly have class x among all those which were classified as class x.

Precision=TP/(TP+FP)

F-Measure:

The harmonic mean of precision and recall. It is an important measure as it gives equal importance to precision and recall.

F-measure=2*recall*precision/precision + recall

Receiver Operating Characteristic (ROC) Curve:

It is a graphical approach for displaying the tradeoff between true positive rate (TPR) and false positive rate (FPR) of a classifier. TPR is plotted along the y axis and FPR is plotted along the x axis. Performance of each classifier represented as a point on the ROC curve.

Using this metrics the result in **Table III** is obtained.

5.2 RESULT:

The datasets were analyzed using the cross validation approach. For each of the datasets listed in Tables I, II, and III, various performance indicators were calculated. The following is a comparison of various decision tree categorization results:

	J48	**Simple CART**	**AD Tree**	**BF Tree**
Time (Seconds)	0.02	0.18	0.01	0.15
Correctly Classified Instances	216	198	211	194
KAPPA Statistic	0.2826	0.0671	0.329	0.0875
MAE	0.3676	0.393	0.3919	0.3887
RMSE	0.4324	0.4587	0.4333	0.4698
RAE %	87.86	93.93	93.66	92.90
RRSE%	94.60	100.36	94.80	102.78

Table: I ERRORS MEASUREMENT FOR DIFFERENT DECISION TREE CLASSIFIERS IN WEKA

Decision Tree	**True Negative**	**True Positive**	**Correctly Classified Instances**
J48	23	193	216
Simple CART	10	188	198
AD Tree	38	173	211
BF Tree	16	178	194

Table: II CONFUSION METRICS FOR DIFFERENT DECISION TREE CLASSIFIERS IN WEKA

Decision Tree	TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	ROC CURVE AREA
J48	**0.755**	**0.524**	0.752	0.755	0.713	0.584
Simple CART	**0.692**	**0.639**	0.632	0.692	0.625	0.593
AD Tree	**0.738**	**0.43**	0.724	0.738	0.727	0.712
BF Tree	**0.678**	**0.605**	0.628	0.678	0.635	0.6

Table: III PERFORMANCE METRICS (Weighted Avg.) FOR DIFFERENT DECISION TREE CLASSIFIERS IN WEKA

6. CONCLUSIONS

We looked at four alternative decision tree classification algorithms in this research. Using a Breast cancer dataset, we investigate J48, Simple CART, AD Tree, and BF Tree decision tree classification methods. We found that J48 successfully identified the most instances216 and took 0.02 seconds, whereas AD Tree took 0.01 seconds and correctly identified 211 occurrences. Simple CART takes 0.18 seconds to identify 198 occurrences properly, while BF Tree takes 0.15 seconds to identify 194 instances correctly. Some of these four classifiers are more accurate, while others take less time. According to their purposes, the most appropriate classifier can be employed.

7. REFERENCES

[1]P.Yasodha, N.R. Ananthanarayanan "**Comparative Study of Diabetic Patient Data's Using Classification Algorithm in WEKA Tool"** International Journal of Computer Applications Technology and Research Volume 3–Issue 9, 554 - 558, 2014, ISSN: 2319–8656

[2] Purva**Sewaiwar, Kamal Kant Verma "**Comparative Study of Various Decision Tree Classification Algorithm Using WEKA"*International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10)*

[3] K. Rajesh, V. Sangeetha "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" **International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012**

[4] http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/breast-cancer.data

[5] Jiawei Han "Data mining Concepts and Techniques" Third Edition

[6] www.cs.waikato

[7]H.Shi, "Best-first decision tree learning", Citeseer 2007.