OUR HERITAGE
ISSN: 0474-9030 Vol-68, Special Issue-27 (Feb. 2020)
5th International Conference On "Innovations in IT and Management"
Organised by: Sinhgad Technical Education Society's
SINHGAD INSTITUTE OF MANAGEMENT AND COMPUTER APPLICATION (SIMCA),
Narhe Technical Campus, Pune, Maharashtra (India) 411041.
Held on 6th & 7th February 2020

A study for Data Mining methods on Brest Cancer

Dr. Rajneeshkaur Bedi
Asst Professor
MITCOE -Pune
rajnibedi@mitcoe.edu.in
9881127706

Ms. Nita Khatri
Asst Professor
Sinhgad College of Science Pune
nitak.scos@sinhgad.edu
9921934538

Abstract

Breast Cancer is leading cancer disease found in most of the Indian woman. The latest data shows increase in theoccurrence of breast cancer and a lot of organizations are taking up this cause of spreading awareness about breast cancer. Many researchers now a days working on developing various techniques and tools from data mining for prediction and analysis. Our paper is study on this related work. All such computer aided development will help oncologists to perform timely detection and resolve the issues in breast cancer treatment.

Introduction

Human body is made up of millions of cells. The ratio and proportion is almost balance throughout the body parts. When these cells starts growing, dividing itself into multiple and spreading into surrounding tissues in un-control way then its termed as 'Cancer'. It can happen in any part of body like lungs, blood, breast, skin etc. are called as lung cancer, blood cancer, breast cancer and skin cancer respectively. Many cancers form a solid tumor which is the masses of tissues. The cancer cells can hide themselves from our immune system or make a use of it to grow more. Cancer is genetic disease where genes affects the controlling of cells growth or division. Brest cancer is very serious in female and this situation is worst in Indian females.Now a days we are finding more and more number of breast cancer patient of younger age group in India which is alarming. This can be seen from the data representation given by [1] as follows:

The growth of breast cancer in the age group of 20 to 50 years increased tremendously as compare to 25 years back. In India female pays less attention to their personal health because of which breast cancer in most of the women are detected at last stage. This make the survival of them difficult as compare to western countries where most of them get detected at stage 1 or 2. The major problem regarding this in Indian scenario is lack of awareness. So, the survival of patient suffering from breast cancer is low.

Figure 1. Statistical survey of breast cancer in India

OUR HERITAGE
ISSN: 0474-9030 Vol-68, Special Issue-27 (Feb. 2020)
5th International Conference On "Innovations in IT and Management"
Organised by: Sinhgad Technical Education Society's
SINHGAD INSTITUTE OF MANAGEMENT AND COMPUTER APPLICATION (SIMCA),
Narhe Technical Campus, Pune, Maharashtra (India) 411041.
Held on 6th & 7th February 2020

Brest Cancer

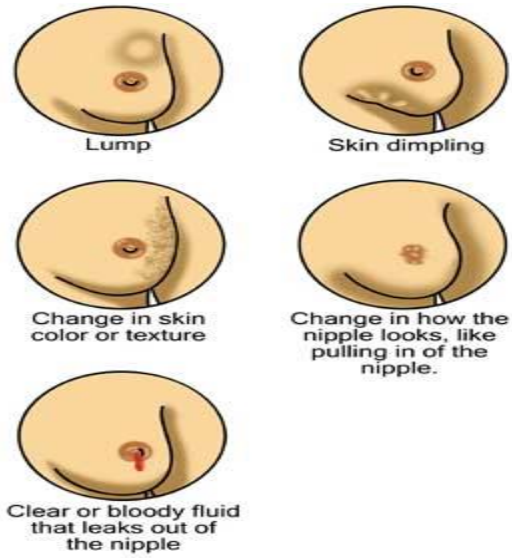
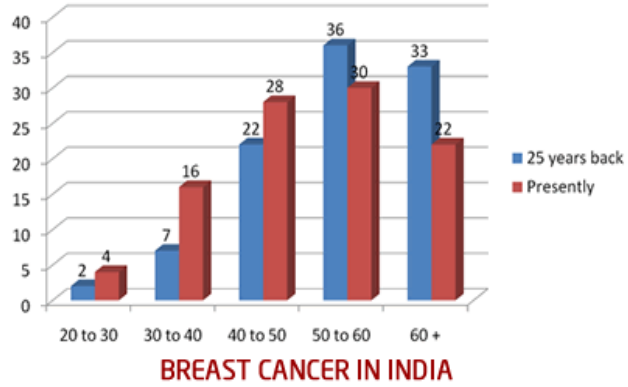


Figure 2. Symptons of breast cancer



This cancer usually develops at two places either inner lining of milk duct or lobules which supply milk called ductal carcinoma and lobular carcinoma respectively. This cancer can happen in male and female but the percentage of female suffering is high. Breast consist of billions of microscopic cells like other part of our body which undergo multiplication/duplication to form a new cell and to replace old dead cell. The common symptoms are like pain in breast or armpit, rashes around nipples, size or shape of breast changes etc as shown in the figure 2 referred from [2]. In any such case it's always recommended to undergo medical diagnosis.

Data Mining Methods

Huge amount of data is collected by various sectors like government agencies, healthcare, social media, IT sector, scientific data, geographical data etc. This data is meaningless unless and until some inference is drawn from it. Data mining is a process of extracting useful information from this data which can be used as a knowledge base for further studies. Knowledge mining from this data can be better understand by the diagram given in figure. 3 referred from [3] called as Knowledge discovery process.

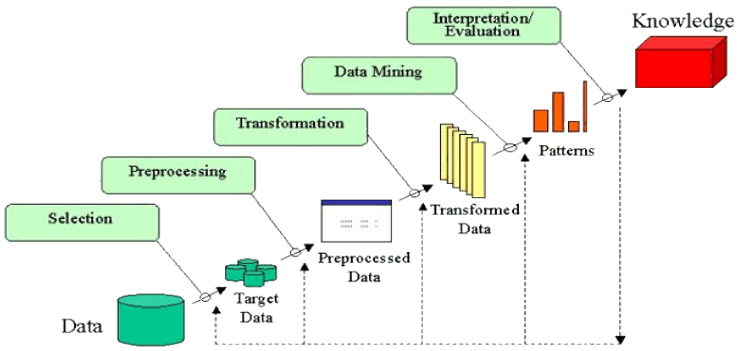


Figure 3: Knowledge discovery Process

The goal of data mining methods [3] [4] and [5] is to find useful patterns from previously unknown facts. Once these patterns are found they can further be used to make certain decisions and suggestions in the field whose data mining is carried out.
Three steps involved are:

- Exploration
- Pattern identification
- Deployment for solutions

Data Mining Algorithms and Techniques
Various algorithms and techniques as shown in figure 4 like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

Types of classification models:

- Classification by decision tree inductionBayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

Figure 4: Data mining Paradigms

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

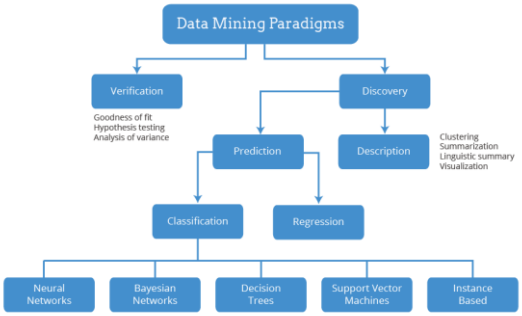
Prediction: Regression technique can be adapted for predication. Types of regression methods

- Linear Regression

- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule



Literature Review

A paper presented by S. Kulkarni and M. Bhagwat [10] evaluates various data mining techniques and their ability to predict whether any particular patient will face a recurrence. The result of their experimentshows the accuracy of various classifiers when applied on the Breast Cancer Dataset available online. According to the results presented in their paper it is clear that k-Nearest Neighbor provides a consistently high accuracy of predicting Breast Cancer recurrence as compared to the jRip classifier. The paper presented by [6] focused on automatic diagnosis system for detecting breast cancer based on association rules (AR) and neural network (NN). In their research, AR is used for reducing the dimension of breast cancer database and NN is used for intelligent classification. The proposed AR + NN system performance is compared with NN model. The dimension of input feature space is reduced from nine to four by using AR. This research demonstrated that the AR can be used for reducing the dimension of feature space and proposed AR + NN model can be used to obtain fast automatic diagnostic systems for other diseases. DursunDelen et.al [8] presented a paper on the prediction model of breast cancer survivability. The focus of their work is based around two data mining algorithm namely ANN + DT and logical regression. The result of their experiment showshow the decision tree (C5) is the best predictor and artificial neural networks and the logistic regression models followed by respectively. The comparative study of multiple prediction models for breast cancer survivability provided an insight into the relative prediction ability of different data mining methods. The work proposed by Shajahaanet.al[9]identified theapplicability of decision treesto predict the presence of breast cancer. The experimental

result shows random tree method gives better result when it is compared for the performance of traditional supervised learning algorithms viz. Random tree,ID3,CART,C4.5 and Naive Bayes. Joana Dizet.al[7] worked on new model to reduce diagnostic tests false-positive. Their work used data mining approach to support oncologists in the process of breast cancer classification and diagnosis. The result compare two breast cancer datasets and find the best methods in predicting benign/malignant lesions, breast density classification, and even for finding identification (mass / microcalcification distinction). Among the different tests classifiers, Naive Bayes was the best to identify masses texture, and Random Forests was the first or second best classifier for the majority of tested groups.

Conclusion:

In this paper, we presented a report on a research effort taken by the research community using different data mining techniques applied for breast cancer detection and survivability. Some of the approaches geared up to overcome human intervention to extract knowledge from the available breast cancer dataset. Few new solutions are also suggested like use of website to support decision system, prediction models etc. Inspite of these there is still a need of new methods and models to work on real time data and self- analysis tools which will be our area of research.

References:

1. http://www.breastcancerindia.net/statistics/trends.html
2. http://www.medicalnewstoday.com/articles/37136.php
3. Han, Jiawei, Jian Pei, and MichelineKamber. Data mining: concepts and techniques. Elsevier, 2011.
4. Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
5. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17, no. 3 (1996): 37.
6. Karabatak, Murat, and M. CevdetInce. "An expert system for detection of breast cancer based on association rules and neural network." Expert systems with Applications 36, no. 2 (2009): 3465-3469.
7. Diz, Joana, GoretiMarreiros, and Alberto Freitas. "Applying Data Mining Techniques to Improve Breast Cancer Diagnosis." Journal of medical systems 40, no. 9 (2016): 203.
8. Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine 34, no. 2 (2005): 113-127.
9. S. Syed Shajahaan, S. Shanthi, V. ManoChitra, "Application of Data Mining Techniques to Model Breast Cancer Data", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 11,November 2013
10. Siddhant Kulkarni and MangeshBhagwat, "Predicting Breast Cancer Recurrence using Data Mining Techniques", International Journal of Computer Applications (0975 –8887)Volume 122, No.23, July 2015