Survey on Data Mining Techniques for Diagnosis and Prognosis of Breast Cancer

Anupama Y.K[1], Amutha .S[2], Ramesh Babu.D.R[3]
[1] Faculty, [2]Prof., [3]Prof.

[1]Anupama Y.K.
Computer Science & Engineering
Dayananda Sagar College of Engineering
Bengaluru, India
anupamayk@gmail.com

[2]Dr.Amutha .S
Computer Science & Engineering
Dayananda Sagar College of Engineering
Bengaluru, India
amuthanandhu@gmail.com

[3]Dr.Ramesh Babu D.R.
Computer Science & Engineering
Dayananda Sagar College of Engineering
Bengaluru, India
bobrammysore@gmail.com

Abstract—Data mining is currently being used in medical systems, growing as a new interest in research community. This paper surveys the application of data mining techniques for diagnosis and prognosis of breast cancer. Each of these has different data sets and different objectives for knowledge discovery. Techniques of data mining (DM) help the medical professionals in decision making for diagnosis of breast cancer in order to avoid surgical biopsy.

Keywords- *Data Mining, Breast Cancer, Diagnosis, Prognosis*

*****

I. INTRODUCTION

The most important aspect in medical field is early and accurate diagnosis of any disease which helps to cure and increase the life expectancy chances. Cancer is one of such disease where accurate diagnosis can reduce the death rate in cancer patients. Generally tumors can be categorized as benign (non-cancerous) and malignant (cancerous). The malignant tumor develops, when cells of the breast tissue iteratively partition without control on its growth.

According to the reports of World Health Organization (WHO), breast cancer is the second most cause of women death both in the developed and the developing countries including India [1]. Breast cancer risks in India revealed that 1 among 28 women develop breast cancer during her lifetime [2]. This is higher in urban areas being 1 among 22 in lifetime compared to rural areas where this risk is relatively much lower being 1 among 60 women developing breast cancer in their lifetime. Some of the risk factors such as age, genetic risk and family history increase the likelihood of a woman developing breast cancer. With early diagnosis, 97% of women can survive for 5 years. But, most of cancer events are diagnosed in the last stage of the illness. So, the techniques for accurate and early diagnosis of breast cancer are required. Data mining is one of the techniques for accurate diagnosis of breast cancer.

II. RESEARCH ARTICLES

The automatic investigation of hazardous illness breast cancer has been considered using the machine learning algorithm [3]. The proposed approach identifies cancer occurrence during the beginning of cancer and also its reoccurrence, which has three stages. The first stage is about, enclosing the data in to number related entities by applying Farthest First clustering algorithm. Computation time took less time, due to decrease in the size of dataset. Followed by in the second stage, deviations from the normality (outliers) are detected from breast cancer dataset (BCD) using Outlier Detection Algorithm (ODA). The Final stage, J48 classification algorithm identifies whether the cancer is benign or malignant from the pre-processed data set. Wisconsin Breast Cancer Dataset (WBCD) and Wisconsin Diagnosis Breast Cancer Dataset (WDBC) show an accuracy of 99.9%, which helps doctor to diagnose the breast cancer.

Walaa Gad [4] proposed a method to improve the diagnosis of breast cancer on WDBC and Wisconsin Prognosis Breast Cancer Dataset (WPBC), in which he combined an unsupervised learning method K-means with SVM a supervised learning method. This method eliminates the inapplicable attributes using feature selection method chi-square. Method also improves the performance by speeding up and also eliminates the curse dimensionality.

Jahanvi Joshi et al. [5] state that k-means clustering algorithm and Farthest First algorithm are useful in early diagnosis of the breast cancer patients. They found that EM (Expectation Maximization) technique is not efficient in diagnosing. The future purpose extended as usage of Orange, Tavera and Rapid Miner tools.

Women with breast cancer can be survived for more than five years, if it is diagnosed in early stage of cancer. This leads to the improving survival rate results up to 97%. The method for this has been proposed by Jamini Malaji et al. [1]. FP-Growth algorithm is used to find the patterns of benign and malignant cells. To predict the possibility of cancer in terms of age the decision tree algorithm is used. Classification accuracy of proposed method is 94% on Wisconsin dataset.

Table 1 shows breast cancer diagnosis accuracy of different data mining techniques.

Table.1. Diagnosis accuracy of BCD

Index of Citation	Year	Methodology (classifiers)	Tool/Software	Dataset	Accuracy
Rakhi Malpani et al.	2011	Association rule mining	WEKA	Breast cancer gene expression (breast cancer patient profile data)	----
D.Lavanya et al.	2012	CART Decision Tree	WEKA	WDBC	92.97%
Gouda I. Salama et al.	2012	SMO+J48+MLP +IBK	WEKA	WDBC,WPBC	77.32%
K.R.Lakshmi et al.	2013	PLS-DA	Tanagra	WPBC,WDBC	96.66%
G. Ravi Kumar et al.	2013	SVM(SMO)	WEKA	WPBC	94.5%
Tintu P B and Paulin R	2013	Fuzzy C-means	MATLAB	WPBC	97.13%
Alaa M. Elsayad and H.A. Elsalamony	2013	SVM	Clementine data mining workbench	WDBC	99.96%
Hemant Palivela	2013	J48	WEKA	City Hospital	75.52%
Al- Imran Ahmed and Md Mahmudul Hasan	2014	J48	WEKA	UCI (http://repository.seasr.org/Datasets/UCI/arff/breast-cancer.arff).)	65.21%
Zehra Karapinar Senturk and Resul Kara	2014	SVM	Rapid Miner 5.0	Wisconsin Breast Cancer	96%
Jahanvi Joshi et al.	2014	K-Means Clustering	WEKA	UCI (http://repository.seasr.org/Datasets/UCI/arff/breast-cancer.arff)	83%
Ronak Sumbaly et al.	2014	J48	WEKA	SEER database	94.56%
Vikas Chaurasia and Saurabh Pal2	2014	SMO	WEKA	Wisconsin breast cancer data set	97%
Jaimini Majali et al.	2015	FP	---	Wisconsin	----
Bojana R et al.	2016	SVM	---	Kragujevac	89%
Animesh Hazra et al.	2016	SVM	---	WDBC	97.39%
Mohammed Abdullah Hassan Al-Hagery	2016	Bayesnet	WEKA	WBC	97.37%
R Delshi Howsalya Devi et al.	2016	Outlier Detection	WEKA	WDBC,WPBC	99.6%
Walaa Gad et al.	2016	SVM-Kmeans	----	WDBC,WPBC	99.8%

III. DATA MINING TECHNIQUE

Data mining technique is defined as the process of discovering interesting patterns and knowledge from the large amounts of data. Technique refers to analyzing data from different viewpoints and abstracting it to get the necessary information. DM technique provides an important means for extracting valuable medical rules hidden in medical data and acts as an important role in clinical diagnosis.

The Fig. 1 shows the typical architecture of data mining. The architecture work flow includes four phases:

- First phase: Cleaning and integration technique applied on the data available from the database to filter the missing and uncertain data. The output of this process is preprocessed data.
- Second phase: Relevant data mining functionalities/tasks can be carried out on the preprocessed data. This process yields patterns.
- Third phase: Patterns evaluated depending on the constraints to extract valuable knowledge.
- Fourth phase: Using knowledge in the decision making of real time applications example medicines, health care, etc.

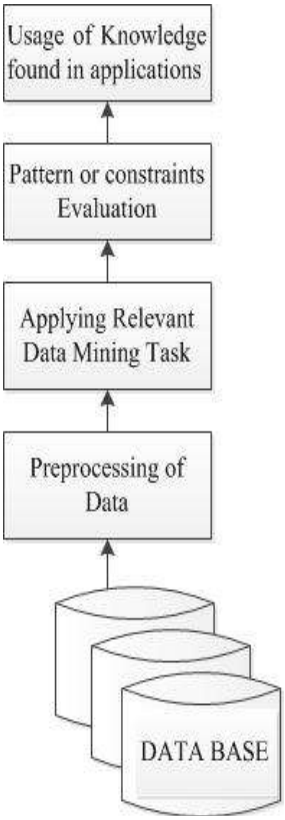

Fig 1. Architecture of Data Mining

A. DM TASKS

Tasks of data mining are helpful in identifying the patterns in the applications. Patterns are the frequency and style of data occurrence. DM tasks can be categorized as predictive and descriptive tasks as shown in the Fig. 2.

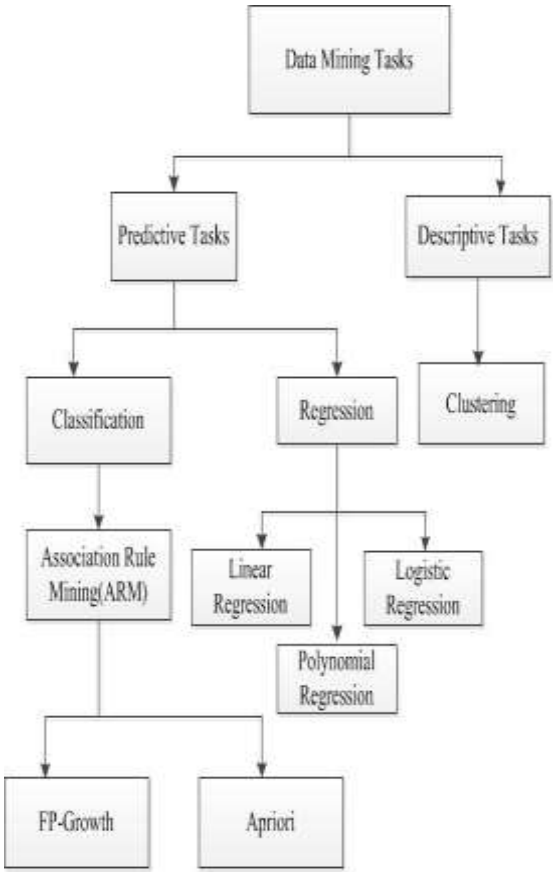

Fig 2. Data Mining Tasks

Predictive Tasks: Predictive mining task is carried out for the deduction of the data from the existing data in order to forecast a model of prediction depending on the constraints. Predictive tasks classified as Classification and Regression. Classifiers are used for classification of data available depending on the sequence of scenario. Techniques of classification traditionally applied for nominal data (categorical/enumerations) and unordered data. Regression technique (RT) can be used for classification, but on quantitative-numerical data and ordered whole numbers. RT reveals the significance state of being connected between dependent and independent variables and also it helps to uncover the influence of more than two independent variables on a dependent variable. Different types of regression are linear regression, logistic regression and polynomial regression.

Linear Regression (LR) is represented by mathematical derivation of Equation 1

$$B=p+q*A+r \quad (1)$$

Where p is an intercept, q is slope of the line and r is error term. This represents a predicted value of target variable based on given predictor. LR shows the relationship between dependent B and one/more independent variable A using a best fit straight line (also known as regression line).

Logistic Regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of B ranges from 0 to 1 and it can represent by following Equation 2.

$$B=p0+p1A+r \quad (2)$$

Polynomial Regression (PR) derives a linear regression equation if the power of independent variable is more than 1. The Equation 3 represents a polynomial equation.

$$B=p+q*A\hat{}2 \quad (3)$$

Polynomial regression technique represents the best fit line is not a straight line. Rather it is a curve that fits into the data points.

Descriptive Tasks: This mining task is carried out to derive a layout of knowledge and analyze its resultant patterns and relationships in huge volume of available data. Clustering task groups the similar data and the remaining data (dissimilar data) used for outlier analysis. For both nominal and quantitative numerical dataset, clustering tasks can be applied.

B. ARM(ASSOCIATION RULE MINING)

The aim of ARM is to identify the useful rules from the large amounts of data. Association rule mining has following logical process and appeal.

- Logical process: Interesting rules are determined in terms of support and confidence.
  1) Support: It reflects the usefulness of discovered rules.
  2) Confidence: It reflects certainty of discovered rules.
- Appeal: Association rules are considered as interesting if they satisfy both minimum support threshold and minimum confidence threshold. These thresholds set by users or domain experts. Additional analysis can be performed to discover interesting statistical correlations between associated items.

C. APRIORI(FREQUENT PATTERN MINING)

The objective of Apriori is to find the frequent-itemsets. Apriori has following logical process and appeal.

- Logical Process: It employs an iterative approach known as level-wise search, where k-item sets are used to explore (k+1) item sets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted by L_1. Next, L_1 is used to find L_2, the set of frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database.
- Appeal: The Apriori procedure performs two kinds of actions, namely, join and prune, as described before. In the join component, L_{k-1} is joined withL_{k-1} to generate potential candidates. The prune component employs the Apriori property (all nonempty subsets of a frequent item set must also be frequent) to remove candidates that have a subset that is not frequent. Finally, all the candidates satisfying the minimum support from the set of frequent item sets is considered.

D. FP-GROWTH(FREQUENT PATTERN MINING)

The goal of FP-Growth algorithm is to identify the frequent item sets without candidate generation from the database. FP-growth has following logical process and appeal.

- Logical Process: The first scan of the database derives the set of frequent items (1-item sets) and their support counts (frequencies).Then according to the predefined minimum_support count the set is sorted in descending order. This is extended to the construction of FP-tree by using the extended prefix-structure of mining the complete set of frequent patterns by pattern fragment growth.
- Appeal: The FP-tree is mined by constructing conditional pattern base (a sub-database, which consists of the set of prefix paths in the FP-tree co-occurring with a suffix pattern), then construct its (conditional) FP-tree, and perform mining recursively on the tree. The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree. A study of the FP-growth method performance shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm.

E. CLUSTERING

Clustering technique is used to identify the object belong to the cluster or not. If not, then it is identified as an outlier. Technique has following logical process and appeal.

- Logical Process: Clustering techniques consider data tuples as objects. They partition the objects into groups or clusters, so that objects within a cluster are similar to one another and dissimilar to objects in other clusters.
- Appeal: The quality of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster.

IV. CONCLUSION

From the survey it is found that there is a wide scope to carry out research work in the field of breast cancer diagnosis using data mining techniques. Design and implementation of novel algorithms to improve the accuracy of breast cancer diagnosis is essential.

REFERENCES

[1] Jaimini Majali, Rishikesh Niranjan, Vinamra Phatak, OmkarTadakhe, “Data Mining Techniques For Diagnosis And Prognosis Of Cancer”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015, pp.613-616

[2] Vikas Chaurasia, Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE), Vol. 2, Issue 1, January 2014, 2456-2465

[3] R Delshi Howsalya Devi, Dr. M Indra Devi, "Outlier Detection Algorithm combined with Decision Tree Classifier for early Diagnosis of Breast Cancer", (IJAET) International Journal of Advanced Engineering Technology E-ISSN 0976-3945, Vol. VII,Issue II,April-June,2016, 93-98

[4] Walaa Gad , "SVM-Kmeans: Support Vector Machine based on Kmeans Clustering for Breast Cancer Diagnosis" International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 05 – Issue 02, March 2016,252-257.

[5] Jahanvi Joshi, Rinal Doshi and Jigar Patel, "Diagnosis of Breast Cancer using Clustering Data Mining Approach" International Journal of Computer Applications (0975 – 8887) Volume 101– No.10, September 2014,13-17.

[6] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", (IJCA) International Journal of Computer Applications, Volume 145 – No.2, July 2016 , 0975 – 8887

[7] Mohammed Abdullah Hassan Al-Hagery, "Classifiers' Accuracy Based on Breast Cancer Medical Data and Data Mining Techniques",(IJBR) International Journal of Advanced Biotechnology and Research ,ISSN 0976-2612, Online ISSN 2278–599X,Vol-7, Issue-2, 2016, pp760-772.

[8] R Delshi Howsalya Devi, Dr. M Indra Devi, "Outlier Detection Algorithm combined with Decision Tree Classifier for early Diagnosis of Breast Cancer", (IJAET) International Journal of Advanced Engineering Technology E-ISSN 0976-3945, Vol. VII,Issue II,April-June,2016, 93-98

[9] Bojana R. Andjelkovic Cirkovic, Aleksandar M. Cvetkovic, Srdjan M. Ninkovic, Nenad D. Filipovic, "Prediction models for estimation of survival rate and relapse for breast cancer patients", (IEEE), vol. 00, no. , pp. 1-6, 2015. doi:10.1109/BIBE.2015.7367658

[10] Meera Narvekara, Shafaque Fatma Syed,"An optimized algorithm for association rule mining using FP tree", (ELSEVEIR-ICACTA) International Conference on Advanced Computing Technologies and Applications , Procedia Computer Science 45 , 2015 ,101 – 110

[11] Zehra Karapinar Senturk , Resul Kara, "Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven Different Algorithms", (CSEIJ)Computer Science & Engineering: An International Journal, Vol. 4, No. 1, February 2014,pp.35-46.

[12] K.R.Lakshmi, M.Veera Krishna, S.Prem Kumar," Performance comparison of data mining techniques for prediction and diagnosis of breast cancer",(AJCSIT) Asian Journal of Computer Science And Information Technology 3 : 5 ,2013, pp. 81 – 87

[13] Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of cancer", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.