Prediction for Breast Cancer Patients using Classification

Thin Thin Yi
University of Computer Studies, Pakokku
thinthinyee.pku@gmail.com

Ei Shwe Zin Naing
University of Computer Studies, Pakokku
einge.yso@gmail.com

Toe Toe Lwin
University of Computer Studies, Thaton
toetoe907@gmail.com

Abstract

Nowadays breast cancer becomes very major disease in many women. Breast cancer is a disease in which cells in the breast grow out of control. Signs of breast cancer include swelling by all or part from a breast, dimple skin, breast or nipple pain and nipple discharge. By using the breast cancer patient data, it generates the rules based on Classification and Adaptive Regression Trees (CART). CART is binary tree and it splits two branches for each node. Gini splitting rule is used for tree growing. Splitting is stopped when number of observations in the node is less than predefined minimum vale (Nmin). In this paper the Gini classification algorithm is used and is helpful in early treatment for the breast cancer patients.

Keyword*:CART(Classification and Adaptive Regression Trees)*

1. Introduction

Breast cancer is a major disease which is found in many women. Breast cancer is a disease in which cancer cells form the tissues of the breast of the woman. The breast is made up of lobes (15 to 20 sections) and ducts. Ducts are thin cube of linking to breast for producing milk. The nipple and areola is outside of the breast and its color is dark than the breast. The most common type of breast cancer begins in the cells of the ducts. Breast cancer is caused by gene changes. Chest x-ray, CT scans, BONE scan and PET scans are used to detect the stages of breast cancer.

Data mining concepts and methods can be applied in various fields like marketing, medicine, real estate, customer relationship management, engineering, web mining etc. Data mining uses many techniques such as decision trees, neural networks, naïve bayes, k-nearest neighbor, and so on. Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering.

The classification and regression tree (CART), an important data mining technique, is a non-parametric model without any pre-defined underlying relationship between the dependent and independent variables, which has been widely employed in many fields of study [1].

2. Building the Classification Model

This section describes the building of the classification model. In general, data classification is a two-step process. In the first step, which is called the learning step, a model that describes a predetermined set of classes or concepts is built by analyzing a set of training database instances. Each instance is assumed to belong to a predefined class. In the second step, the model is tested using a different data set that is used to estimate the classification accuracy of the model. If the accuracy of the model is considered acceptable, the model can be used to classify future data instances for which the class label are not known. At the end, the model acts as a classifier in the decision making process. There are several techniques that can be used for classification such as decision tree, Bayesian methods, rule based algorithms, and neural network.

Decision tree classifiers are quite popular techniques because the construction of tree does not require any domain expert knowledge or parameter sett, and is appropriate for exploratory knowledge discovery. Decision tree can produce a model with rules that are human-readable and interpretable. Decision tree has the advantages of easy interpretation and understanding for decision makers to compare with their domain knowledge for validation and justify their decision. Some of the decision tree classifiers are C4.5, ID3, CART, NBTree, and others.

3. CART

CART stands for classification and regression trees introduced by Breiman. It is also based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree [2]. For building decision trees, CART uses so-called learning sample – a set of historical data with pre-assigned classes for all observations. For example, learning sample for credit scoring system would be fundamental information about previous borrows (variable) matched with actual payoff results (classes) [3, 4].

The CART methodology is technically known as binary recursive partitioning [3, 4]. The process is binary because parent nodes are always split into exactly two child nodes and recursive because the process can be repeated by treating each child node as a parent. The key elements of a CART analysis are a set of rules for:

(1) Constriction of maximum tree
(2) Choice of the right tree size
(3) Classification of new data by constructed tree

3.1. Constriction of Maximum Tree

The CART procedure for tree growing is technically known as binary recursive partitioning. The method is a

form of binary partitioning because the data are always subdivided into two parts. It is recursive because the process is repeated for each subdivision of the data, continuation until further partitioning is impossible or is limited by some criterion set by the analyst.

The classification and regression tree (CART), an important data mining technique, is a non-parametric model without any pre-defined underlying relationship between the dependent and independent variables, which has been widely employed in many fields of study [5].

Classification tree is built in accordance with splitting rule – the rule that performs the splitting of learning sample into smaller parts. We already know that each time data have to be divided into two parts with maximum homogeneity.

Maximum homogeneity of child nodes is defined by so-called impurity function. In theory there are several impurity functions, but only two of them are widely used in practice: Gini splitting rule and Twoing splitting rule. But this paper uses Gini splitting rule.

3.1.1. Gini Splitting Rule

Gini splitting rule works faster than towing splitting rule. Gini works well for noisy data [6]. Gini splitting rule (or Gini index) is most broadly used rule. Gini index measures the divergences between the probability distributions of the target attribute's values [7]. The Gini index has been used in various works [9] and it is defined as:

$$\text{Gini}\,(y, S) = 1 - \sum_{cj \in \text{dom}(y)} \left[\frac{|\sigma_{y=cj}\, S|}{|S|}\right]^2 \quad (1)$$

Consequently the evaluation criterion for selection the attribute a_i is defined as:

$$\text{GiniGain}\,(a_i, S) = \text{Gini}(y, S) - \sum_{vi,j \in \text{dom}(ai)} \frac{|\sigma_{ai=vi,j}.S|}{|S|} . \text{Gini}\,(y, \sigma_{ai=vi,j}. S) \quad (2)$$

3.2. Choice of the Right Tree Size

Maximum trees may turn out to be of very high complexity and consist of hundreds of levels. Therefore, they have to be optimized before being used for classification of new data. Tree optimization implies choosing the right size of tree. In this case, splitting is stopped when number of observations in the node is less than predefined required minimum Nmin. Obviously the bigger Nmin parameter is the smaller the grown tree. On the one hand this approach works very fast, it is easy to use and it has consistent results. But on the other hand, it requires the calibration of new parameter Nmin. In practice Nmin is usually set to 10% of the learning sample size.

3.3. Classification of the New Data

As the classification or regression tree is constructed, it can be used for classification of new data. The output of these stages is an assigned class or response value to each of the new observations. By set of questions in the tree, each of the new observations will get to one of the terminal nodes of the tree. A new observation is assigned with the dominating class/response value of terminal node, where this observation belongs to [10].

4. Overview of the Breast Cancer Prediction System

Firstly, the system needs to collect breast cancer patient data. These collected data are training dataset. Classification and Adaptive Regressions Tree using Gini Index is applied to the patient data. The minimum threshold value is important to define to prune the tree. Normally, the number of minimum sample size is the at least ten percent of the entire data. Calculation with CART tree is also called rule generation. So the output of these steps is a set of rules. The resulted rules are used for classification of the test dataset. Rules are stored in Rule Database. The system flow diagram is mentioned in the Figure 1.

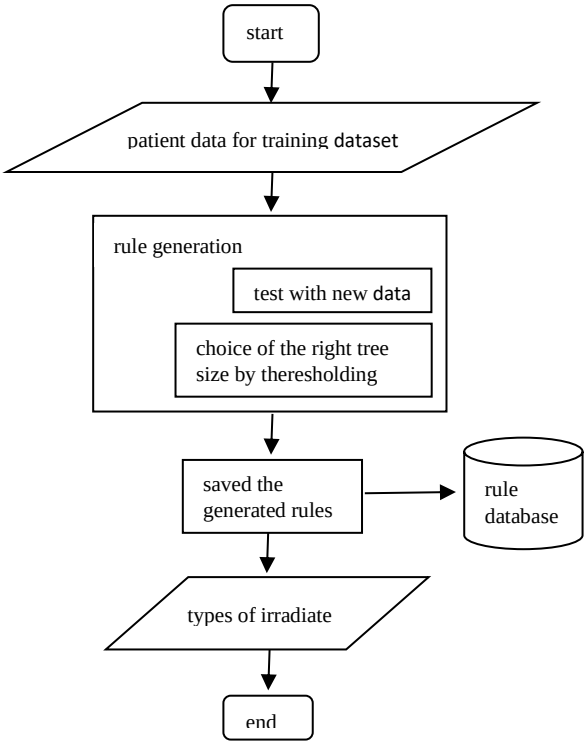


Figure 1. System flow diagram of the breast cancer patient for training dataset

For testing, the test dataset is entered and these are classified with the rules. The output of the system is the irradiate. The output of CART is form of probability. In the system, CART produces the output as two classes such as yes and no. The system flow diagram for testing dataset is mentioned in the Figure 2.

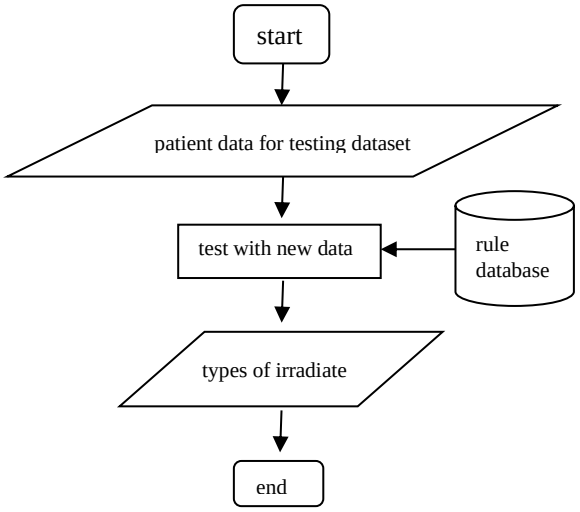


Figure 2. System flow diagram of the breast cancer patient for testing dataset

4.1. Implementation using CART

For explaining the calculation step by step, breast cancer patient data onto 13 sample data is used as in Table 1. Breast cancer dataset carried out from the UCI web data repository. The dataset includes 286 instances and 9 attributes.

Table 1. Breast cancer dataset attribute

o	men opa-use	i nv-nodes	n ode-caps	d eg-malig	bre ast-quad	ir radi-ate
	prem eno	0 -2	n o	3	right -up	n o
	prem eno	6 -8	y es	2	right -up	n o
	prem eno	6 -8	y es	2	left- up	y es
	ge40	0 -2	n o	3	right -up	n o
	ge40	0 -2	n o	2	left- low	y es
	prem eno	3 -5	n o	3	right -up	y es
	prem eno	0 -2	n o	1	left- low	y es
	Lt40	0 -2	n o	3	left- up	n o
	ge40	0 -2	n o	1	right -low	n o
0	ge40	0 -2	n o	2	left- up	y es
1	ge40	6 -8	y es	2	right -up	n o
2	prem eno	0 -2	n o	2	left- up	n o
3	prem eno	0 -2	n o	2	left- low	y es

The dataset attributes descriptions are as followed.

Table 2. Description of breast cancer attribute

Attribute Name	Description
Menopause	The period in a woman's life when menstruation ceases
Inv-nodes	Node size in main portion of the breast.
Nod-caps	Node is present or not in cap of the breast.
Deg-malig	Stage of breast cancer
Breast-quad	Portion of the breast for example left-up, left-low, right-up, right-low.
Irradiate	Present or not (yes, no)

In Table 1, the total data row count is 13. We defined the minimum row count for tree growing as 3 (Nmin=3) that is more than 10% of training data. At first, Gini Gain of 13 sample data is calculated as follows.

$$2 \text{ classes } \left(\text{yes}\frac{6}{13}, \text{no } \frac{7}{13}\right)$$

$$\text{Gini(D)} = 1 - \left(\frac{6^2}{13} + \frac{7^2}{13}\right) = 0.497$$

Gini Gain for the irradiate data is calculate from (1). It value is 0.497. After that, choosing root from five attributes (menopause, inv-nodes, node-caps, deg-malig and breast-quad). The root of the tree is one with the smallest Gini Gain comparing with others. We must calculate the Gini Gain for each attribute as follows.

$$1^{st}\text{Attribute premeno} = \frac{7}{13}, \text{ge40} = \frac{5}{13}, \text{lt40} = \frac{1}{13}$$

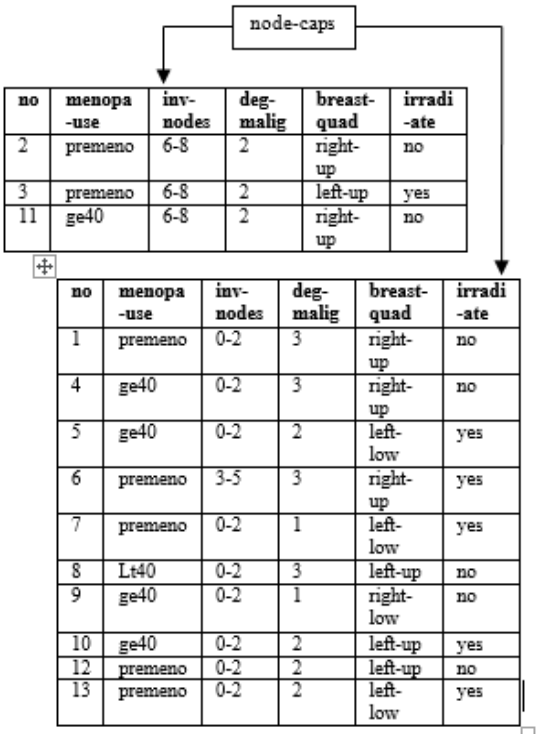
$$\text{Gini(menopause)} = \frac{7}{13}\left[1 - \left(\frac{4^2}{7} + \frac{3^2}{7}\right)\right] + \frac{5}{13}\left[1 - \left(\frac{2^2}{5} + \frac{3^2}{5}\right)\right]$$

$$= 0.2367 + 0.1846 = \text{o}.4483$$

$$\text{Gain (menopause)} = 0.497 - 0.448 = 0.049$$

First attribute is menopause and Gini Gain is 0.049 as shown in above. Second attribute is inv-nodes and Gini Gain is 0.052. Third attribute is node-caps and Gini Gain is 0.01. Fourth attribute is deg-malig and Gini Gain is 0.041. Fifth attribute is breast-quad and Gini Gain is 0.22. All of these gain values are calculated from (2). For comparing the five outputs, Gini Gain for node-caps is the smallest value among them as show in the following. Node-caps is chosen as root of the tree.

Gain (menopause) = 0.049
Gain (inv-nodes) = 0.052
Gain (node-caps) = 0.01
Gain (deg-malig) = 0.041
Gain (breast-quad) = 0.22



no	menopa-use	inv-nodes	deg-malig	breast-quad	irradi-ate
2	premeno	6-8	2	right-up	no
3	premeno	6-8	2	left-up	yes
11	ge40	6-8	2	right-up	no

no	menopa-use	inv-nodes	deg-malig	breast-quad	irradi-ate
1	premeno	0-2	3	right-up	no
4	ge40	0-2	3	right-up	no
5	ge40	0-2	2	left-low	yes
6	premeno	3-5	3	right-up	yes
7	premeno	0-2	1	left-low	yes
8	Lt40	0-2	3	left-up	no
9	ge40	0-2	1	right-low	no
10	ge40	0-2	2	left-up	yes
12	premeno	0-2	2	left-up	no
13	premeno	0-2	2	left-low	yes

Figure 3. Trees growing of CART

$$\text{deg} - \text{malig node2} = 1 - \left[\frac{1^2}{2} + \frac{0^2}{2}\right] = 0.75$$

$$\text{deg} - \text{malig node2} = 1 - \left[\frac{1^2}{2} + \frac{2^2}{2}\right] = -0.25$$

$$\text{Gini} - \text{child} = 0.75 * \frac{1}{4} - 0.5 * \frac{3}{4} = 0$$

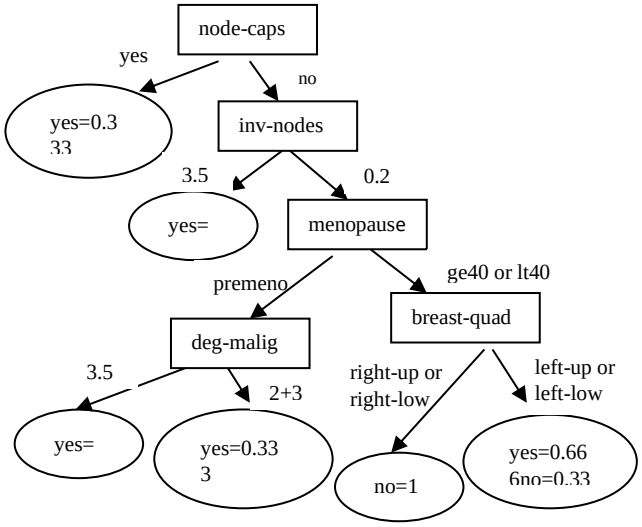


Figure 4. CART tree with Nmin=3

We can draw a tree like Figure 3. Node-caps is as root and it has two values yes and no. It splits into two values. The size of sample is 3 and 10. Right side is larger than Nmin. So, it need to continuous of growing tree.

Sample calculation of Gini splitting rule is shown by splitting menopause, deg-malig and breast-quad. In deg-malig, it has 3 values (1, 2, and 3). For best splitting, we need to calculate as in Table 3. It has two types of splitters. We choose the smaller error rate for higher homogeneity. One part of deg-malig is 1 and other part is 2 and 3. So, the right part of table 3 is chosen as best splitting. Similarly, breast-quad has 4 values (left-up, left-low, right-up, right-low). It also has two types of splitters. One part of breast-quad is right-up and right low. Other part is left-up and left-low.

Table 3. Splitting "deg-malig"

	deg-malig1		deg-malig2	
	1	2+3	1	2+3
yes	1	1	1	1
no	1	1	0	2
Gini	0.5		0	

$$\text{deg} - \text{malig node1} = 1 - \left[\frac{1^2}{2} + \frac{1^2}{2}\right] = 0.25$$

$$\text{deg} - \text{malig node1} = 1 - \left[\frac{1^2}{2} + \frac{1^2}{2}\right] = 0.25$$

$$\text{Gini} - \text{child} = 0.5 * \frac{2}{4} + 0.5 * \frac{2}{4} = 0.5$$

According to Figure 4, six rules are got from the training data. The resulted rules for training data are followed. These rules are used for classifying new data.

1. Node-caps is yes then irradiate is yes=0.333 and no=0.666.
2. Node-caps is no and inv-nodes is 3.5 then irradiate is yes=1.
3. Node-caps is no and inv-nodes is 0.2 and menopause is premeno and deg-malig is 1 then irradiate is yes=1.
4. Node-caps is no and inv-nodes is 0.2 and menopause is premeno and deg-malig is 2 or 3 then irradiate is yes=0.333, no=0.666.
5. Node-caps is no and inv-nodes is 0.2 and menopause is ge40 or lt40 and breast-quad is right-up or right-low then irradiate is no=1.
6. Node-caps is no and inv-nodes is 0.2 and menopause is ge40 or lt40 and breast-quad is left-up or left-low then irradiate is yes=0.666, no=0.333.

4.2. Classification of New Data by Constructed Tree using Patient Data

Rules which are extracted from classification and adaptive regressive tree (CART) use for classifying and prediction of new data.

Table 4. Test data

menopa-use	inv-nodes	node-caps	deg-malig	breast-quad	irradi-ate
ge40	0-2	no	2	left-up	?

Table 4 is used for testing data. The testing data match with rule 6. "If node-caps is no and inv-nodes is 0.2 and menopause is ge40 or lt40 and breast-quad is left-up or left-low then irradiate is yes=0.666, no=0.333. So, the predicted irradiate may be yes=0.666 and no=0.333.

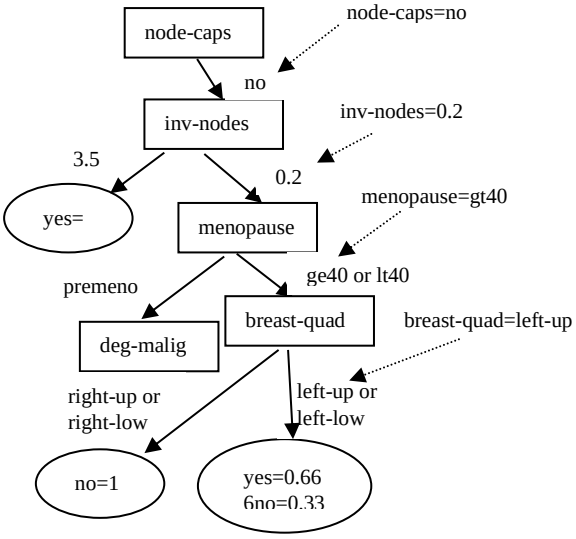


Figure 5. Testing with new data

If test data result is 'yes', the system predict that the irradiation treatment is required to kill breast cancer cells of this patient. If test data result is 'no', the irradiation treatment is not required for that patient.

5. Experimental Results

CART is predicting the probability of irradiate with percentage. We observe that the greater the number of Nmin, the shorter the tree size. Alternatively, the smaller the number of Nmin, the larger the size tree and more rules are inducted. The classification accuracy of breast cancer patient system using CART tree is varied based on the number of Nmin. The greater the number of minimum node (Nmin), rules are more general. So, it is better accuracy than using rules inducted from the smaller number of minimum node (Nmin).

Classification is using the reliable holdout method 70% of data are used as training and 30% are tested. Classification accuracy received 72% of testing. This system using CART works very fast, it is easy to use and it has consistent results.

6. Conclusion

Classification technique in data mining is an interesting topic to the researchers. Decision trees are so popular because they produce classification rules that are easy to interpret than other classification methods.

Prediction of breast cancer patient data is formulated with CART classification tree. It uses five attributes. As the method of splitting rule, Gini Index is used. Each attribute has to be divided into two parts in maximum homogeneity. Optimization by minimum number of points (N min) method is used for pruning tree. Classification and Adaptive Regression Tree can be extended by using another splitting method such as towing splitting. Moreover, it can be considered more attribute like age, tumor-size and so on. We will try to extend the experimental works for other cancer datasets to give prediction and treatment decision.

Acknowledgment

Our thanks go to the experts and my colleges who supported this article accomplish. We would like to show our gratitude to all reviewers for giving an opportunity to publish the paper.

References

[1] T.Beshah, D.Ejigu, A.Abraham, V.Snasel, "Knowledge Discovery from Road Traffic Accident Data in Ethiopia: Data Quality, Ensembling and Trend Analysis for Improving Road Safely", Department of Computer Science Ostraua, Czech Republic 2012.

[2] J.R Quinlan, "Introduction of decision tree", Journal of Machine learning", pp 81-106, 1986.

[3] Roger J.Lewis, "An Introduction to Classification and Regression Tree (CART) Analysis", Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California.

[4] Tibebe Beshah Tesema, Ajith Abraham And Crina Grosan, "Rule mining and classification of road traffic Accidents using adaptibe regression trees", Chung-Ang University, Korea, 2005.

[5] A.T. Kashani, A. Shariat-Mohaymany, A.Ranjbari. "A data mining approach to identify key factors of traffic injury severity", School of Civil engineering Narmak, Tehran, Iran Dec.21, 2010.

[6] Roman Timofeev, "Classification and Regression Trees (CART) Theory and Applications", Master Thesis, Berlin, December 20, 2004.

[7] Dillon Matthew Carty, "An Analysis of Boosted Regression Trees to Predict the Strength Properties of Wood Composites", University of Tennessee, Knozville.

[8] Lior Rokach, Oded Maimon, "Decision Trees", Department of Industrial Engineering, Tel-Aviv University.

[9] Charles Elkan, "Predictive analytics and data mining", May 28, 2013.

[10] F.Stulajter, "Prediction in Nonlinear Regression Models", vol LxVI, (1997).