DATA MINNING PROJECT
BREAST CANCER CLASSIFICATION USING DATA MINNING APPROACH

18MCA0109
SURAJ SAHA
suraj.saha2018@vitstudent.ac.in

18MCA0097
RIYA SHAJI
riya.shaji2018@vitstudent.ac.in

18MCA0088
ABHIJEET GIRI
abhijeet.giri2018@vitstudent.ac.in

ABSTRACT

Breast cancer poses serious threat to the lives of people and it is the second leading cause of death in women today and the most common cancer in women in developing countries in Nigeria where there are no services in place to aid the early detection of breast cancer in Nigerian women. Countries such as Belgium, Luxemborg and Nethelands tops the list of having the most cases of breast cancer. It is and was alaways wise to start cure before detection than after.Thus a number of studies have been undertaken in order to understand the prediction of breast cancer and types of cancer the person is having using data mining techniques. This paper studies about breast cancer prediction based on data mining methods to discover an effective way to predict breast cancer. The objective of this paper is to compare and identify an accurate model to predict the incidence of breast cancer based on various patients' clinical records. We will be applying various data mining models this paper and perform comparision of all those method to find out a optimally method for classification of cancer. Furthermore, feature space is highly discussed in this paper due to its high influence on the efficiency and effectiveness of the learning process. To test the influence of feature space reduction, a hybrid between principal component analysis (PCA) and related data mining models is proposed, which applies a principle component analysis method to reduce the feature space. The results performed by this analysis demonstrate a comprehensive trade-off between these strategies and also provides a detailed evaluation on the models. It is expected that in real application, physicians and patients can benefit from the feature recognition outcome to prevent breast cancer.

INTRODUCTION

According to WHO cancer has been responsible for the deaths of millions of peopleworldwide with an estimated increase of 50% for developing countries and for 70% of the total deaths due to cancer. According to Parkin in [1] developing nations only possess 5% of global funds for cancer control and very few human and material resources are also available in such countries. Breast cancer is a type of cancer which affects the breast tissue which is most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk. Breast cancer is caused by a number of factors called risk factors; they are classified as either modifiable (those that can be controlled like habits, environmental hazards, etc) or non-modifiable factors (those that cannot be controlled like, gender, family history etc). According to the Collaborative Group on Hormonal Factors in Breast Cancer presented in 2002, the primary risk factors for breast cancer are being female and of an older age. Other potential risk factors include: family history of breast cancer, age of menarche (first occurrence of menstruation), age of first birth, age of menopause, body weight, alcohol consumption, exposure to radiation, higher hormonal levels and diet. In the United States, there were 1,665,540 new cancer cases and 585,720 cancer deaths in 2014. Approximately 30% of cancer diagnosed in women was breast cancer, which led to approximately 15% of cancer deaths in 2014 [2]. With the increasing development of biomedical and computer technologies, various clinical factors related to breast cancer have been recorded. This overall crisis led to the development of optimised data minning tricks and methods in order to classify and predict the presence of cancer or the type of cancer the patient is suffereing from.

LITERATURE SURVEY

The paper [3] presented is a study about breast cancer prediction based on data mining methods to discover an effective way to predict breast cancer. The objective of the author of this paper is to compare and identify an accurate model to predict the incidence of breast cancer based on various patients' clinical records. Four data mining models such as support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, AdaBoost tree. To test the influence of feature space reduction, a hybrid between principal component analysis (PCA) and related data mining models is proposed, which applies a principle component analysis method to reduce the feature space. To evaluate the performance of these models, two widely used test data sets are used, Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995). 10-fold cross-validation method is implemented to estimate the test error of each model. The results performed by this analysis demonstrate a comprehensive trade-off between these strategies and also provides a detailed evaluation on the models.

The author of this paper [4] aims to review on various data mining techniques that are specifically considered on breast cancer prediction. A varieties of data mining techniques are highlighted and a detailed description have benn highlighted such as Decision Tress, Artificial Neural Network, Genetic Algorithm and Support Vector Machine. Each method possesses its own merits and demerits. This paper intends to provide the reviews conducted by the various experts in the field of data mining systems. From the above study, we can infer that there is a still lack of early diagnosis, accuracy, sensitivity and specificity of the breast cancer data.

In this paper [5] two different data mining classification techniques was used for the prediction of breast cancer risk and their performance was compared in order to evaluate the best classifier. Experimental results shows that the J48 decision trees is a better model for the prediction of breast cancer risks for the values of accuracy, recall, precision and error rates recorded for both models. Hence, an efficient and effective classifier for breast cancer risks has been identified while the number of attribute covered by the classifier can be increased by increasing the sample size of the training set and hence the development of a more accurate model.

In [6], the author have utilized the genetic systems for the forecast of breast tumor. This system is hybrid with the decision tree, ANN and logistic regression. They used 699 records acquired from the breast cancerous patients at the University of Wisconsin. They utilized 9 indicator variables and 1 result variable for the information investigation with 10-fold cross approval. The researchers asserted that their genetic prediction model gives precision as much as 99%.

Lipo Wang, Feng Chu in [7] proposed the cancer prediction using gene expression data. They found the minimum gene probability. Two approaches were proposed namely, gene selection and gene ranking scheme. Based on the ranking scores, the prediction of the malignant breast cancer is detected. They also employed T-test and class separablity.

The author of this paper [8] utilized Support Vector Machine as a nonlinear mapping to move the training data into the view of high dimensional spaces. This new dimension allowed searching for linear optimal hyper plane. The SVM discovered this hyper plane utilizing support vectors and edges. The Support Vector Machines (SVM) is a general class of learning architectures, propelled by the statistical hypothesis.

PROBLEM DESCRIPTION

In this problem, we are using Wisconsin Diagnostic Breast Cancer (WDBC) data set that provides many attributes for the classification on diagonisis of Breast Cancer types, i.e, Malignant and Bengin. Therefore we will be applying different classification approaches such as logistic regression, SVM, decision trees and atlast we will be implementing it by deep learning approach. Using Neural Network and Stochastic Gradient Descent, the classification approach seems optimal for large number of data set. Moreover, in order to avoid Curse of Dimensionality, we have to select minimum number of attributes as possible that affects the result most using Principal Component Analysis (PCA).

REFERENCES

[1] American Cancer Society (2005). "**Breast Cancer Facts & Figures 2005–2006**" (PDF). http://web.archive.org/web/20070613192148/http://www.cancer.org/downloads/STT/CAFF2005BrFacspdf2005.pdf. 13 June 2007. Retrieved 2013-02-26.

[2] American Cancer Society (2007). "**Cancer Facts & Figures 2007**" (PDF). 10 April 2007. http://web.archive.org/web/20070410025934/, http://www.cancer.org/downloads/STT/CAFF2007PWSecured.pdf. Retrieved 2012-11-26.

[3] Haifeng Wang, Sang Won Yoon, "**Breast Cancer prediction using Data Mining Methods",** Proceedings of the 2015 Industrial and Systems Engineering Research Conference, S. Cetinkaya and J. K. Ryan, eds. October 2015.

[4] M Deepika, L Mary Gladence, and R Madhu Keerthana, "**A Review on Prediction Of Breast Cancer Using Various Data Mining Techniques**", Research Journal of Pharmaceutical, Biological and Chemical Sciences, ISSN: 0975-8585, pg-808, Januray- February 2016.

[5] Peter Adebayo Idowu, Kehinde Oladipo Williams, Jeremiah Ademola Balogun and Adeniran Ishola Oluwaranti, "**Breast Cancer Risk Prediction Using Data Mining Classification Techniques",** Transactions on Network and Communication, ISSN: 2054-7420, Vol. 3, Issue 2, March 10 2015.

[6] RuiXu, Anagnostopoulos, G.C. And Wunsch, D.C.I.I.,"**Multiclass Cancer Classification Using Semi supervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data**", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol.4, No.1, Pp. 65-77, 2007.

[7] Lipo Wang, Feng Chu, And Wei Xie, "**Accurate Cancer Classification Using Expressions Of Very Few Genes**", IEEE/ ACM Transactions On Computational Biology And Bioinformatics, 4, 40-52, 2007.

[8] XiaoweiSonga, Arnold Mitnitskib,c, JafnaCoxb, Kenneth Rockwood - **Comparison of Machine Learning Techniques with Classical Statistical Models in Predicting Health Outcomes**, MEDINFO 2004 M. Fieschi et al. (Eds) Amsterdam: IOS Press © 2004 IMIA