Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Study on Data Mining Techniques for Breast Cancer Prediction

Harshnika Bhasin
M.Tech Scholar
All Saint's College of Technology ,Bhopal, India

Abstract — This paper presents a study of different techniques of information mining algorithms used for the aim of predicting carcinoma because it is understood to any or all that prediction of carcinoma survivability has been a difficult research problem for several researchers. Since the early dates of the related analysis, a lot of advancement has been recorded in many related fields. For an instance, a sincere thanks to existing biomedical technologies, higher instructive prognostic factors are being measured and recorded; because of low value computer components and software system technologies, high volume good quality information is being collected and keeps automatically; and at last thanks to higher analytical strategies, those voluminous information is being processed effectively and with efficiency. Therefore, the most important objective of this manuscript is to report on a research project where we have a tendency to take advantage of these available technological advancements to develop prediction models for carcinoma survivability.

Keywords — Cancer prediction, Breast cancer detection, association rule mining, pattern analysis,hybrid apriori.

I. Introduction

The cancer generally develops once cells of a part of the body begin to grow out of managing. These additional cells make a mass of tissue, referred to as a growth or neoplasm. Tumors are benign or malignant. The science whose goal is to perform classification of objects into a variety of classes or categories is referred to as Pattern Recognition. Objects may be pictures, signal waveforms or any variety of measures that must be classified [1].

The importance of the study is that; the carcinoma refers to severe malignancies that develop in one or each breast and it is the most typical type of cancer among ladies in developed countries. According to U.S. Cancer Society one in eight ladies can develop carcinoma throughout their period. The matter with carcinoma identification is that despite radiographies breast imaging and screening has allowed for a lot of correct identification of carcinoma, 100% to a half-hour of malignant cases don't seem to be detected for numerous reasons. There are basically two types of errors typical in examining mammograms. They're False Positives (FP) and False Negatives (FN). The Computer-Aided designation (CAD) will cut back each the FP and also the FN identification rates. CAD is an application of pattern recognition aiming at aiding doctors in creating diagnostic decisions. Identification designation is formed by the doctor. Our aim is to utilize a pattern recognition system so as to help medical specialist with a "Second" opinion by closing the mammographic mass options that the majority indicates malignancy.

Cancer is basically a malignant cell that becomes a significant reason behind the death and hardly prevented [2, 3]. India is one of the growing carcinoma epidemics with an increasing range of younger ladies turning into susceptible to the unwellness. An international study estimates that by 2030, the amount of recent cases of carcinoma in India can increase from the present 115,000 to around 200,000 annually. In keeping with Globocan knowledge (International Agency for analysis on Cancer), India is on prime of the table with 1.85 million years of healthy life lost because of carcinoma. The study confirmed conclusions from earlier research: that carcinoma is currently the second most typical cancer diagnosed in Indian ladies after cervical cancer. Studies have additionally shown that Indian ladies develop carcinoma roughly a decade sooner than women in western countries. Poor survival is also mostly explained by lack of or restricted access to early detection services and treatment. In medical domains data processing approaches are increasing quickly because of the advance effectiveness of those approaches to classification and prediction systems, new and novel analysis directions are known for more clinical and scientific research. Usually cancer analysis relies upon statistical models failed to reach to spread in medical as a result of the use of these tools aren't belongs to a medical community. A knowledge classification method using information obtained from well-known knowledge has been one amongst the foremost intensively studied subjects in statistics. There are several techniques to predict and classification carcinoma pattern.

Most women have over one identified a risk issue for carcinoma, however, can ne'er get the unwellness. The foremost common risk factors for carcinoma isn't solely being feminine and growing older. There is also over one reason behind carcinoma. These could be:

- Being a lady
- Getting older
- Having an hereditary mutation within the brca1 or brca a pair of carcinoma cistron
- Lobular cancer in situ (LCIS)
- A personal history of breast or female internal reproductive organ cancer

- A case history of breast, female internal reproductive organ or glandular cancer
- Having high breast density on a mammogram
- Having a previous diagnostic assay showing atypical dysplasia
- Starting change of life when age fifty five
- Never having youngsters
- Having your first child when age thirty five
- Radiation exposure, frequent x-rays in youth
- High bone density
- Being overweight when change of life or gaining weight as an adult

II. Related data

Many researchers are applying numerous algorithms and techniques like Classification, Clustering, Regression, AI, Neural Networks, Association Rules, decision Trees, Genetic algorithmic rule, Nearest Neighbor technique, etc., to assist health care professionals with improved accuracy within the identification of carcinoma. In our study, we've got used the carcinoma dataset from the University Medical Centre, Institute of medical specialty, Ljubljana, Yugoslavia. Thanks to M. Zwitter and M. Soklic for providing the information. This literature showed that there are many studies on the survivability prediction downside of carcinoma. These studies have applied various approaches to the given drawback and achieved high classification accuracies.

There is plenty of work done for numerous diseases like cancer like shown in paper [4]. As a technique employed in it's terribly convenient since the decision Tree is easy to grasp, works with mixed knowledge varieties, models, non-linear functions, handles classiest, and most of the promptly obtainable tools use it. Even within the paper [5] that I referred discusses however information warehousing, data processing, and decision support systems will scale back the national cancer burden or the oral complications of cancer therapies. For this goal to be achieved, it 1st are necessary to observe populations; collect relevant cancer screening, incidence, treatment, and outcomes data; determine cancer patterns; justify the patterns, and translate the reasons into effective diagnoses and coverings. A successive paper that I referred [6] contains the analysis of the breast masses in a very series of pathologically evidenced tumors using data processing with a decision tree model for classification of breast tumors. Accuracy, sensitivity, specificity, positive prognostic value and negative prognostic value are the five most typically used objective indices to estimate the performance of diagnosing results. Sensitivity and specificity are the foremost two necessary indices that a doctor involved concerning. With sensitivity 93.33% and specificity 96.67%, the proposed technique provides objective evidences for good diagnoses of breast tumors. Pascal Boilot [7] and his team report on the utilization of the Cyranose 320 for the detection of microorganism inflicting eye infections using pure laboratory cultures and also the screening of microorganism associated with ENT infections using actual hospital samples.

A. Sudha et al [8] offers an inspiration regarding major critical diseases and their identification using data processing with minimum range of attributes and creates awareness regarding diseases that end up in death. K. Balachandran et al [9] Early detection of the cancer sickness is crucial in identifying and treating the patient. Thus, it's terribly essential that the common person who has some symptoms and risk factors are higher to endure checkup by a specialist at the earliest.

Delen et al. [10] had taken 202,932 carcinoma patients' records that then pre-classified into 2 teams of "survived" 93,273 and "not survived" 109,659. The results of predicting the survivability were within the range of 93 accuracy. Tan AC's [11] used C4.5 decision tree, bagged decision tree on seven in public accessible cancerous micro array knowledge, and compared the prediction performance of those strategies. Liu Ya-Qin's [12] experimented on carcinoma information using C5 algorithmic rule with bagging to predict carcinoma survivability.

Jinyan LiHuiqing Liu's [13] experimented on female internal reproductive organ, tumor information to diagnose cancer using C4.5 with and without bagging. Dong-Sheng Cao's [14] planned a replacement decision tree based mostly ensemble technique combined with feature choice technique backward elimination strategy with sacking to seek out the structure activity relationships within the area of Chemometrics associated with pharmaceutical trade. My Chau Tu's [15] planned the utilization of bagging with C4.5 formula, bagging with a Naive Bayes formula to diagnose the center sickness of a patient. My Chau Tu's [16] used bagging formula to spot the warning signs of cardiopathy in patients and compared the results of decision tree induction with and while not bagging.

Tsirogiannis's [17] applied bagging algorithmic program on medical databases using the classifier neural networks, SVM'S and decision trees. Results exhibits improved accuracy of bagging than while not bagging. Pan wen [18] conducted experiments on ECG information to spot abnormal high frequency cardiograph exploitation decision tree algorithmic program C4.5 with bagging. Kaewchinporn C's [19] conferred a replacement classification algorithmic program TBWC combination of decision tree with bagging and clustering. This algorithmic program is experimented on 2 medical datasets: cardiocography1, cardiocography2 and alternative datasets not associated with medical domain. Chaurasia and Pal [20, 21] conducted a study on the prediction of coronary failure risk levels of the heart disease information with data processing technique like Naïve Bayes, J48 call tree and bagging approaches and CART, ID3 and decision table. The result shows that bagging techniques, performance is a lot of more accurate than Bayesian classification and J48.

III. Breast Cancer Dataset

The data used in this study are provided by the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. My special thanks go to M. Zwitter and M. Soklic for providing the data for this research work. The dataset has 10 attributes and total 286 rows,

1. Age: patient's age at the time of diagnosis;
2. Menopause: menopause status of the patient at the time of diagnosis;
3. Tumor size: tumor size (in mm);
4. Inv-nodes: range 0 - 39 of axillary lymph nodes showing breast cancer at the time of histological examination;
5. Node caps: penetration of the tumor in the lymph node capsule or not;
6. Degree of malignancy: range 1-3 the histological grade of the tumor. That is grade: 1 predominantly that consist of cancer cells, grade: 2 neoplastic that consist of the usual characteristics of cancer cells, grade: 3 predominately that consist of cells that are highly affected;
7. Breast: breast cancer may occur in either breast;
8. Breast quadrant: if the nipple considers as a central point the breast may be divided into four quadrants;
9. Irradiation: patient's radiation (x-rays) therapy history.
10. Class: no-recurrence or recurrence depending reappearing symptoms of breast cancer in the patients after treatment.

IV. Proposed Work

To overcome the issues of the previous work; A prognostic tool is planned that helps oncologist in identification of carcinoma and so helps oncologist in deciding in treatment technique. Controlling the mortality and enhancing the survivability of carcinoma patients is the main objective of this model. This model may facilitate in predicting carcinoma at the initial stage that saves lots of valuable time. The proposed model uses data cleansing algorithmic program to clean information and to get error free, correct and complete knowledge for prediction. Therefore, machine learning classifiers are trained with error free information which reinforces the prediction accuracy of classifiers once checked with test knowledge. Moreover, in situations once good oncologists aren't out there for the patient, prognostic model created with machine learning techniques will support different doctors in deciding and within the format of patient initial treatment with none delay.

The planned tool framework uses numerous machine learning techniques and decision support system for precise and correct prediction. The tool uses a feature selection rule to boost the parameters like accuracy, preciseness and Sensitivity. Next step is to train the classifier to evaluate the classifier performance on specific information set in order that the classifier will give correct results in an economical manner. In the proposed work hybrid apriori is implemented for the purpose of extracting the frequent pattern, this algorithm is pretty faster than conventional apriori and improved apriori.

Proposed Apriori Hybrid Algorithm

forall *large k-itemsets* l_k, $k \geq 2$ do begin
    $H_1 = \{$*consequents of rules derived from* l_k *with one item in the consequent* $\}$;
    call Hybrid-Rulegen (l_k, H_1);
end

Function Hybrid-Rulegen $(l_k$*:large k-itemset* , H_m*: set of m-item consequents)*
if $(k > m+1)$ then begin
    H_{m+1} = *apriori-gen*(H_m);
    forall $h_{m+1} \in H_{m+1}$ do begin
        $conf = support(l_k)/support(l_k - h_{m+1})$;
        if $(conf \geq minconf)$ then
            output *the rule* $(l_k - h_{m+1}) \Rightarrow h_{m+1}$
            with confidence $=$ *conf and support* $=$ $support(l_k)$
        else
            delete h_{m+1} *from* H_{m+1}
    end
    call Hybrid-Rulegen (l_k, h_{m+1});
end

V. Conclution

This paper provides a study of assorted technical and review papers on breast cancer identification and prognosis problems and explores that data processing techniques supply nice promise to uncover patterns hidden in the information that can facilitate the clinicians in decision making. From the above study it is determined that the accuracy for the diagnosis analysis of numerous applied data processing classification techniques is very acceptable and may facilitate the medical professionals in deciding for early identification and to avoid biopsy.In this paper, we tend to propose an economical hybrid model for the prediction of the positivity and negativity of the carcinoma relying upon the training data set. Numerous association rule mining algorithms are combined to boost the accuracy within the prediction of positive and negative. Approach like this helps the doctors in better decision making by which precious life of several patients can be saved.

References

[1]. Rafael C. Gonzalez, Richard E. Woods. "Digital Image Processing", New Jersey, Prentice Hall, 2002.

[2]. Delen D, Patil N. Knowledge extraction from prostate cancer data. The 39thAnnual Hawaii International Conference on System Sciences; 2006; 1-10.

[3]. National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program Public-Use Data (1973-2008). Cancer Statistics Branch; 2011.

[4]. Nevine M. Labib, and Michael N. Malek, "Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia", World Academy of Science, Engineering and Technology 8 2005

[5]. G. A Forgionne, A. Gagopadhyay, and M. Adya, "Cancer Surveillance Using Data Warehousing, Data Mining, and Decision Support Systems", Topics in Health Information Management, vol. 21(1); Proquest Medical Library, August 2000

[6]. W. Kuo, R. Chang, D. Chen and C. C. Lee, "Data Mining with Decision Trees for Diagnosis of Breast Tumor in Medical Ultrasonic Images", Breast Cancer Research and Treatment, Dordrecht, vol. 66, Iss. 1, Mar 2001.

[7]. Pascal Boilot, Evor L. Hines, Julian W. Gardner, Member, IEEE, Richard Pitt, Spencer John, Joanne Mitchell, and David W. Morgan, "Classification of Bacteria Responsible for ENT and Eye Infections Using the Cyranose System", IEEE SENSORS JOURNAL, vol. 2, NO. 3, JUNE 2002.

[8]. A.Sudha "Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability" International Journal of Computer Applications (0975 – 8887) Volume 41– No.17, March 2012.

[9]. K.Balachandran, Dr. R.Anitha "Supervised Learning Processing Techniques for Pre-Diagnosis of Lung Cancer Disease" ©2010 International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 4.

[10]. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.

[11]. Tan AC, Gilbert D. "Ensemble machine learning on gene expression data for cancer classification", Appl Bioinformatics. 2003;2(3 Suppl):S75-83.

[12]. Liu Ya-Qin, Wang Cheng, Zhang Lu," Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data" , 3rd International Conference on Bioinformatics and Biomedical Engineering , 2009.

[13]. Jinyan LiHuiqing Liu, See-Kiong Ng and Limsoon Wong," Discovery of significant rules for classifying cancer diagnosis data", Bioinformatics 19(Suppl. 2) Oxford University Press 2003.

[14]. Dong-Sheng Cao, Qing-Song Xu ,Yi-Zeng Liang, Xian Chen, "Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity", Chemometrics and Intelligent Laboratory Systems.

[15]. My Chau Tu, Dongil Shin, Dongkyoo Shin ,"Effective Diagnosis of Heart Disease through Bagging Approach", 2nd International Conference on Biomedical Engineering and Informatics,2009.

[16]. My Chau Tu, Dongil Shin, Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms" Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009.

[17]. Tsirogiannis, G.L, Frossyniotis, D, Stoitsis, J, Golemati, S, Stafylopatis, A Nikita,K.S," Classification of Medical Data with a Robust Multi-Level Combination scheme", IEEE international joint Conference on Neural Networks.

[18]. Pan Wen, "Application of decision tree to identify a abnormal high frequency electrocardiograph", China National Knowledge Infrastructure Journal, 2000.

[19]. Kaewchinporn .C, Vongsuchoto. N, Srisawat. A " A Combination of Decision Tree Learning and Clustering for Data Classification", 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE).

[20]. V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.

[21]. V. Chauraisa and S. Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.j.SciTech,, Vol.1, pp. 208-217, 2013.

