RESEARCH ARTICLE OPEN ACCESS

Review on Breast Cancer Prediction Using Data Mining Algorithms

Nitasha
Department of Computer Science and Engineering
Beant College of Engineering and Technology
Punjab Technical University, Gurdaspur
Punjab - India

ABSTRACT
Breast Cancer is the main deadest problems faced by the not only women but men too. To predict and identify the tumor at earlier stage and to cure the cancer earlier is the most important need at this time. There are different techniques and algorithms used for the predicting of the breast cancer as a tumor at earlier stage. As in this review paper, we survey different classification algorithms and clustering algorithms of data mining and machine learning used to predict the Breast Cancer. Out of which shows that classification algorithms are better than the clustering algorithms depending upon the datasets.

Keywords:- Breast Cancer, Classification, SVM, Naïve Bayes, KNN, Random Forest

I. INTRODUCTION

The human body is composited of millions of cells. When the irregular growth of cells starts it form tumor which leads to the cancer. The Breast cancer is one of the deadest diseases among men and women. According to the study of 2012-2014, 38.5% men and women have been diagnosed by the cancer [1, 2]. In today, time the main reason for the death on this globe is cancer [3, 4]. The research reported that the cancer death rate is 70% in the developing countries as compared to the other developed countries [5, 6]. The research reported that the cancer death rate is 70% in the developing countries as compared to the other developed countries [7, 8]. Among the cancer, Breast cancer is rapidly increasing following the ovaries and cervical cancers [9-11]. Breast cancer is not only found in the women, it is also found in the men. Breast cancer can be classified in two ways- Benign and Malignant. When the tumor is identify earlier and can be diagnosed at the earlier stage is known as benign condition. When the tumor starts spreading to the other parts and cannot be cured is known as malignant condition.

It is important to predict the breast cancer symptoms so that it can be treated at earlier stage before it leads to the death. In today, where the technology has reached at such height that can cure the other disease. So it is more important to predict the breast cancer earlier and treated at earlier stage. The large number of data related to the breast cancer is analyzed and then the datasets are evaluated using different technologies such as Machine Learning and Data mining techniques.

In this paper, Section I contains the introduction of Review on Breast cancer Predication using Data Mining Techniques, Section II contains the Methodology and Algorithms used in prediction of breast cancer, Section III contains the Literature survey about the Breast Cancer prediction using Data mining techniques by different authors, Section IV contains the conclusion about the different Data mining techniques used for prediction of Breast Cancer.

II. METHODOLOGY AND ALGORITHMS

1. Naïve Bayes

Navies Bayes is a statically and probabilistic classifiers, which is based on Bayes Theorem. Every feature of the attribute is independent to each other attribute. It is a classification technique which was designed to classify the high-dimensional datasets.
The probability of events are calculated as

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Where y and X are different features. y is known as class variable and X is the evidence i.e. probability of event prior evidence.

2. Decision Table

Decision Table is a classification algorithm which build tree structure data format. The datasets are divided into sub nodes or sub leaf. Each sub nodes represent the instances of the datasets. The leaf node is called as class label. This algorithm is implemented as J48 in WEKA application. Some datasets include missing values are not calculated by the algorithms but this algorithm show accurate result still having missing values.

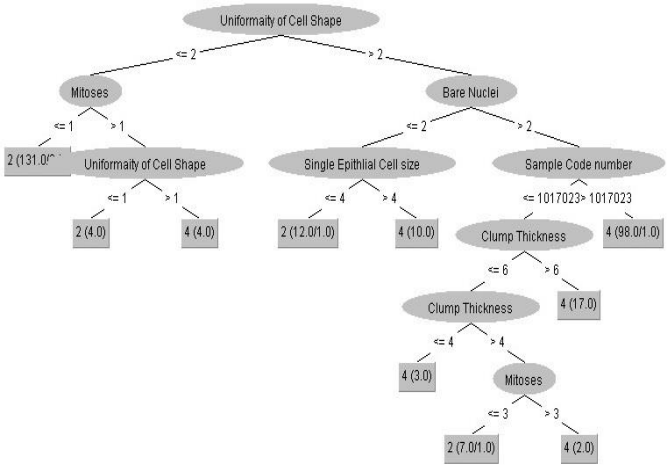

Fig 2.2.1: Diagram for the Decision Table algorithm

3. Support Vector Machine

Support Vector Machine is called as SVM. It is Supervised Machine learning algorithm used for classification and regression. In SVM algorithm the data set is plotted in the form of n-dimensional (where n is no. of instances in dataset). It differentiates different features in the hyper plane and compares them very well. It is used in handwritten digital recognition, image recognition, face detection, Bioinformatics and many more.

4. K-Nearest Neighbor

KNN is a lazy model because it does not learn anything during the training phase and learns in testing phase. It is instance-based learning. It is a non-parametric learning which memorize the resultant of classify off unseen data. It is used for classification and regression algorithms. The output of the classification are in form of 1, -1and 0. This algorithm is used for pattern recognition and intrusion detection. It takes more time to compute the result so it is less efficient among the others.

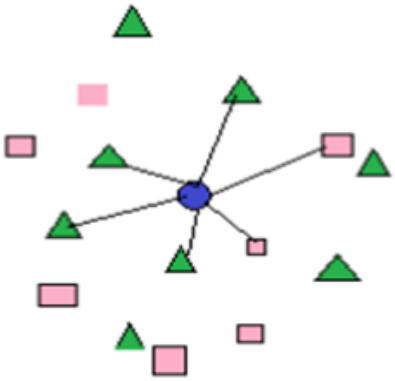

Fig 2.4.1: Diagram for the KNN algorithm

5. Random Forest

Random Forest algorithm is collections of different decision tree which assemble to build a forest know as Random Forest. In Random Forest, each node is split using the best nodes among the others random nodes. It is unaffected by the missing values and noise present in the input data. It is a recursive approach in which the instances are taken randomly from the data sets. It has better stability then the single decision tree and handles the data minorities which make it more efficient to use for the breast cancer prediction.

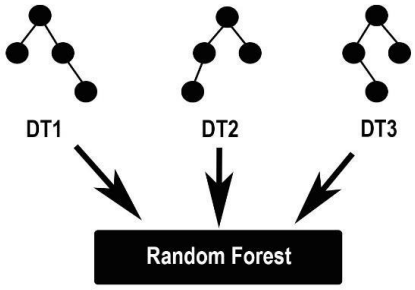

Fig 2.5.1: Diagram for the Random Forest

III. LITERATURE SURVEY

M. Navya Sri et al [2019] [12], comparative analysis has been made between Decision Tree J48 algorithm and Bayesian classification to determine the breast cancer among the women showing the resultant that J48 have75.875% accuracy and 75.17% of Bayesian using WEKA tool. The data set includes 32 attributes and 286 instances.

Akshya Yadav et al [2019][13], discussed comparative study of six Machine Learning algorithms for Breast Cancer Prediction namely as SVM, Artificial Neural Network, K-nearest Neighbor, Decision tree and Random Forest using Anaconda Python tool. They made comparison and on the performance based conclude that SVM and Random Forest has high accuracy, whereas Naives Bayes classifier have highest precision.

K. Goyal et al[2019][14], used different test option such as cross validation and percentage spilt to give better result of comparison of Adaboot, SVM, Naives Bayes, Decision Tree, J48 algorithms. The data set contains 32 attributes and using feature selection show the result. As a resultant Random Forest gave highest accuracy and Adaboot least.

D S Jacob et al [2018][15], give a survey of breast cancer prediction using data mining technology. It shows that classification algorithm perform better resultant than the clustering algorithms in predicting breast cancer. Shows that SVM and C5.0 have same accuracy.

Chintan et al [2013][16], using the data mining classifiers namely Naives Bayes, DT, KNN using different parameters to predict the cancer and as a resultant shows that Naives Bayes is more superior to other two. For the superior result, focus on the accuracy and lowest computing time taken.

Comparsion of varies algorithms used with different tool for data mining

Author	Year	Tool	Algorithms
K.Goyal et al	2019	WEKA	Adaboost, SVM, Random Forest, Naïve Bayes, Decision Tree, J48, Logistic Regression
Akshya Yadav et al	2019	Anaconda, Python	SVM, NBC,ANN,KNN,DT,RF
M.Navya Sri et al	2019	WEKA	J48, Bayesian
DS Jacob et al	2018	WEKA	C5.0, KNN, Naïve Bayes, SVM
Chitan Shah	2013	WEKA	Naïve Bayes, Decision tree, KNN

IV. CONCLUSION

In this paper, the different datamining and machine learning algorithms are studied to predict the breast cancer using different data sets and different data mining algorithms. The accuracy depends upon the data mining algorithms. Future work can be done to analyze the data set of breast cancer using datamining classification algorithms because classification algorithms show better result than the clustering algorithms. To focus on the accuracy and precision, classification algorithms can show the better performance in predicting Breast Cancer.

REFERENCES

[1] M. Balonode et al, The Stanford digital library metadata architecture. INt.J.Digital.Libr, pp.108-121(1997).

[2] C.Paper et al, Prediction Models for Estimation of survival rate and replace for Breast Cancer patients, March 2016 (2015).

[3] K.B Bruce et al, Comparing object Encodings in Theoretical Aspects of Computer Software., Springer, Berlin, New York, 1997, Vol.1281, pp.415-438.(1997)

[4] K.K et al, Machine Learning applications in cancer prognosis and predication, Computer Structure Biotechnology (2015).

[5] J V Leeumen et al, Computer Science Today. Recent Trends and developments. Lecture notes

in Computer Science, Springer, Berlin, vol. 1000, (1995).

[6] RJ. Kate et al, Stage-Specific Predictive models for breast cancer survivability. International Journal Med.Information, pp.304-311(2017).

[7] Z. Michaewicz et al, Genetic algorithms and Data Structures Evolution Programs, 3rd edition Springer Berlin (1996).

[8] K. Ahmed, Prediction of Breast Cancer risk Level with risk factor in perspective to Bangladeshi women using data mining, pp.36-41 (2013).

[9] A.Maheshwari et al, Gyneclogical Cancers: A summary of published Indian data. South Asian Cancer, vol 5(3), pp.112-120 (2017).

[10] H. Asri et al, Using Machine Learning Algorithms for Breast Cancer risk Prediction and Diagnosis. Computer science 83(Fams), pp.1064-1069,(2016).

[11] E.Gauthier et al, Breast Cancer Risk Score: A data mining Approach to improve Readability, The international conference on data mining, pp.15-21, (2011).

[12] M.Navya Sri,J.S.V.S. Hari Priyanka, D.Sailaja and M.Ramakrishna Murthy, A Comparative Study of Breast Cancer Data Set Using Different Classification Methods, Springer Nature Singapore Pte Ltd. ,pp. 175-181,(2019).

[13] Akshya et al, Comparative Study of Machine Learning Algorithms for Breast Cancer Predication- A review, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 5, Issue 2, pp.979-985 (2019).

[14] Kashish Goyal et al, Comparative Analysis of Machine Learning Algorithms for Breast Cancer Prognosis, Springer Nature Singapore Pte Ltd., pp.727-734, (2019).

[15] Dono Sara Jacob et al, A Survey on Breast Cancer Prediction Using Data Mining Techniques, IEEE Conference on Emerging Devices and Smart System, pp.256-258 (2018)

[16] Chintan shah et al, Comparsion of data mining algorithms classification for breast Cancer prediction, pp.1-4, (2013)