Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN: 2348-4470
p-ISSN: 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 02 February -2018

A THERORETICAL SURVEY ON BREAST CANCER PREDICTION USING DATA MINING TECHNIQUES

Disha Patel[1], Mr Bhavesh Tanawala[2], Mr Pranay Patel[3]

[1]*Computer Department, BVM Engineering College, V.V.Nagar, Gujarat, India*
[2] *Computer Department, BVM Engineering College, V.V.Nagar, Gujarat, India*
[3] *Computer Department, BVM Engineering College, V.V.Nagar, Gujarat, India*

Abstract — *Breast cancer has reason for the leading cause of death in lady in various countries .The popular effective way to decrease breast cancer deaths is detect it as earlier as possible. An early diagnosis method requires a more accurate and user reliable diagnosis techniques those are allow physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. Actually, Cancer research is a clinical and biological research area. Predicting the results of a disease is one of the best interesting and challenging problems where to develop data mining applications. The main goal of this study is to summarize various review and technical articles on predicting breast cancer.*

Keywords- *Data Mining techniques: Classification and Clustering techniques, Breast Cancer: Diagnosis; prognosis.*

I. INTRODUCTION

Cancer is a term used for diseases in which abnormal cells divide without control and are able to invade other tissues .Cancer cells spread to other parts of the body through the blood and lymph systems. Cancer is no just one disease but many diseases. There are more than 100 different types of cancer .Most cancers are named for the organ or type of cell in which they start. [1][2]

There are two general types of cancer tumors namely:

- Benign
- Malignant

A benign tumor is a mass of cells that lacks the ability to invade neighboring tissue or metastasize. They do not spread into nearby tissues. A malignant tumor is produce breast cancer. A collection of cancer cells from the cells the breast. And it is invade neighboring tissues and organ blood.

1.1 Breast Cancer

Breast cancer is one of the second leading causes of cancer death in women. Despite the fact that cancer is preventable and curable in primary stages , the vast number of patients are diagnosed with cancer very late. [1] Breast cancer is one of the most commonly occurring epithelial malignancies in women and there are estimated 1 million new cases and over 400,000 deaths annually worldwide. In the past twenty years, the incidences of breast cancer continue to rise. Then, the diagnosis and treatment of the breast cancer have become an extremely urgent work to do. [3]

The purpose of this study is to build diagnosis model for breast cancer. In other word, we intend to search the relationship between breast cancer and its symptoms. Data mining techniques are applied to build the prediction model and in data mining fields, searching the relationships between diseases and their symptoms is a classification problem. [3]

1.2 Symptoms of Breast Cancer:

- A lump in a breast
- A rash around (or on) one of the nipples
- A swelling (lump) in one of the armpits
- An area of thickened tissue in a breast
- The size or the shape of the breast changes
- One of the nipples has a discharge; sometimes it may contain blood
- The nipple changes in appearance; it may become sunken or inverted
- A pain in the armpits or breast that does not seem to be related to the woman's menstrual period
- Pitting or redness of the skin of the breast; like the skin of an orange

II. RELATED WORKS AND LITERATURE SURVEY

Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree Subrata Kumar Mandal Information Technology Department, Jalpaiguri Government Engineering College, Jalpaiguri, West Bengal, India, International Journal Of Engineering And Computer Science . Feb., 2017. In these paper Author have applied techniques namely data cleaning, feature selection, feature extraction, data discretization and classification for predicting breast cancer as accurately as possible. Our study reveals that Logistic Regression Classifier gives the maximum accuracy with are duced subset of features (four) and time complexity of this algorithm is least compared to other two classifiers.[1]

Intelligent Breast Cancer Prediction Model Using Data Mining Techniques Runjie Shen , Yuanyuan Yang , Fengfeng Shao , Department of Control Science & Engineering Tongji University Shanghai, China . 2014 IEEE. In this paper, we intend to build a diagnostic model of breast cancer by using data mining techniques. A feature selection method, INTERACT is applied to select relevant features for breast cancer diagnosis, and the support vector machine is used to build the classification model. Two diagnostic models are built with and without feature selection for the sake of proving the significance of the feature selection. Through the experiments, the accuracy of the diagnostic model with feature selection is improved obviously compared with the model without feature selection .So as compare to other techniques with feature selection (INTERACT) is best accuracy 92%.[3]

A Comparative survey on data mining techniques for breast cancer diagnosis and prediction. Hamid karim khani zand Department of computer engineering, Iran University of science and technology, Tehran, Iran. These paper using SEER data set and compare the data mining techniques. And also they prediction on breast cancer data. So they summaries techniques and conclusion is classification algorithm is best prediction as compare to clustering algorithm.[5]

A Study on Prediction Of Breast Cancer Recurrence Using Data Mining Techniques. Uma Ojha Computer Science Department ARSD College, Delhi University Delhi-India And Dr. Savita Goel Sr.System Programmer IIT Delhi. IEEE 2017. In this paper, they use different data mining algorithms to predict all those cases of breast cancer that are recurrent using Wisconsin Prognostic Breast Cancer (WPBC) dataset from the UCI machine learning repository .The C5.0 and SVM were best predictor algorithm of 0.813 and Fuzzy clustering came worst predictor of 0.3711. And also used KNN, PAM and EM .but best classifier is C5.0 and SVM to predict highest accuracy. The result indicates the decision tree and SVM is the best predictor with 81% accuracy. [6]

Data Mining Techniques in Multiple Cancer Prediction Dr. A. R. Pon Periasamy Associate Professor of Computer Science Nehru Memorial College Puthanampatti, Trichy (DT) Tamilnadu, India K. Arutchelvan Assistant Professor / Programmer Department of Pharmacy AnnamalaiUniversity,ChidamparamTamilnadu, India ,IJARC May 2017 . In this paper,. And also used different data mining techniques which are classification, clustering and association mining. See5 is having the higher precision on correlation. [7]

III. METHODOLOGY

3.1. Data Source:

In Order to find the best predictor model that can predict the recurrence cases of breast cancer, the authentic dataset has been used. In Wisconsin Breast Cancer Database (1991) University of Wisconsin Hospitals, There are 699 Number of instances, 10 plus class attributes and 16 missing values.

3.2. Data mining:

Data mining is the process of extracting interesting patterns and knowledge from the data. This paper focuses on using some of the clustering and classification models to predict the chances of recurrences and survivability of the diseases. A short description of these algorithms and their specific research. [6]

There are main two techniques which are classification algorithms and clustering algorithms. The classification is a function that assigns items in a collection to target categories or classes. The goal is to accurelty predict the target class for each case in the data. Some classification algorithms are Decision tree, Naïve Bayes, KNN, and Neural Network and so on. Clustering is an unsupervised learning method is different from classification. It is mainly used for analyzing data.[5] It deals with finding a structure in a collection of unlabeled data. There are some used clustering algorithms like a K-means, Fuzzy means, PAM, EM (Exception Maximization) and so on.

3.2.1. Decision tree:

In Decision tree, it’s a structure that includes root node, branches, and leaf nodes. Each node denotes a test on an attribute, each branch denotes the outcomes of a test and each leaf nodes holds a class label. There are using different algorithm which are ID3, C4.5 and c5.0. The C5.0 algorithm is a decision tree that recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. [9]

The classifier is tested first to classify unseen data and for this purpose resulting decision tree is used. C4.5 algorithm follows the rules of ID3 algorithm. Similarly C5 algorithm follows the rules of algorithm of C4.5. C5 algorithm has many features like:

- The large decision tree can be viewing as a set of rules which is easy to understand.
- C5 algorithm gives the acknowledge on noise and missing data.
- Problem of over fitting and error pruning is solved by the C5 algorithm.
- In classification technique the C5 classifier can anticipate which attributes are relevant and which are not relevant in classification.[10]

3.2.2. KNN (K-Nearest Neighbor):

The KNN is a non-parametric method used for classification and regression .In both cases the input consist of the k closet training data in the feature space. The output depends on whether KNN is used for classification and regression. If K=1 then the object is simply assigned to the class of that single nearest neighbor so they know is KNN classification and otherwise the output is the property value for the object is called KNN regression.

The object is classified by a majority vote of its neighbors with the object being assigned to the class most common among K nearest neighbors.

3.2.3. Naïve Bayes:

It is a quick method for creation of statistical predictive models. NB is based on the Bayesian theorem. These classification techniques analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationship between the attribute values and the class. The probability of each class is a computed by counting how many times its occurs in the dataset. This is called “prior probability” P(C=c).

3.2.4. K-Means:

It is one among the algorithms of partitioning methods. It is exceptionally straightforward and it can be effortlessly utilized for taking care of a large portion of the useful issues. It is the best squatted error based clustering algorithm. The n observation into k-sub classes defined by centeriod. Its NP-hard problem.

3.2.5. Fuzzy Means:

It’s used more than one cluster .it is a soft clustering algorithm .it is collection of finite elements. Clustering involves assigning data points to clusters such that items in the same cluster are as possible, while items belonging to different clusters are as possible.

3.3. COMPARISON TABLE

3.3.1. Comparison on clustering algorithm:

Algorithms	Confusion Matrix	Accuracy	Sensitivity	Specificity
K-means	N R N 100 48 R 23 23	0.6340	0.8130	0.3239
EM	N R N 117 31 R 31 15	0.6840	0.7905	0.3260
PAM	N R N 64 84 R 29 17	0.4175	0.4324	0.1683
Fuzzy c-means	N R N 50 98 R 24 22	0.3711	0.5934	0.1833

Table 1 comparison on clustering algorithm [4]

3.3.2. Comparison on classification algorithm:

Algorithms	Confusion Matrix	Accuracy	Sensitivity	Specificity
C5.0	N R N 47 0 R 11 0	0.8130	1.0	0.0
KNN	N R N 47 0 R 11 0	0.7068	0.8297	0.2
Naïve Bayes	N R N 47 0 R 11 0	0.5344	0.5319	0.2142
SVM	N R N 47 0 R 11 0	0.8103	0.8404	0.1036

Table 2 Comparison on classification algorithm [4]

IV. CONCLUSON

The main goal medical data mining algorithms is to get best algorithms that describe given data from multiple aspects. Cancer is a term used for diseases in which abnormal cells divide without control and are able to invade other tissues. From the various research paper, it is identified that the comparison between classification and clustering algorithms and also find out the best algorithm to be based on their accuracy. So that according to my survey the best algorithm is classification decision tree (C5.0) algorithm based on their accuracy.

ACKNOWLEDGMENTS

Disha Patel would like to express thanks to Mr. Bhavesh Tanawala and Mr.Pranay Patel Assistant Professor in the Computer Engineering Department, Birla Vishvakarma Mahavidyalaya for supporting the outstanding information of this research.

REFERENCES

[1] Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree Subrata Kumar Mandal Information Technology Department, Jalpaiguri Government Engineering CollegeJalpaiguri, West Bengal, India,International Journal Of Engineering And Computer Science . Feb., 2017

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

[3] Intelligent Breast Cancer Prediction Model Using Data Mining Techniques RunjieShen , Yuanyuan Yang , Fengfeng Shao , Department of Control Science & Engineering TongjiUniversityShanghai, China . 2014 IEEE

[4] A comparative survey on data mining techniques for breast cancer diagnosis and prediction Hamid Karim Khani Zand Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran. Indian Journal of Fundamental and Applied Life Sciences ,2015

[5] A Survey on Breast Cancer Analysis Using Data Mining Techniques B.Padmapriya, T.Velmurugan 2014 IEEE

[6] A Study on Prediction Of Breast Cancer Recurrence Using Data Mining Techniques. Uma Ojha Computer Science Department ARSD College, Delhi University Delhi-India And Dr. SavitaGoelSr.System Programmer IIT Delhi. IEEE 2017

[7] Data Mining Techniques in Multiple Cancer Prediction Dr. A. R. PonPeriasamy Associate Professor of Computer Science Nehru Memorial College Puthanampatti, Trichy (DT) Tamilnadu, India K. Arutchelvan Assistant Professor / Programmer Department of Pharmacy AnnamalaiUniversity,ChidamparamTamilnadu, India ,International Journal of Advanced Research in Computer Science and Software Engineering May 2017

[8] Breast Cancer Prediction using Data Mining Techniques JyotsnaNakte Student, Dept. of Information Technology MCT Rajiv Gandhi Institute ofTechnology Mumbai, India, VarunHimmatramka Student, Dept. of Computer

Engineering MCT Rajiv Gandhi Institute of Technology Mumbai, India , International Journal on Recent and Innovation Trends in Computing and Communication nov-2016

[9] C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning Rutvija Pandya Diploma Computer Engineering Department, Gujarat Technological University Atmiya Institute of Tech & Sci Rajkot Jayati Pandya Bachelar in Computer science and Application,Saurashtra University K.P.Dholakiya Infotech Amreli International Journal of Computer Applications (0975 – 8887) Volume 117 – No. 16, May 2015.

[10] Performance Analysis of Data Mining Classification Techniques on Public Health Care Data Tanvi Sharma1, Anand Sharma2, Prof. Vibhakar Mansotra M. Tech Research Student, Dept. of Computer Science & I.T., University of Jammu, Jammu, J&K, India Research Scholar, Dept. of Computer Science & I.T., University of Jammu, Jammu, J&K, India Professor, Dept. of Computer Science & I.T., University of Jammu, Jammu, J&K, India IJIRCE June 2016.