Prediction of Breast Cancer Disease Using Decision Tree Algorithm

1Nwokocha, Nathan, 2Ledisi Kabari & 3Agaba, Francis

1Department of Computer Science Education
Federal College of Education (Technical) Omoku, Rivers State, Nigeria
2Ken Saro-Wiwa Polytechnics, Bori, Rivers State, Nigeria

ABSTRACT
The Healthcare industry is generally “information rich”, but unfortunately not all the data mined can be used for discovering hidden patterns and effective decision making. Advanced data mining techniques are used to discover knowledge in database and for medical research, particularly in Heart disease prediction. This paper has analysed prediction systems for Breast Cancer disease using Decision tree algorithm and WEKA 3.8 as a machine learning tool. The system uses medical terms such as Age, menopause, tumour size, inv-nodes, irradiant, node-caps, deg-malig, breast, breast-quad and class. These are the 10 attributes including class applied to predict the likelihood of patient having a breast cancer disease. The class is either no recurrence event or recurrence event.

Keywords: Breast cancer, Classification, Prediction, Decision trees.

INTRODUCTION
According to WHO (2002) cancer has been responsible for the death of millions of people Worldwide, with an estimated increase of 50% for developing countries and 70% of the total death are due to cancer. According to Parkin, Ferlay, Hamdi-Cherif, Sitas, Thomas and Wabinga (2003) developing nations only possess 5% of global funds for cancer control. It is opined that very few human and material resources are available in these countries (Grey & Sener, 2006).

Breast cancer is a malignant tumour (a collection of cancer cells) arising from the cells of the breast. Although breast cancer predominantly occurs in women, it can also affect men. This article deals with breast cancer in women. Breast cancer and its complications can affect nearly every part of the body.

Breast cancer does not always produce symptoms; women may have cancers that are so small which do not produce masses and can be felt or by other recognizable changes in the breast. When symptoms do occur, a lump or mass in the breast is the most common symptom. Other possible symptoms include

- nipple discharge or redness,
- changes in the skin such as puckering or dimpling,
- Swelling of part of the breast.

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.

This study aims at using data mining techniques to classify breast cancer risks using datasets of patients’ information from UCI Repository which contains the risk factors and the cancer classes. The decision trees classification of breast cancer was performed using the WEKA software.
Literature Review
Rajesh and Anand (2012) employed SEER dataset for the diagnosis of breast cancer using the C4.5 classification algorithm. The algorithm was used to classify patients into either pre-cancer stage or potential breast cancer cases. Random tests were performed on the dataset which contained information for 1183 patients including the age of diagnosis, regional lymph nodes measures, and sequence number of tumours, dimension of primary tumours and contiguous growth of the primary tumours. The analysis involved the use of three random 500 records form the pre-processed data of 1183 and was used as training data and the lowest error rate achieved was 0.599. During the testing phase, the C4.5 classification rules were applied to a test sample and the algorithm showed had an accuracy of 92.2%, sensitivity of 46.66% and a specificity of 97.4%. Future enhancement of the work will require the improvisation of the C4.5 algorithm to improve classification rate to achieve greater accuracy.
Shajahan et al (2013) worked on the application of data mining techniques to model breast cancer data using decision trees to predict the presence of cancer. Data collected contained 699 instances (patient records) with 10 attributes and the output class as either benign or malignant. Input used contained sample code number, clump thickness, cell size and shape uniformity, cell growth and other results physical examination. The results of the supervised learning algorithm applied showed that the random tree algorithm had the highest accuracy of 100% and error rate of 0, while CART had the lowest accuracy with a value of 92.99% but naïve bayes’ had the an accuracy of 97.42% with an error rate of 0.0258.
Mangasarian, Street and Wolberg (1995) performed classification on both diagnostic and prognostic breast cancer data. The classification procedure adopted by them for diagnostic data is called Multi Surface Method-Tree (MSM-T). This used a linear programming model to iteratively place a series of separating planes in the feature space of the examples. The procedure is recursively repeated. Moreover they have approached the prognostic data using Recurrence Surface Approximation (RSA) which used linear programming to determine a linear combination of the input features which accurately predicts the Time-To-Recur (TTR) for a recurrent breast cancer case. The training separation and the prediction accuracy with the MSM-T approach was 97.3% and 97% respectively, whereas the RSA approach was able to give accurate prediction only for each individual patient. Their drawback was the inherent linearity of the predictive models.

Data Mining Process
Data mining is defined as a process of discovering hidden valuable knowledge by analysing large amounts of data which is stored in databases or data warehouse using various data mining techniques such as machine learning, artificial intelligence(AI) and statistical. The two basic data mining tasks are: descriptive data mining tasks which help to understand the characteristic properties of dataset and predictive data mining tasks which are used to perform predictions based on available dataset. Predictive data mining is the chosen data mining task for this study.

METHODOLOGY
The breast cancer disease dataset has been taken from the UCI Repository, https://archive.ics.uci.edu/ml/machine-learning. This dataset has 286 instances and 10 attributes. The attributes are:
- Age,
- Menopause,
- Tumour size,
- Inv-nodes,
- Irradiant,
- Node- caps,
- Deg- malig,
- Breast,
- Breast- quad
- Class
Data mining Algorithm used
Decision Trees
Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets, while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor is called root node. Decision trees can handle both categorical and numerical data.

Performance Evaluation
The performance evaluation criteria allow the measurement of the accuracy of the model developed using the training dataset. The results of the classification are recorded on a confusion matrix. A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the vertical. All correct classifications lie along the diagonal from the north-west corner to the south-east corner, also called True Positives (TP) and True Negatives (TN). The other cells are called the False Positives (FP) and False Negatives (FN).

Experimental Result
The experimental results of this study used the classifier and the WEKA software data mining tool was employed. In the experiment, two classes were used and therefore a 2x2 confusion matrix was applied. Class a = No recurrence event and Class b = Recurrence event. The result obtained from use of the WEKA software is displayed in fig 5. The result summary of the classifier output is shown in Figure 6. Table 1 indicates the accuracy of the result in percentage using the Decision Tree, while fig 7 shows the result in pie-chart.
Figure 2: Result Summary of classifier output

Table 1: Accuracy Table

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correctly Classified Instance (%)</th>
<th>Incorrectly Classified Instance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>66%</td>
<td>34%</td>
</tr>
</tbody>
</table>
CONCLUSION
In this study a data mining classification technique was used for the prediction of breast cancer disease. Experimental result shows that the correctly classified instances for breast cancer are greater than the incorrectly classified instances considering 286 instances and 10 attribute.

REFERENCES