ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

A Survey on Breast Cancer Diagnosis Using Data Mining Technique

Vivek Barot[1], Prof.Niku Brahmbhatt[2]

PG Scholar, Department of IT, G.H.Patel College of Engineering and Technology, V.V.Nagar, Anand, India[1]

Assistant Professor, Department of Computer, JG College of Computer Application, Ahmedabad, India [2]

ABSTRACT: Data mining is the process to obtain knowledge from a data set and transfigure it into an understandable structure. Earlier there are many data mining techniques have been proposed to design correct classification systems for various medical issues. From that classification is especially useful in the area of medical diagnosis for decision making. The breast cancer is a serious ailment found among females all over the world. Breast cancer is being extremely injurious to all females, may be a reason for loss of breasts or even cost their life. This paper reviews the results of different classification techniques such as Support vector machine, naive bayes, Decision tree, neural networks to diagnosis breast cancer.

KEYWORDS: Data mining, classification, breast cancer.

I. INTRODUCTION

Many medical organizations generate and collect a large quantity of data. The medical field is considered as one of the leading areas for applying data mining to identify some significant properties of data[4]. Data mining tasks can be divided into two categories: Descriptive and predictive. Descriptive means to discover interesting patterns in the data and predictive means to predict the behavior of the model from available data. Some common data mining tasks are Classification, Regression, Clustering, Rule generation, Sequence analysis[9]. The aim of classification is to accurately predict the target class label of an object for which the class label is unknown. There are two steps in classification first is model construction in which a set of predetermined classes from database generates training set and second is model usage where training set is used to predict label of unknown objects.

The second leading cause of death among women is breast cancer. Breast cancer is one of the most common cancers among Egyptian women; as it represents 18.3 % of the total general of cancer cases in Egypt and a percentage of 37.3 % of breast cancer is considered a treatable disease[2]. The age of breast cancer affection in Egypt and Arab countries is prior ten years compared to foreign countries as the disease targets women at the age of 30 in Arab countries, while affecting women above 45 years in European countries.

The remaining part of this paper is organized as follows: Section II Literature Survey about the recent contribution made in breast cancer diagnosis, Section III is about Performance analysis of different classifiers, Section IV is about Conclusion and future scope.

II. LITERATURE SURVEY

There are many approaches have been used to diagnosis breast cancer using data mining techniques. Some of this are presented here.

Salama, Gouda I., M. B. Abdelhalim, and Magdy Abd-elghany Zeid [2] compare different classifiers decision tree, Multi-LayerPerception, Naive Bayes, Sequential Minimal Optimization, and Instance-Based for K-Nearest neighbor on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy and confusion matrix

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

based on 10-fold cross validation method. And also combine classifiers to get better accuracy, the experimental results show that in the classification using fusion of MLP and J48 with the PCA is superior to the other classifiers using WBC data set.

Ghosh, Soumadip, Sujoy Mondal, and Bhaskar Ghosh [3] applied different classification techniques namely, MLP using Backpropagation Neural Network and Support Vector Machine on Breast Cancer Wisconsin dataset from the UCI machine language repository for detection of breast cancer. The author concluded that SVM classifier has the potential to significantly improve the conventional classification methods for use in medical or in general, Bioinformatics field.

Elouedi, Hind, et al.[4] proposed a hybrid diagnosis approach of breast cancer based on decision trees and clustering. First they have done Extraction of the malignant instances and apply K-means algorithm to split the malignant instances. After that they apply decision tree algorithm (C4.5) to every cluster and compare the different accuracies calculated in each case. Combined results of benign and malignant are applied to the C4.5 algorithm to find out the results of classification based on the confusion matrix and the global and detailed accuracy values. They have gotten better results than those obtained with the original dataset.

Mittal, Dishant, Dev Gaurav, and Sanjiban Sekhar Roy [5] proposed an effectively hybridized classifier which is made by combining an unsupervised artificial neural network (ANN) method named self-organizing maps (SOM) with a supervised classifier called stochastic gradient descent (SGD) for breast cancer diagnosis. Also they compare there results with three supervised machine learning techniques decision tree (DTs), random forests (RF) and support vector machine (SVM). The hybrid model builds up by integrating stochastic gradient descent with self-organizing maps gave an accuracy of 99.521% over training set and 99.686% over the testing set.

Phetlasy, Sornxayya, et al.[8] proposed a hybrid method for data classification with two different classifier algorithms. The first classifier responds to reduce FN, and the second classifier is in charge of reducing FP. they implement their experiment with five popular algorithms, two algorithms for one combination so that they obtain 20 cases in total. The five algorithms are Sequential Minimal Optimization (SMO) for SVM, Naïve Bayes (NB), decision tree J48, Instance-based learning IBK algorithm for K-Nearest Neighbor, and Multilayer Perceptron (MLP) for Neural Network. The author obtained the best result of 99.13% accuracy.

Lavanya D., and K. Usha Rani [9] studied a hybrid approach: CART decision tree classifier with feature selection and boosting ensemble method has been considered to evaluate the performance of classifier. They applied CART algorithm, CART with Feature Selection Method and CART with Feature selection and Boosting on different Breast cancer data sets and compared the Accuracy.

Sarvestani, A. Soltani, et al.[10], applied self-organizing map(SOM), radial basis function network (RBF), general regression neural network (GRNN) and probabilistic neural network (PNN) are tested on the Wisconsin breast cancer data (WBCD) and on the Shiraz Namazi Hospital breast cancer data (NHBCD). And concluded RBF and PNN were proved as the best classifiers in the training set. However the PNN gives the best classification accuracy when the test set is considered.

Shah, Chintan, and Anjali G. Jivani [11] conducted the comparison between three algorithms namely Compares Decision tree, Bayesian Network and K-Nearest Neighbor with help of WEKA (The Waikato Environment for Knowledge Analysis), which is an open source software. The author concluded that Naïve Bayes is a superior algorithm compared to the two others because it takes lowest time i.e. 0.02 seconds and at the same time is providing highest accuracy.

III. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS

Different data mining techniques have been used to help healthcare professionals in the diagnosis of breast cancer. Table 1 shows a various data mining techniques used in the diagnosis of breast cancer over different breast cancer datasets.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

TABLE I
DATA MINING TECHNIQUES USED ON DIFFERENT BREAST CANCER DATASETS

Author	Year	Technique	Accuracy
Salama Gouda I., et al.	2012	Naïve Bayes	95.9943%
		MLP	95.279%
		J48	95.1359%
		SMO	96.9957%
		IBK	94.5637%
Ghosh Soumadip, et al.	2014	SVM	96.85%
		MLP BPN	95.71%
Elouedi Hind, et al.	2014	C4.5	95.14%
Nematzadeh Zahra, et al.	2015	Svm-linear	98%
		Svm-mlp	97.11%
		Svm-rbf	97.14%
		Neural Network	98.09%
		Decision Tree	95.36%
		Naïve Bayes	97.23%
Lavanya D, et al.	2012	CART	94.84%
		CART With FS	96.99%
		CART With FS and Boosting	95.56%
Sarvestani A. Soltani, et al.	2010	MLP	90%
		PNN	100%

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

IV. CONCLUSION AND FUTURE SCOPE

In this survey paper, we have studied different data mining techniques for diagnosis breast cancer. From this survey, we got the information about how to apply various data mining technique to diagnosis of breast cancer. In future an Intelligent system can be developed by using SVM and Neural Network which gives better accuracy.

ACKNOWLEDGMENT

Vivek Barot remains thankful to Prof. Niku Brahmbhatt(Assistant professor Niku Brahmbhatt,JG College of Computer Application) for their useful discussions & suggestions during the preparation of this technical paper.

REFERENCES

[1] Lakshmi, B. N., and G. H. Raghunandhan. "A conceptual overview of data mining." Innovations in Emerging Technology (NCOIET), 2011 National Conference on. IEEE, 2011.
[2] Salama Gouda I., M. B. Abdelhalim, and Magdy Abd-elghany Zeid. "Experimental comparison of classifiers for breast cancer diagnosis." Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on. IEEE, 2012.
[3] Ghosh Soumadip, Sujoy Mondal, and Bhaskar Ghosh. "A comparative study of breast cancer detection based on SVM and MLP BPN classifier." Automation, Control, Energy and Systems (ACES), 2014 First International Conference on. IEEE, 2014.
[4] Elouedi Hind, et al. "A hybrid approach based on decision trees and clustering for breast cancer classification." Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of. IEEE, 2014.
[5] Mittal Dishant, Dev Gaurav, and Sanjiban Sekhar Roy. "An effective hybridized classifier for breast cancer diagnosis." 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2015.
[6] Nematzadeh Zahra, Roliana Ibrahim, and Ali Selamat. "Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques." Control Conference (ASCC), 2015 10th Asian. IEEE, 2015.
[7] Kesavaraj, G., and S. Sukumaran. "A study on classification techniques in data mining." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.
[8] Phetlasy, Sornxayya, et al. "Sequential Combination of Two Classifier Algorithms for Binary Classification to Improve the Accuracy." 2015 Third International Symposium on Computing and Networking (CANDAR). IEEE, 2015.
[9] Lavanya D., and K. Usha Rani. "Ensemble decision making system for breast cancer data." International Journal of Computer Applications 51.17 (2012).
[10] Sarvestani A. Soltani, et al. "Predicting Breast Cancer Survivability using data mining techniques." Software Technology and Engineering (ICSTE), 2010 2nd International Conference on. Vol. 2. IEEE, 2010.
[11] Shah Chintan, and Anjali G. Jivani. "Comparison of data mining classification algorithms for breast cancer prediction." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.