ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques

Sumalatha.G, Archana.S,

Assistant Professor, Dept. of Computer Technology, Sri Krishna Arts and Science, College, Coimbatore, Tamil Nadu, India

PG Scholar, Dept. of Computer Technology, Sri Krishna Arts and Science College, Coimbatore,Tamil Nadu, India.

ABSTRACT: A cancer is a disease caused by uncontrolled division of abnormal cells in a part of the body.They are various cancer in the world.one among them is breast cancer. Nowadays Breast cancer becomes very major disease in many women not only in India but also in other country The main objective of this paper is to early diagnosis of the breast cancer patients.For early prevention and detection of the breast cancer patients, data mining techniques is used.Former determination of Breast Cancer spares incredible lives, falling flat which may prompt to other forceful issues bringing on sudden fatal end. Its cure rate and expectation depend predominantly on the early identification and finding of the infection.The selection of suitable clustering data mining technique is a challenge for the diagnosis of breast cancer.This paper becomes very helpful to doctor for diagnosis breast cancer and also helpful to patients for early treatment.

KEYWORDS: breast cancer,data mining,J48 decision tree,zeroR,Weka.

I. INTRODUCTION

Cancer is one of the most common diseases in the world that results in mainstream of death caused by unrestrained growth of cells in any of the tissues or parts of the body. Breast cancer is the most widespread cancer among Women. two types of breast cancer, i.e. malignant and benign.the malignant tumor develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. Only premature detection of cancer at the benign stage and avoidance from spreading to other parts in malignant stage could save a person’s life. Although breast cancer is the second leading cause of cancer death in women, still the endurance rate is high once it is detected early.With early diagnosis, 97% of women survive for 5 years or more. There has to be the availability of exact and accurate data, so that a model with accurate model helps the doctors to predict and diagnose the cancer whether it is benign or malignant at the early stage. This will really save time for the physicians and improve their efficiency. This paper predominantly discusses the possibility to identify the breast cancer condition whether it is benign or malignant even at very early stage.The prediction condition is based on the attributes related to the breast cancer. There are three major steps that have been used in this paper i.e. collection of datasets, data preprocessing and classification. This paper explains the various phases of data mining that is performed on the dataset

II. RELATED WORKS

A literature review showed that there have been several studies on the survival prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and recurrence using data mining approaches such as decision trees [5,6]. Delen et al. used artificial neural networks, decision trees and logistic regression to develop prediction models for breast cancer survival by analyzing a large dataset, the SEER cancer incidence database [6]. Lundin et al. used ANN and logistic regression models to predict 5, 10, and 15 -year breast cancer survival. They studied 951 breast cancer patients and used tumor size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis, and age as input variables [7]. Pendharker et al. used several data mining techniques for exploring interesting patterns in breast cancer.

DOI: 10.15680/IJIRCCE.2017. 0506006

In this study, they showed that data mining could be a valuable tool in identifying similarities (patterns) in breast cancer cases, which can be used for diagnosis, prognosis, and treatment purposes [4]. These studies are some examples of researches that apply data mining to medical fields for prediction of diseases

III.PROBLEM STATEMENT

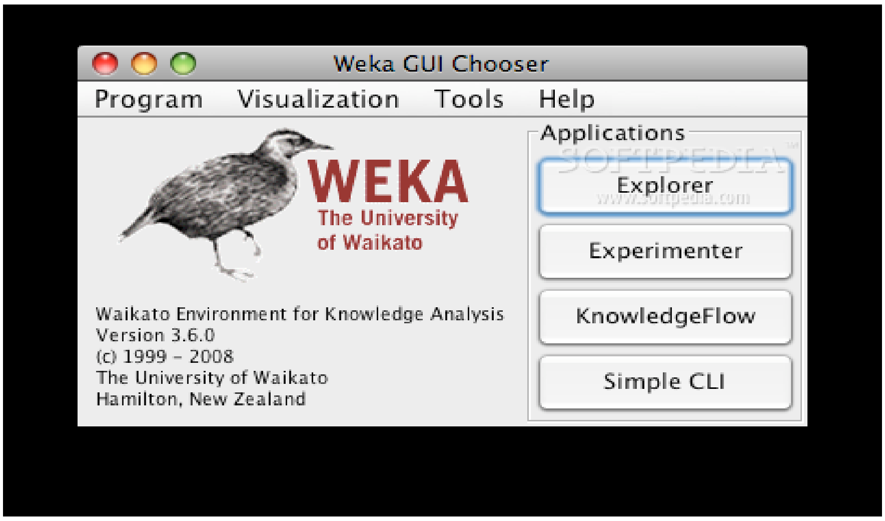
Breast Cancer is one of the leading cancer developed in many countries including India. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased drastically in last few decades. The main issue pertaining to its cure is early recognition. Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue.

IV.DATA MINING TECHNIQUES

Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. Clustering is a process of separating dataset into subgroups according to their unique features.

V. WEKA

Weka is a collection of machine learning algorithms for data mining tasks. the algorithm may be directly applied to the dataset or from the our own java code. weka contains a tool for data pre-processing, classification, regression, clustering, association rule and visualization. it also well suited for developing machine learning schemes.



Waikato Environment for Knowledge Analysis
Version 3.6.0
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand

VI. ATTRIBUTES USED

Analysis of Breast Cancer has been carried upon 10 attributes, namely,

Attributes	values
clumpThickness,	1-10
cell size uniformity	1-10
cell shape uniformity	1-10
marginal adhesion	1-10
single epithelial cell size	1-10
size of bare nuclei,	1-10
BlandChromatin,	1-10
NormalNucleoli,	1-10
Mitoses,	1-10
class	Beningn or malignant

When pathologists examine FNA(fine needle aspirate) tissues samples in breast cancer diagnosis,they consider the above nine attribute.each of the attributeis assigned to number from 1-10 by the pathologists.the larger the number the greater the likelihood of malignancy.no single measurement can be used to determine whether itis benign or malignant.

Analysis of result:

Clump thickness indicates that radius was computed by averaging the length of radial line segments from the center of the nuclear mass to each of the points of the nuclear border. For cell size, perimeter was measured as the distance around the nuclear border which is considered to be uniform. For measuring the cell shape, area is measured by counting the number of pixels in the interior of the nuclear border and adding one-half of the pixels on the perimeter. Marginal adhesion is measured bycombining the perimeter and area to give a measure of the compactness of the cell nuclei using formula: perimeter^2/area.

The analysis have been carried on using two algorithms namely, J48 and ZeroR. Total instances for ZeroR analysis is 699.

VII. ZEROR

ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is effective for determining a baseline performance as a yardstick for other classification methods. It constructs a frequency table for the target and selects its most frequent value.

ZeroR result analysis:

- Percentage Split = 66 %
- Total Instances = 699
- Attributes = 10
- Test mode: split 66.0% training set, remainder test
- ZeroR predicts class value: benign

Summary for ZeroR decision tree:

Correctly Classified Instances	152 (63.8655%)
Incorrectly Classified Instances	86 (36.1345 %)
Mean absolute error	0.4548
Root mean squared error	0.481
Total Number of Instances	38
Relative absolute error	94.07%
Root Relative squared error	97.41

Accuracy measures of zeroR decision tree:

TP rate	**FP rate**	**precision**	**recall**	class
1	1	.639	1	benign
0	0	0	0	malignant

Confusion matrix for zeroR decision tree:

classifier	**benign**	**malignant**
Session 1	152	**0**
Session **2**	86	**0**

VIII. J48

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples while the terminal nodes tell us the final value (classification) of the dependent variable. The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is resolute by, the values of all the other attributes. The additional attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

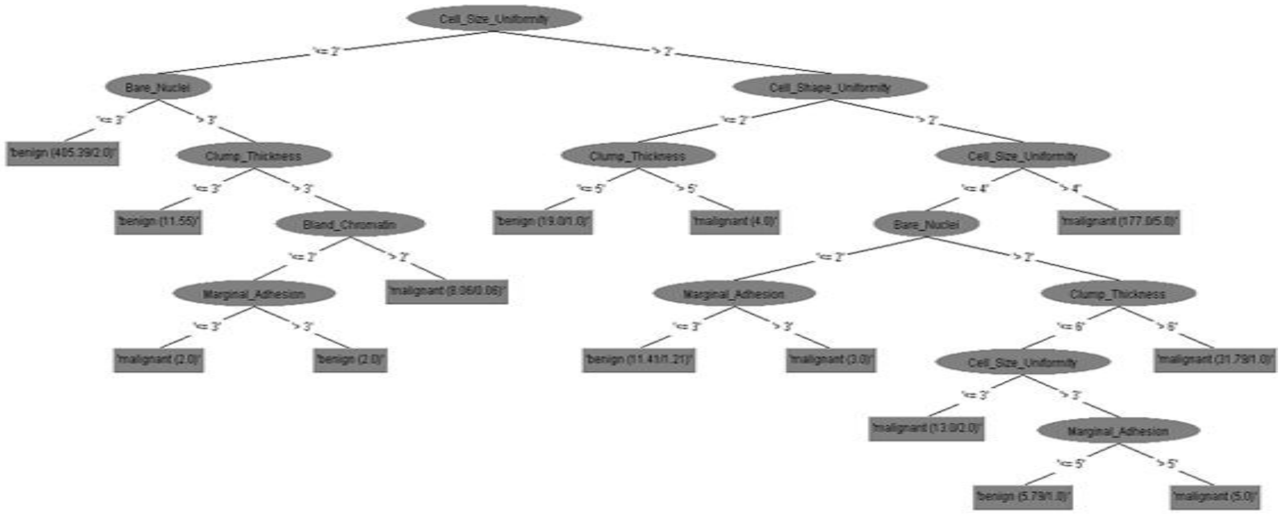
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

J48 Decision tree:



J48 Result Analysis

Test mode: split 66.0% train, remainder test

Correctly Classified Instances	227 (95.3782 %)
Incorrectly Classified Instances	11(4.6218 %)

Summary for J48:

Mean absolute error	0.0671
Root mean squared error	0.2124
Total Number of Instances	238
Relative absolute error	14.7632%
Root Relative squared error	44.1621%

Accuracy measures of J48**:**

TP rate	**FP rate**	**precision**	**recall**	class
0.954	0.047	0.973	0.954	benign
0.953	0.046	0.921	0.953	malignant

DOI: 10.15680/IJIRCCE.2017. 0506006

Confusion matrix for J48**:**

classifier	benign	malignant
Session 1	145	7
Session **2**	4	82

IX. CONCLUSION

Breast Cancer has become the foremost cause of death worldwide for womens. The most successful way to reduce cancer deaths is to detect it earlier. Many people avoid cancer screening due to the cost involved in taking numerous tests for diagnosis.This prediction system may provide easy and a cost effective way for screening cancer and may play a significant role in earlier diagnosis process for different types of cancer and provide effective preventive approach.This system can also be used as a source of record with detailed patient history in hospitals as well as help doctors to premeditated on particular therapy for any patient.

REFERENCES

[1] G. Holmes; A. Donkin and I.H. Witten (1994). "Weka: A machine learning workbench". Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.

[2] Vikas Chaurasia “Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability” International journal of Computer Science and Mobile Computing (IJCSMC), Vol.3, Issue. 1, January 2014, pg.10-22, ISSN: 2320-088X

[3]Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo and RueffJose, “A Data Mining approach for detection of high-risk Breast Cancer groups,” Advances in Soft Computing, vol. 74, pp. 43-51,2010.

[4] G. Holmes; A. Donkin and I.H. Witten (1994). "Weka: A machine learning workbench". Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.

[5] Vikas Chaurasia “Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability” International journal of Computer Science and Mobile Computing (IJCSMC), Vol.3, Issue. 1, January 2014, pg.10-22, ISSN: 2320-088X

[6] Ritu Chauhan “Data clustering method for Discovering clusters in spatial cancer databases” International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010.

[7] Dechang Chen “Developing Prognostic Systems of Cancer Patients by Ensemble Clustering” Hindawi publishing corporation, Journal of Biomedicine and Biotechnology Volume 2009, Article Id 632786.

[8] S M Halawani “A study of digital mammograms by using clustering algorithms” Journal of Scientific & Industrial Research Vol. 71, September 2012, pp. 594-600.