Detection of Breast Cancer using Data Mining Tool (WEKA)

Jyotismita Talukdar
Centre of Information Technology
University of Technology and Management
Shillong, India
jyotismita4@gmail.com

Dr. Sanjib Kr. Kalita
Dept. of Computer Science
Gauhati University
Assam, India
sanjib959@rediffmail.com

Abstract — Breast cancer has become the primary reason of death in women in developed countries. Breast cancer is the second most common cause of cancer death in women worldwide. The high incidence of breast cancer in women has increased significantly during the last few decades. In this paper we have discussed various data mining approaches that have been utilized for early detection of breast cancer. Breast Cancer Diagnosis is distinguishing of benign from malignant breast lumps. We have approached the diagnosis of this disease by using Data mining technique. Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns. The most effective way to reduce breast cancer deaths is to detect it earlier. This paper discusses the early detection of breast cancer in three major steps of determining the breast cancer. They include (i) collection of data set, (ii) preprocess of the data set and (iii) classification. Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on medical datasets using multiple classifiers and feature selection techniques. A good amount of research on breast cancer datasets is found in literature. Many of them show good classification accuracy. For classification we have chosen J48.All experiments are conducted in WEKA data mining tool. Data-Sets are collected from online repositories which are of actual cancer patient

Key Words- Breast Cancer, Data Mining, WEKA, J48 Decision Tree, ZeroR

INTRODUCTION

In this research paper we have proposed the diagnosis of breast cancer using data mining techniques. Breast cancer is the most common cancer among Women. Out of the 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. Industrialized nations such as the United States, Australia, and countries in Western Europe witnessed the highest incidence rates of breast cancer. Although breast cancer is the second leading cause of cancer death in women, still the survival rate is high once it is detected early. With early diagnosis, 97% of women survive for 5 years or more. In the healthcare industry, it is vital to understand the gradual developments of such tumor. There has to be the availability of precise 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 primarily 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 10 attributes in the data set used in this paper. These data will help

the physicians to decide which attributes are more important for early prediction. 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. We have used WEKA as a data mining tool.

PROBLEMSTATEMENT

Breast Cancer is one of the leading cancer developed in many countries including India. Though the survival 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 significantly in last few decades. The main issue pertaining to its cure is early detection. Hence, apart from medicinal solutions some Data Science solution needs to be incorporated for resolving the death causing issue.

TheoreticalBackground

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 useful for determining a baseline performance as a benchmark for other classification methods. It constructs a frequency table for the target and selects its most frequent value.

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 decided by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset.

Medical Data Mining

In 2011, the case of Sorrell v. IMS Health, Inc., decided by the Supreme Court of the United States, ruled that pharmacies may share information with outside companies. The practice was authorized under the 1st Amendment of the Constitution, protecting the "freedom of speech." However, the passage of the Health Information Technology for Economic and **Clinical Health Act** (HITECH Act) helped to initiate the adoption of the electronic health record (EHR) and supporting technology in the United States. The HITECH Act was signed into law on February 17, 2009 as part of the American Recovery and Reinvestment Act (ARRA) and helped to open the door to medical data mining. Prior to the signing of this law, estimates of only 20% of United States based physician were utilizing electronic patient records. Soren Brunak notes that “the patient record becomes as information-rich as possible” and thereby “maximizes the data mining opportunities.” Hence, electronic patient records further expands the possibilities regarding medical data mining thereby opening the door to a vast source of medical data analysis.

Data Mining Tasks

i. Classification

ii. Clustering
iii. Association Rule Discovery
iv. Sequential Pattern Discovery
v. Regression
vi. Deviation Detection

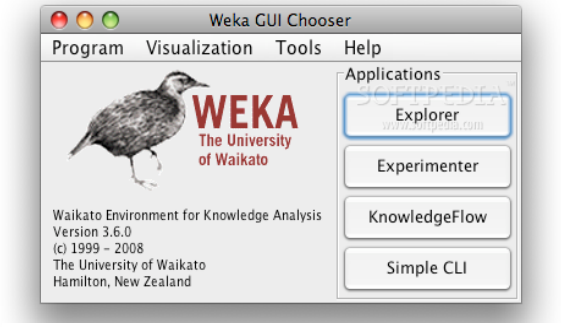

Figure 1: WEKA data mining tool

Results of Analysis

The analysis of Breast Cancer has been carried upon 10 attributes, namely, ClumpThickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, size of bare nuclei, BlandChromatin, NormalNucleoli, Mitoses, class. 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. Following is the detailed analysis of both ZeroR and J48 algorithm.

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

Evaluation on test split for ZeroR

Table 1: Summary for ZeroR decision tree

Correctly Classified Instances	152 (63.8655%)
Incorrectly Classified Instances	86 (36.1345 %)
Kappa statistic	0.013
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%

Detailed Accuracy by Class

Table 2: Accuracy measures for ZeroR decision tree

TP Rate	FP Rate	Precision	Recall	Class
1	1	.639	1	benign
0	0	0	0	malignant

Confusion Matrix

Table 3: Confusion matrix for ZeroR decision tree

Classifier	Benign	Malignant
Session 1	152	0
Session 2	86	0

J48 Result Analysis

Test mode: split 66.0% train, remainder test

Table 4: Instances for J48

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

Evaluation on test split for J48

Table 5: Summary for J48

Kappa statistic	0.9006
Mean absolute error	0.0671
Root mean squared error	0.2124
Relative absolute error	14.7632 %
Root relative squared error	44.1621 %
Total Number of Instances	238

Detailed Accuracy by Class

Table 6: Accuracy measures for 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

Confusion Matrix

Table 7: Confusion matrix for J48

Classifier	Benign	Malignant
Session 1	145	7
Session 2	4	82

J48 Decision Tree

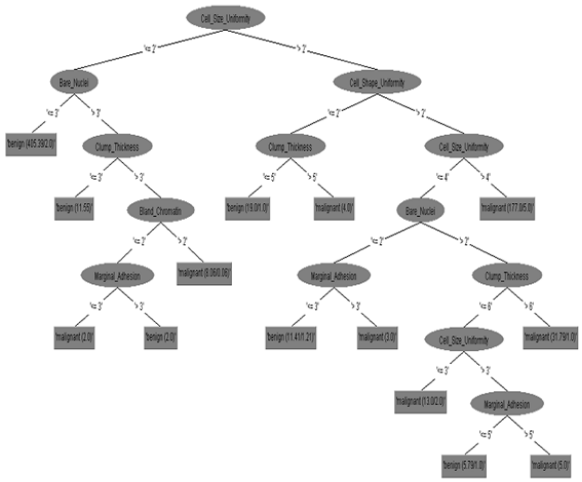

Figure 2: J48 decision tree

By using data mining we can predict the occurrence of breast cancer most efficiently. For early detection, we must know the attributes that are present in the pathology report. We have developed patterns via which , we can select the important attributes of Breast Cancer for early, efficient and accurate detection of it so that it can be properly medicated upon in time. The present study could be extended with more number od patients integrating more alike institutions or organizations. With the help of clous computing facilities the results so obtained could be shared among the institutions and thus helping the disgnosis process most affectively.

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] S.R. Garner; S.J. Cunningham, G. Holmes, C.G. Nevill-Manning, and I.H. Witten (1995). "Applying a machine learning workbench:

Experience with agricultural databases". Proc Machine Learning in Practice Workshop, Machine Learning Conference, Tahoe City, CA, USA. pp. 14–21. Retrieved 2007-06-25.

[3] P. Reutemann; B. Pfahringer and E. Frank (2004). "Proper: A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners". 17th Australian Joint Conference on Artificial Intelligence (AI2004). Springer-Verlag. Retrieved 2007-06-25.

[4] Breast Cancer Wisconsin (Original) Data-Set[Online] Available : https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

[5] [1] Ian H. Witten; Eibe Frank; Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco. Retrieved 2011-01-19.

[6] http://134.208.26.59/INA/Cancer_Diagnosis.pdf