Empirical Study on Data Mining Techniques For Breast Cancer Diagnosis

Ms. Sindhuja C.
M. Phil Research Scholar,
Department of Computer Science
Auxilium College (Autonomous), Vellore-6.

Ms. Kavitha S.
Assistant Professor,
Department of Computer Science
Auxilium College (Autonomous), Vellore-6.

Abstract

Data mining is a process used by companies to turn raw data into useful information. It is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends. This paper studies various data mining techniques for breast cancer Diagnosis. This paper reviews about the parameters and techniques used to diagnosis breast cancer in begnin or malignant stage. Various data mining techniques were used in diagnosis like k-nearest, Bayes classifier, fuzzy-c-means, neural network, thresholding etc. has been explored in this paper.

Keywords: Breast cancer diagnosis, Data mining Techniques.

I. Introduction

Breast cancer is a cancer that forms in the cells of the breasts. After skin cancer, breast cancer is the most common cancer diagnosed in women in the United States. Breast cancer can occur in both men and women, but it's far more common in women. Substantial support for breast cancer awareness and research funding has helped created advances in the diagnosis and treatment of breast cancer. Breast cancer survival rates have increased, and the number of deaths associated with this disease is steadily declining, largely due to factors such as earlier detection, a new personalized approach to treatment and a better understanding of the disease. Breast cancers can start from different parts of the breast. Most breast cancers begin in the ducts that carry milk to the nipple (ductal cancers). Some start in the glands that make breast milk (lobular cancers). There are also other types of breast cancer that are less common.

A small number of cancers start in other tissues in the breast. These cancers are called sarcomas and lymphomas and are not really thought of as breast cancers. Although many types of breast cancer can cause a lump in the breast, not all do. Many breast cancers are found on screening mammograms which can detect cancers at an earlier stage, often before they can be felt, and before symptoms develop. There are other symptoms of breast cancer you should watch for and report to a health care provider. It's also important to understand that most breast lumps are benign and not cancer (malignant). Non-cancerous breast tumours are abnormal growths, but they do not spread outside of the breast and they are not life threatening. But some benign breast lumps can increase a woman's risk of getting breast cancer.

A tumor is a mass of abnormal tissue. There are two types of breast cancer tumors: those that are non-cancerous, or 'benign', and those that are cancerous, which are 'malignant'. When a tumor is diagnosed as benign, doctors will usually leave it alone rather than remove it. Even though these tumors are not generally aggressive toward surrounding tissue, occasionally they may continue to grow, pressing on organs and causing pain or other problems. In these situations, the tumor is removed, allowing pain or

complications to subside. Malignant tumors are cancerous and aggressive because they invade and damage surrounding tissue. When a tumor is suspected to be malignant, the doctor will perform a biopsy to determine the severity or aggressiveness of the tumor.

Nowadays in all fields of sciences including genetics, education, earth science, agriculture and medicine the amount of data is increasing dramatically. Analyzing these huge amount of data to extract the novel and usable information or knowledge is very complicated and time consuming task. Data mining techniques are useful for this matter. Data mining and knowledge discovery in databases (KDD) are extracting novel, understandable and useful information, knowledge or patterns from huge amount of available data . In the other words, data mining has capabilities for analysing the large datasets, finding unexpected or hidden relationships between various attributes and summarizing the extracted information more understandable and useful to data users or owners. In the traditional model for transforming data to knowledge, some manual analysis and interpretation are executed. For example, in .3medical centres, generally doctors or specialists manually analyse current trends, disease and health-care data, then make a report and use this report for decision making or planning for medical diagnosis, treatments and etc. The problem of this type of data analysis is that, this form of manual data analysis is slow, expensive, time consuming, and highly subjective.

Recently the data mining techniques like thresholding, ANN, neural networks, genetic algorithms, Bayes classifier provide the accurate results. Many researchers has made a smoothing view by collecting the previous data offering a great promise to uncover patterns hidden in the data that can help the clinicians in decision making. The accuracy for the diagnosis analysis of various applied data mining classification techniques is highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy. The best model can be obtained after building several different types of models, or by trying different technologies and algorithms.

The paper is planned as follows. Section II reviews the data mining techniques. Section III discusses the related work. Section IV has made a comparison table of existing data mining techniques, methods and algorithms. Section V contains the gaps in the literature work. Section VI contains the conclusion and future work.

II. Data Mining Techniques

Data Mining is the process of turning raw data into appropriate and meaningful information. It involves exploring and analysing large blocks of information to glean meaningful patterns and trends. Various researchers have studied and work on data mining techniques to evaluate the diagnosis of breast cancer.

KNN

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

Naïve Bayes Classifier

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive. We can derive probability models by using Bayes' theorem (credited

to Thomas Bayes). Depending on the nature of the probability model, you can train the Naive Bayes algorithm in a supervised learning setting.

Data mining in Infosphere Warehouse is based on the maximum likelihood for parameter estimation for Naive Bayes models. The generated Naive Bayes model conforms to the Predictive Model Markup Language (PMML) standard. A Naive Bayes model consists of a large cube that includes the following dimensions:

- Input field name.
- Input field value for discrete fields, or input field value range for continuous fields.
- Continuous fields are divided into discrete bins by the Naive Bayes algorithm.
- Target field value.

Support Vector Machine

In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).The given labeled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

Decision Tree

Decision tree is one of the predictive modelling technique used in data mining. It aids to divide the larger dataset into smaller dataset indicating a parent-child relationship. Each internal node defined as inner node is labelled with an input feature. The inner nodes which exhibit many types of attribute test, bifurcations exhibit the test outcomes and leaf nodes particularly exhibit the category of a specific type. Decision tree can handle both numerical and categorical data. It is well suited with large datasets. Higher accuracy in decision tree classification technique depicts that the technique can simulate. It is able to optimize variety of input data such as nominal, numeric and textual.

Fuzzy-C-Means

The FCM is also known as fuzzy *c*-means nebulous because it uses fuzzy logic so that each instance is not associated with only one cluster, but has a certain degree of membership for each of the existing centroids. Clustering is a field of research belonging to both data analysis and machine learning major domains. The Fuzzy C-Means algorithm (FCM) is used in the areas like computational geometry, data compression and vector quantization, pattern recognition and pattern classification.

III. Related Work

Ch. Shravya et.al [1] proposed the logistic regression, support vector machine, K-nearest neighbour. The comparison was made on the basis of accuracy, precession, sensitivity, specificity, false positive rate the efficiency is calculated. Logistic regression relies simple equation which shows the linear relation between the independent variables. KNN is an non- parametric algorithm relies on the nearest training points rather than parameters. Support Vector Machine is well in doing pattern recognition problems and it is used as an training algorithm. The datasets which are used for experiment contain 32 attributes where a lot of decreasing in multi- dimensional data to few dimensions. This analysis signify the integration of multi- dimensional data along with different classification. The SVM is best for predictive analysis with an accuracy of 92.7%.

Meriem Amrane et.al [2] proposed K-means classifier and Naïve Bayesian classifier. This paper states that 8% of women are diagnosis with breast cancer. BC is characterized by the mutation of genes, constant pain, changes in the size, color (redness), skin texture of breasts. The main aim is to find out whether the cancer is in malignant or begnin stage. The performance is calculated based on the accuracy, training process and testing process. The result of KNN(97%) has higher than the Bayesian(96.19%) .

Sangeetha R et.al[3] proposed Otsu's thresholding, K-nearest neighbour, Bayes classifier. Image processing on breast cancer is not new, but the lack of the detection in early stage is still a complicated. The accuracy won't be 100% and doesn't provide the end to end solution. In this they are provided with techniques of digital image processing and detect in early stage of cancer, they completely provide the accuracy. Otsu's thresholding algorithm will segment the microcalcification the image with high accuracy. The microcalcification are grouped into clusters using K-nearest neighbour clustering algorithm. The extracted features are classified whether in begnin or malignant using Bayes classifier. The methodology identified the microcalcification clusters at an very early stage with very high accuracy and leads to true positive and negative values.

Pavel Kral et.al [4] proposed the novel method for breast cancer detection from mammographic images based on local binary patterns(LBP) with the features of classifier and thresholding. In experimental setup for detecting of cancer they had used the SVM classifier with polynomial kernel. The classification algorithm used is the thresholding for detection of mammographic. For optimal LBP cell size possibility they had used binary classifier and to find out which breast size has more cancerous. The proposed method was calculated with MIAS and DDMS database. The best threshold values is set to 0.5 whereas the accuracy of f-measures is 84% and resulted that LBP is more efficient and effective.

Kanchan Lata Kashyap et.al[5] proposed fuzzy-c-means with thresholding technique. In this paper they discussed about the approach for automatic detection of abnormalities in the mammograms. Region on Interest(ROI) is used for abnormality detection in image processing technique. The data noise removal was done with median filtering. The image processing suspicious ROI segmented using fuzzy-c-means along with thresholding. Sharpen images is used further to segment the suspicious mass region. Segmentation is performed to partition the image into homogenous region. Support vector machine is used to identify the segment region whether it is in normal or abnormal region. They used the mammographic images analysis society(MIAS) data set contains 322 mammogram images. They conclude that moment based is better than the region based and tamura based region.

Nithya R et.al [6] proposed ensembles classifier, SVM,SMO,NB tree. The detection and the characterization of breast cancer had been done with the assists of Computer Aided Design(CAD). They stated that rather than using a single classifier they had proven that ensembles are better. Ensembles classifier is the combination of the serval classifier known as multiple classifier, it improves the classification performance. It contains bagging, multi-boost and random subspace. It improves the prediction accuracy. It helps in determining the patient whether they have breast cancer or not. They done the experiment using Wisconsin breast cancer dataset(WBCD). The parameters used to calculate sensitivity, specificity and classification accuracy. Support vector machine and Sequential minimal optimisation is effective for boundary classification. It is based on decision boundary between different classes. NB tree is more accurate than Naïve Bayes or decision tree classifier. They had made a accuracy of base classifier and ensembles for breast cancer classification and tested on WBCD. The performance was seen developed and later used to check the accuracy. They conclude by stating that SMO has gained more accuracy.

IV. Comparison Table

Ref No	Author	Paper	Year	Techniques	Observation	Advantages	Limitations
1	Ch. Shravya, K Pravalika, Shaik subhani	Prediction of breast cancer using supervised machine learning Techniques	2019	Bayesian Classifier, K-nearest neighbour	To find the systematic and objective prognostic.	It provides high accuracy	In case of larger dataset Naïve Bayesian can't provide more accuracy
2	Merien Amrane, Ikram Gagaoua	Breast cancer classification using machine learning	2018	SVM Logistic regression	For classification of begnin or malignant tumor used to learn from the past and predict new inputs.	Predicted efficient performance and good accuracy	Logistic regression is completely mathematical and difficult to understand.
3	Sangeetha Dr. Srikanta Murthy K	Novel approach for early detection of breast cancer using image processing techniques	2017	Otsu's Thresholding, Bayes classifier.	To detect cancer in early stage with true positive and negative values.	It provide the end to end solution.	The performance level are not stated and suitable only for image segmentation
4	Pavel Kral, Ladislav lenc	LBP Feature for breast cancer detection	2016	SVM, Thresholding	The compared the mammographic images using MIAS and DDSM datasets	Accuracy of thresholding is high and can help the radiologists to work.	It is not possible to compare the reported results among themselves.
5	Ranchan Lata Kashyap, Manish Kumar, Bajpai, Pritee Khanna	Breast Cancer detection in digital mammogram	2015	Fuzzy-c-means, Region of Interest (ROI)	Automatic detection of abnormalities in mammograms	FCM has more accuracy and effective	They will provide the temporal information not the exact one's
6	R Nithya and B shanthi	Data mining technique for diagnosis of breast cancer	2014	NB tree, SMO, Ensemble	Detection and characterization using CAD	Improves the performance.	Necessary to solve QP-problem scaling

V. Gaps in Literature

The majority of the pre -existing techniques has certain restrictions and problems, because it has neglected many of the points some of them are:

- The use of selected techniques can be done to enhance the accuracy rate further for prognosis and diagnosis of breast cancer.
- The majority of the existing techniques are limited to most of the substantial features of breast cancer using image segmentation.

VI. Conclusion

This paper presents an evaluation for diagnosis of breast cancer by applying numerous data mining techniques and methods. Many existing evaluation methods are studied. Various algorithms have been reviewed for diagnosing the breast cancer and hence made a comparison in this KNN and SMV have high frequency. As a result instead of using the image segmentation for diagnosis of breast cancer we can also use the data.

Reference

[1] Ch. Shravya, K Pravalika, Shaik subhani "Prediction of breast cancer using supervised machine learning Techniques"(2019), ISSN: 2278-3075, Volume-8 Issue-6.

[2] Merien Amrane, Ikram Gagaoua "Breast cancer classification using machine learning", LRDSI Laboratory, University of Blida 1, Blida, Algeria(2018), 78-1-5386-5135-3/18.

[3] Sangeetha R, Dr. Srikanta Murthy K "Novel approach for early detection of breast cancer using image processing techniques", PES Institute of Technology, South Campus Bangalore, India. 978-1-5090-4715-4/17(ICISC-2017)

[4] Pavel Kral, Ladislav lenc "LBP Feature for breast cancer detection", University of West Bohemia Plzeˇn, Czech Republic, 978-4673-9961-6(ICIP 2016)

[5] Kanchan Lata Kashyap, Manish Kumar Bajpai, Pritee Khanna "Breast Cancer Detection in Digital Mammograms", Computer Science & Engineering, Indian Institute of Information Technology, Design & Manufacturing (2015) 978-1-4799-8633-0.

[6] R. Nithya and B. Santhi " A Data Mining Techniques for Diagnosis of Breast Cancer Disease", School of Computing, SASTRA University, India(2014) 18-23, 2014 ISSN 1818-4952.

[7] AlirezaOsarech, BitaShadgar,"A Computer Aided Diagnosis System for Breast Cancer",International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011

[8] MandeepRana, PoojaChandorkar, AlishibaDsouza, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.

[9] Haifeng Wang and Sang Won Yoon – Breast Cancer Prediction using Data Mining Method, IEEE Conference paper

[10] D.Dubey ,S.Kharya, S.Soni and –"Predictive Machine Learning techniques for Breast Cancer Detection", International Journal of Computer Science and Information Technologies,Vol.4(6),2013,1023-1028.