International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(2), pp.321-325
DOI: http://dx.doi.org/10.21172/1.82.043
e-ISSN:2278-621X

APPLICATIONS OF DATA MINING CLASSIFICATION TECHNIQUES ON PREDICTING BREAST CANCER DISEASE

Dr. G. Rasitha Banu[1], Dr.Prakash[2], Ms.Illham Bashier[3] and Ms.Summera[4]

Abstract: Breast cancer is a malignant growth in the breast tissue. Breast cancer is a second leading death among women today after the cervical cancer. If the breast cancer is diagnosed and treated properly then we can save human life. Data mining is a process of finding hidden pattern in huge volumes of databases. It is very useful in healthcare organization. There are several classification algorithms are available in data mining. In our work, we have used data mining classification algorithms namely Zero R, One R, decision stump and J48 to predict breast cancer and performance measures can be analyzed through confusion matrix. In our work, the J48 Algorithm is giving higher accuracy than other algorithms.

Keywords— Data Mining, Breast Cancer, Decision Trees, Classification, Prediction, accuracy, WEKA

I. INTRODUCTION

Breast cancer is highly heterogeneous disease. It represents the second important cause of cancer death in women today. Breast cancer occurs both in female and male. But very rare among men. It is the most common type of cancer in women, both in the developed and developing countries. In developed countries 1 in 8 is suffering from Breast cancer. Breast cancer is very serious malignant tumor originating from the breast cells. There are two types of Breast cancer namely benign and malignant. Benign breast cancer: This is non-invasive types of breast cancer is rarely a threat to life, occurs in the lining of milk ducts. In this type cancer doesn't spread to neighboring tissues and remains in ducts that's why called as ductal carcinoma. Malignant Breast cancer: This is invasive type of breast cancer begins in the lobules of the breast so named as lobular carcinoma. It spreads from where it began in the breast lobules to surrounding normal tissues and is a threat to life. Sometime they grow back even after removal. The symptoms of breast cancer is a lump in the breast or underarm, that persist after menstrual cycle. This lump is usually painless. Another symptom is retraction of nipple or discharge in nipple. Puckering of skin on the breast is also considered the symptom of breast cancer. Medial scientists consider that mammography screening as the most

[1] *Faculty of Public Health and Tropical Medicine, Department of HIM&T, Jazan University, Jazan.*
[2] *Dept.of CS&IS, Jazan University, Jazan.*
[3] *FPHTM, Department of Health Education, Jazan University, Jazan.*
[4] *FPHTM, Department of HIM&T, Jazan University, Jazan.*

dependable method of early detection of breast cancer. Women may survive for long time, with an early detection of breast cancer. Since breast cancer is complex disease it is likely to be caused by a combination risk factors. Age, genetic factor and heredity are non-preventable risk factors associated with breast cancer. Preventable risk factors are overweight, hormone replacement therapy, alcohol and smoking. Other risk factors are Radiation exposure, late pregnancy at older age, high bone density and post-menopausal hormone therapy. Data Mining is the process of extracting hidden and useful information from large databases. There are many data mining techniques are available such as classification, clustering, association rule mining and so on. Classification is the process of classifying similar objects. There are many classification Algorithms. This paper focuses on how data mining techniques are applied to predict breast cancer disease.

II. DATA COLLECTION

The breast cancer dataset used in this work is collected from the website (http://repository.seasr.org/Datasets/UCI/arff)[3]. This dataset consists of 286 instances and 10 Attributes. The Attributes are shown below.

S.No	Attributes	Type
1	age	nominal
2	menopause	nominal
3	Tumor size	nominal
4	Inv_ nodes	nominal
5	Node_caps	nominal
6	Deg_ malig	nominal
7	Breast	nominal
8	Breast quad	nominal
9	irradiat	nominal
10	class	Non recurrence event, recurrence event

Table1: Data Set for Breast Cancer

III. METHODS AND MATERIALS

a. Classification

Classification is one of the data mining Technique. It is used to group the instances which belong to same class. It is a supervised learning, in which predefined training data is available. Most popular data mining classification techniques are decision trees and neural networks.

b. Decision tree

Decision tree is one of the classification technique in data mining. It is tree-like graph. [5] The internal node denotes a test on attribute, each branch represents an outcome of the test,

and the leaf node represent classes. It is a graphical representation of possible solutions based on condition from these solutions optimum course of action is carried out. In our work, we have used four decision tree classifier such as decision stump, One R, Zero R and J48 to classify the breast cancer data set.

IV. EXPERIMENTS WITH WEKA

The open source software Waikato Environment for knowledge Analysis 3.7(WEKA) is used for experiment. It is a collection of machine learning algorithms for data mining tasks. Weka tool contains many data preprocessing filters, classifiers, regression, clustering, association rule mining algorithms, selected attributes and visualization. Weka can downloaded from the website[3].

Performance Measure of Classifiers

In our experiment data is supplied to classifier of J48,One R,Zero R Algorithm and decision stump to classify the data. The classifiers performance is evaluated through Confusion Matrix.

a. Confusion Matrix

It is used for measuring the performance of classifiers. In the confusion matrix, correctly classified instances are calculated by sum of diagonal elements TP (True Positive) and TN (True Negative) and others as well as FP (false positive) and FN (False Negative) are called incorrectly classified instances.

b. Accuracy

It is defined as the ratio of correctly classified instances to total number of instances in the breast cancer dataset.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

V. RESULT ANALYSIS

There are totally 286 records in the Breast cancer dataset. Among these 286 instances, 201 instances which are belongs to non-recurrent event and 85 instances are belongs to recurrent event.

The following Table 2 represents confusion matrix for ZERO R Algorithm

Target class	Non recurrence	recurrence
Non recurrence	201	0
Recurrence	85	0

Table2: Confusion matrix for ZeroR Algorithm

In ZERO R classifier, the correctly identified instances are 201 and incorrectly identified instances are 85.

The following Table 3 represents confusion matrix for OneR Algorithm.

Target class	Non recurrence	recurrence
Non recurrence	166	35
recurrence	63	22

Table3: Confusion matrix for OneR Algorithm

In OneR classifier, the correctly identified instances are 188 and incorrectly identified instances are 98.

The following Table 4 represents confusion matrix for J48Algorithm.

Target class	Non recurrence	recurrence
Non recurrence	193	8
Recurrence	62	23

Table4: Confusion matrix for J48 Algorithm

In J48 classifier, the correctly identified instances are 216 and incorrectly identified instances are 70.

The following Table 5 represents confusion matrix for Decision Stump Algorithm.

Target class	Non recurrence	recurrence
Non recurrence	201	0
Recurrence	85	0

Table5: Confusion matrix for Decision Stump Algorithm

In Decision Stump classifier, the correctly identified instances are 201 and incorrectly identified instances are 85.

The following Table 6 depicts detailed accuracy for J48 and decision stump , One R and Zero R algorithm.

Algorithm	Accuracy
One R	70.27%

Zero R	65.73%
J48	75.52%
Decision Stump	68.53%

Table6: Accuracy of Algorithms

The following chart1 shows the Accuracy of classifiers.

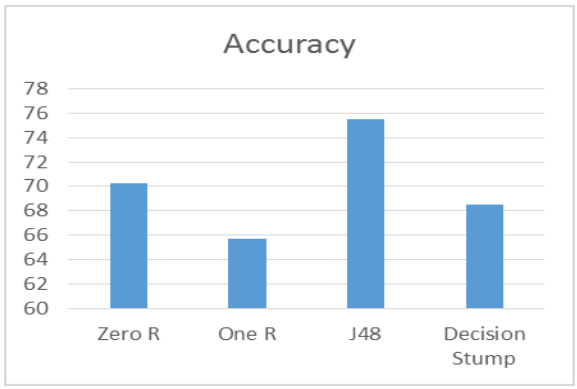

Chart1: Accuracy of classifiers

In this chart, X axis represent the algorithm and Y axis represent the accuracy. It shows that the accuracy of J48 is 75.52% which is more than other Algorithms.

VI. CONCLUSION AND FUTURE SCOPE

Diagnosis of disease is a very challenging task in the field of health care. Many data mining techniques are used in decision making process. In our work, we have used dimensionality reduction to select the subset of attributes from original data and we have applied J48, One R, Zero R and decision stump data mining classification techniques which are used to classify the recurrent and non-recurrent breast cancer disease. The performance of classifiers are evaluated through the confusion matrix in terms of accuracy. The J48 Algorithm gives 75.52% which is providing better accuracy than other Algorithm. As a future work the same technique is used to apply for other disease datasets such as heart disease, Diabetes, Lung cancer and so on.

REFERENCES

[1] Jiawei Han, Kamber Micheline (2009). Data mining: Concepts and Techniques, Morgan Kaufmann Publisher.

[2] "UCI Machine Learning Repository of machine learning database", University of California, school of Information and Computer Science, Irvine.

[3]Available from: http://www.cs.waikato.ac.nz/ ml/weka/.[Last accessed on 24].

[4] G. Ravi Kumar et.al " An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 4 August 2013

[5] A. Bellachia and E.Guvan "Predicting breast cancer survivability using data mining techniques", Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining, 2006.

[6] A. Endo, T. Shibata and H. Tanaka (2008), Comparison of seven algorithms to predict breast cancer survival, Biomedical Soft Computing and Human Sciences, vol.13, pp.11-16.

[7] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).