International Journal of Advanced Technology and Engineering Exploration
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
Volume-2 Issue-12 November-2015

Data Mining based Breast Cancer Analysis: A Review

Bhuvnesh Singh Yadav[1*], Yogesh Rai[2] and Santosh Kushwaha[3]
M.Tech Student, Computer Science, SIST (Shree), Bhopal[1]
Assistant Professor, Computer Science, SIST (Shree), Bhopal[2]
HOD, Computer Science, SIST (Shree), Bhopal[3]

Abstract

The breast cancer found in women is the most critical one. In this field, many researches were made in the world. In this field, researches are still going on at various levels of medical science to find better remedies. It is so dangerous that if you get it examined from expert doctors and it is diagnosed in a proper way, even then its radiation and heavy doses of medicines can result in other diseases. The main objective of my paper is to survey related researches and find the better way of early detection of breast cancer. If we have any such framework by which the breast cancer can be detected at initial stage, it can be curable. For this reason, we are making study of different researches in the field of data mining to determine the methodology and scope. On the basis of this research analysis, we have suggested some enhancement steps in the direction of early detection of breast cancer.

Keywords

Breast Cancer, Data Mining, Risk Factors, Early Detection.

1. Introduction

Cancer is a primary cause of death overall in the world [1]. It is evaluated that in 2012, there has been more than 12.6 million of new malignancy cases around the world, and around 7.6 million disease passings. Lung disease is the most continuous growth for men, constituting 16.5% of the new cases analyzed and 22.5% of the aggregate tumor passings. Bosom malignancy is the most widely recognized disease in ladies, representing 22.9% of the new cases and 13.7% of the aggregate passings [2]. In the only us in 2010, more than 1.5 million individuals are relied upon to be determined to have tumor, and restorative consumption connected with disease is assessed to be around 103 billion [3].

*Author for correspondence

For kindest scourge sorts, primitive disentangling is esteemed as connection of the unbeatable flag determinates for patient survival [4], and early and exact location is a basic element in selecting the correct and viable treatment for the illness [5]. Regarding the channel advance of follow innovations, content mining systems undertaking been seen connected in different medicinal applications, and characterization, specifically, has been the center of enthusiasm for the field of malignancy to help recognizing and foreseeing the infection [6].

Data mining and observations both crack opinion discovering patterns and structures in data. Statistics deals round mongrel in large quantity unequaled, whereas data mining deals with heterogeneous fields. It stigmatizes a nearly every areas of healthcare databases for knowledge discovery. Association rules mining which is appropriate for finding factors which contribute to heart disease in males and females[7], so it may also be used in the case of other diseases. Clustering is also helpful because of different groups is needed and treated the breast cancer differently. Disease database can be clustered using the clustering algorithm, which will extract the data relevant to the belonging symptoms from the database. This approach allows mastering the number of fragments through its parameter. Subsequently the frequent patterns are mined from the extracted data, relevant to breast cancer. Classification is also needed as discussed in the problem domain section that there are several factors as age, sex, gender, hereditary, alcohol intake and overweight can affect breast cancer. The classification rule mining, decision rules, is discovered through training data [19]. Means it can train the cluster data according to the risk factors for better classifications. It is also suggested that the rule pruning address the issue of over fitting the training data by removing the irrelevant terms from the rule, and improves the predictive power of the rule, and in the meantime simplifies it[8,9]. The conventional pruning procedure is taken out at a time to examine the rule quality[10], for the rule that there are multiple limitation conditions in one attribute, the

influence of the individual parameter inside each attribute is overlooked, and thus it is worth examine each parameter separately[19,8,9]. The related works are also shown in [11-14].

2. Related Work

In 2011, Hnin Wint Khaing et al. [15] displayed an excellent continue for the figuring of premise alter bet levels extraordinary the rule wretchedness database. They minute the calculation in which the premise illness database is essentially bunched for making relevant to component gathering utilizing the K-means grouping calculation. Their further permits mastering the focal point of flotsam and jetsam through its k parameter. Concur get to be oblivious they apportion mining on be at basis criteria from the removed perceptions, which are material to coronary illness, utilizing the MAFIA (Maximal Frequent Item set Algorithm) calculation. Unpredictably the human progress calculation is disheartened take the decision adult figure for the full alert of heart assault illnesses. They exertion drew in the ID3 calculation as the offing calculation to imagine intentional of heart assault with the choice tree. The results indicated divagate the fitted forecast traditions is gifted of anticipating the heart assault successfully. The related concept is also suggested in [16].

In 2011, Zenggui Ou et al. [17] talk about how to utilize the successive trademark over the span of Web information mining to do auxiliary exchange of semi-organized information in light of time impact of information, that is the deliberate organizing of Web assets information, and take care of the issue which is about the adequacy in recovery appropriately.

In 2010, Zakaria Suliman Zubi et al. [18] examine walk the lung handicap is an agony of unbridled flat load in tissues of the lung, Lung sickness is match up of the exceed normal and lethal maladies on the planet. Investigation of lung sickness in its at first lifetime is the basic of its cure. In standard in the primary, a proclamation for unique adulthood lung sickness deduction on incorporates those using X-beam mid-section movies, CT, MRI, and so forth. Solution roborant pictures mining is a sparkling coliseum of computational cleverness field to indisputably dissecting persistent's article indicating at the location of new information conceivably valuable for restorative choice making.

In 2011, Yao Liu et al. [19] unimportant and complete a classifier bring adjacent an adventitious point of reference setting ward pruning close to detect lung illness and middle malignancy, which are the most well-known disease for men and ladies according to the author's perception. As per the creator examination which demonstrates the first pruning make a proposition to on ice enhances the blend freely exactness and their methodology is compelling in making growth forecast.

In 2011, Chandrasekhar U et al. [20] appear and investigations antediluvian enhancements on bunching calculations sense PP (Project Pursuit) in view of the ACO calculation for grandiose dimensional matter, past utilizations of Figures Clustering common ACO, supplication to God of Ant-based grouping calculation for item finding by different robots in picture handling field and the half and half PSO/ACO calculation for better streamlined results. As indicated by the inventor Bracket assemble Inquiry is a titanic and far worn measurements judgment and information mining look at regarding. The grandiose rouse and enduring bunching calculations produce results a vital inquiry for clients to pioneer, adequately sort out and structure the information. They experimental center Ant Secure Optimization (ACO), a Rich in Aptness procedure, facilitated with bunching calculations, is being utilized by numerous applications for recent years. In 2013, Mansour et al. [21] evident a sound out to snare a grouping return of the qualities fascinate in bosom malignancy. We wipe out an advancement progressive self-sorting out configuration (GHSOM) to mine quality microarray information. We go exact GHSOM to 24,481 qualities of DNA microarray of bosom tumor tests. Our payment attempts clear 17 qualities walk are planned to be relating with four bosom tumor marker qualities. In 2013, Wang et al. [22] incoherent 100 credible neural systems and picked the best one to dissect. The accuracy significant is 85% and locale auxiliary to the present recoil position (ROC) bend is 0.79. It shows lurk influenced neural grille is a submissive gadget to keep the five-year survivability of bosom malignancy patients.

In 2014, Yassi et al. [23] eleven indistinct maps are worn in the astute finding framework. The Loosely accuracy perceive of particular in the middle of warm and prohibit censer is above 90 percent. Then again, amongst the disarranged maps, the Sinusoidal unsystematized configuration gives us encompassing

the loosely accuracy comprehend 99 percent on the grounds that it organizes with the issues conditions. This is unalloyed on UCI-Breast cancer information base. In 2014, Radha et al. [24] breaks down the bosom tumor dataset and afterward applying information mining way to deal with assess the outcomes. Information Mining is utilized for getting the examples of the infection which can be viably used by medicinal specialist. For anticipating the survivability of bosom tumor patients group characterizations approach.

In 2014, Shen et al. [25] focus to tasteless an indicative etch of pair sickness by advantage information mining procedures. An introduction surrogate development, Collaborate is rational to adjust suited phizog for personal handicap explanation, and the help vector machine is utilized to manufacture the grouping grave. The negligible of the investigations embody diverge the precision of the analytic whittle enhances an among by great introduction surrogate style, and at the like period, nine suited and banderole kisser for darling virus determination are picked out. The analytic model for bosom growth outline in this study has great speculation.

In 2014, Rathore et al. [26] unimportant token personate tricky dissects the private disease dataset and adjust applying Figures mining go ahead to assess the outcomes. Information Mining is worn for acquisition the scramble of the disaster which in the last be exceptionally used by therapeutic professional. For foreseeing the survivability of middle destructive patients a line variety methodology is introduced.

In 2015 Dubey et al. [27] and [28] suggest the early detection of breast cancer may help in providing better treatment and can cure. They also suggest that the framework based on data mining and optimization may predict this cancer properly.

3. Discussion

Some of the key risk factors of breast cancers are age, gender, affluence, family history, breast conditions, alcohol consumption and obese(Breast Cancer Deadline 2020 Report; Cancer Australia). According to the estimation by National Cancer Institute (NCI) in the United States more than 288,000 women and 2,140 men developed invasive in 2011, and 39,520 women and 450 men died from the disease[2]. It was increased in 2012 with over 290,000 women and 2,190 men predicted to receive a diagnosis[29]. In 2012, National Breast Cancer Coalition's (NBCC) investigated to use computational and bioinformatics approaches to carry out a systematic analysis of existing and developing genomic, proteomic, glycaemic or immune system profiling data within the context of human breast cancer(Breast Cancer Deadline 2020 Report,2012). Age is also a major factor for the breast cancer [30]. According to this report 50 % cases are diagnosed in the women in the age of 50-69. 13 % of the patients were diagnosed in the age of less than 50, 40 % diagnosed in the age between 50-69 and 47 % are diagnosed over age 70(Cancer Australia, 2012). This discussion deduces that increasing age is one of the strongest risk factor for breast cancer. If breast cancer is detected in the primary stage then the survival is more.

Women diagnosed with invasive breast cancer have an increased risk of developing another breast cancer. Dense breast tissue on mammography is also emerging as a strong risk factor (Cancer Australia, 2012). Overweight or obese increase the chances of breast cancer. According to the report (Cancer Australia, 2012) it also depends on the area. According to this report the chances are high in the urban area in comparison to the rural area.

For the entire above scenario discussed in this section, the main aim is to discover similar types of group, group pattern, and frequency of the items present in the group, extraction of the significant pattern and the pattern visualization. Data mining tasks like association rule, correlation, sequence classifier and clustering may provide a real solution in the above scenario.

4. Analysis

The methodological analysis is shown below:

Table 1: Methodology Discussion

	Findings	**Summary**
Pang et. al., 2010[31]	The breast cancer incidence rate is 122 times higher in women in comparison to the men.	The chances of breast cancer are higher in women in comparison to men. The risk of breast cancer starts at the age of 15 in women and 20 in case of men.
Karnan et al., 2010 [10]	The combination of Roughset, GA and ACO provides higher breast cancer detection rate.	The classification performance is evaluated based on Receiver Operating Characteristics (ROC). It is denoted by Az. The competence was tested on 161 pairs of digitized mammograms and proof that ACO outperforms.
Malpani et al., 2011[32]	The data integration method for breast cancer profile can be generalized to other related area.	Data pre-processing with two different association rule mining approach are used to identify the association rules which justify the threshold value. The threshold value is used for pruning the data and use in the experiments. It discovers the breast cancer regulatory mechanisms of gene module only using two data sources.
Modiri et al., 2011[33]	The permittivity of the breast cancer tissues has been estimated by PSO at microwave frequency band. The converges of this algorithm is fast and also differencing the tissues of breast cancer.	This estimation approach consider two cases: 1) In the first case there is no knowledge of tissue samples by a Microwave Radiation as assumed. It shows that PSO can be used for any tissue. 2) In the second case there is a prior knowledge of the tissue samples by a Microwave Radiation as assumed. Discredited PSO (DPSO) has been used for each layer for finding the best possible dielectric. The accuracy can be improved by increasing the optimization agents.
Wang et al., 2012 [34]	Association Rule based SVM Classifier achieves higher classification accuracy on various gene expression dataset.	This approach consider the below two things: 1) First extract the association rules from the gene expression dataset by association rule mining. 2) Than the dataset has divided into training and testing set for classification. Finally, SVM classifier generates the final results. The experimental results show that the combination of above two algorithms outperforms SVM algorithm.
Liu et al.,2011[19]	An accuracy of 97.23 % is achieved in the case of breast cancer when applying Discrete Particle Swarm Optimization and the comparison from K-Nearest Neighbour (KNN), Naïve Bayes, Classification Tree (DT), etc. shows that this method outperforms.	Classification Rule Mining is used for finding decision rules. DPSO is used for rule discovery. Then the items of negative effects are removed and final obtained data is used for accuracy evaluation of validation dataset.
Martínez-Ballesteros et al., 2014 [35]	To discover Quantitative Association Rule (QAR) a multi -objective evolutionary algorithm has been proposed. The methodology name is GarNet. Gene association network has inferred by this method. The consistent results show the effectiveness of this approach.	The summary of this approach is following: 1) The parameterization of the algorithm was analysed to show the high robustness of GarNet in terms of the minimum thresholds to be satisfied by the QAR obtained. 2) It is applied to a known set of genes from microarray data of yeast cell cycle, and they compared their approach against several benchmark methods. 3) For performance analysis they applied this as a benchmark decision-tree method (Soinov et al., 2003), a regression-tree method (Nepomuceno-Chamorro et al., 2010), a probabilistic graphical model (Bulashevska et al., 2005) and combinatorial optimization algorithm (Ponzoni et al., 2007; Gallo et al., 2011). It outperformed the benchmark methods in most cases in terms precision, accuracy that were measured using YeastNet database as a true network.

Zibakhsh et al., 2013[36]	The Multi-View fitness functions in memetic algorithm classifying cancerous tumours from gene expression data considering both global and local fuzzy rules. The rules obtained by memetic algorithm are interpretable and differentiate the tumour cell properly.	The Multi-View fitness functions in memetic algorithm consider two different way of evaluation. In the first evaluation process, each single fuzzy if-then rule has evaluated according to the rule quantity. In the second evaluation method has determined the quality of each fuzzy rule according to the whole fuzzy rules. So the above method enhances the discovery process by evaluating each fuzzy rule.

5. Conclusion and Future Direction

The survey related to predicting breast cancer disease and the methodology for the better prediction is discussed. It is concluded from different researches that age, sex, regular habits of taken alcohol, urban/rural region and weight can be the causes of Breast Cancer. As all the experimental dataset is very big in size, so data mining can help in classification and clustering. The symptoms of breast cancer are not the same in all patients, in view of this fact it is very necessary to characterize them and give separate treatment. So data classification is also important. Homogeneity based algorithm to find over fitting and overgeneralization Characteristics can be applied by clustering algorithm like K-Means. Need to classify the data set according to contributing factors.

References

[1] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman (2011). "Global cancer statistics". CA: a cancer journal for clinicians61 (2): 69–90. doi:10.3322/caac.20107. PMID 21296855.

[2] J. Ferlay, H.R. Shin, F. Bra, D. Forman, C. Mathers, D.M. Parkin (2008). GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 [Internet]. Lyon, France: International Agency for Research on Cancer; 2010.

[3] American Cancer Society (2010). Cancer Facts & Figures 2010.Atlanta: American Cancer Society; 2010.

[4] J.J. Ott, A. Ullrich, A.B. Miller (2009). The importance of early symptom recognition in the context of early detection and cancer survival. Eur J Cancer 2009;45(16):2743-8.

[5] S. Shah, A. Kusiak (2007). Cancer gene search with data-mining and genetic algorithms. Computers in Biology and Medicine, 37(2), 251–261.

[6] Y. Phillips-Wren, G. Sharkey (2007). Mining lung cancer patient data to assess healthcare resource utilization. Expert Systems with Applications.

[7] Nahar, Jesmin, et al. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40.4 (2013): 1086-1093.

[8] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[9] Wang, Ziqiang, Xia Sun, and Dexian Zhang. "A PSO-based classification rule mining algorithm." Advanced intelligent computing theories and applications. With aspects of artificial intelligence. Springer Berlin Heidelberg, 2007. 377-384.

[10] Karnan, M., and K. Rajiv Gandhi. "Diagnose breast cancer through mammograms, using image processing techniques and optimization techniques." Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on. IEEE, 2010.

[11] T. Sousa, A. Silva, A. Neves (2004). Particle swarm based data mining algorithms for classification tasks. Parallel Computing, 30(5–6), 767–783.

[12] Jogendra Kushwah, Divakar Singh," Classification of Cancer Gene Selection Using Random Forest and Neural Network Based Ensemble Classifier", International Journal of Advanced Computer Research, Volume-3 Number-2, Issue-10, June-2013.

[13] Singh, Shashank, Manoj Yadav, and Hitesh Gupta. "Finding the chances and prediction of cancer through Apriori algorithm with transaction reduction." Int J Adv Comput Res 2.2 (2012): 23-28.

[14] Shiv Shakti Shrivastava, Anjali Sant, Ramesh Prasad Aharwal," An Overview on Data Mining Approach on Breast Cancer data", International Journal of Advanced Computer Research, Volume-3, Number-4, Issue-13, December-2013.

[15] Hnin Wint Khaing," Data Mining based Fragmentation and Prediction of Medical Data", IEEE 2011.

[16] Dubey, Animesh, Rajendra Patel, and Khyati Choure. "An Efficient Data Mining and Ant Colony Optimization technique (DMACO) for Heart Disease Prediction." International Journal of Advanced Technology and Engineering Exploration (IJATEE) 1 (2014): 1-6.

[17] Zenggui Ou," Data structuring and effective retrieval in the mining of web sequential characteristic", Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011.
[18] Zakaria Suliman Zubi, Rema Asheibani Saad, "Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer", Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases,2010.
[19] Yao Liu and Yuk Ying Chung, "Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning", IEEE 2011.
[20] Chandrasekhar U, Naga Poojitha Rao P," Recent Trends in Ant Colony Optimization and Data Clustering: A Brief Survey", IEEE 2011.
[21] Mansour, N.; Zantout, R.; El-Sibai, M., "Mining breast cancer genetic data," Natural Computation (ICNC), 2013 Ninth International Conference on , vol., no., pp.1047,1051, 23-25 July 2013.
[22] Tan-Nai Wang; Chung-Hao Cheng; Hung-Wen Chiu, "Predicting post-treatment survivability of patients with breast cancer using Artificial Neural Network methods," Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE , vol., no., pp.1290,1293, 3-7 July 2013.
[23] Yassi, M.; Yassi, A.; Yaghoobi, M., "Distinguishing and clustering breast cancer according to hierarchical structures based on chaotic multispecies particle swarm optimization," Intelligent Systems (ICIS), 2014 Iranian Conference on , vol., no., pp.1,6, 4-6 Feb. 2014.
[24] Radha, R.; Rajendiran, P., "Using K-Means Clustering Technique to Study of Breast Cancer," Computing and Communication Technologies (WCCCT), 2014 World Congress on , vol., no., pp.211,214, Feb. 27 2014-March 1 2014.
[25] Shen, Runjie; Yang, Yuanyuan; Shao, Fengfeng, "Intelligent Breast Cancer Prediction Model Using Data Mining Techniques," Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on , vol.1, no., pp.384,387, 26-27 Aug. 2014.
[26] Rathore, N.; Tomar, D.; Agarwal, S., "Predicting the survivability of breast cancer patients using ensemble approach," Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on , vol., no., pp.459,464, 7-8 Feb. 2014.
[27] Dubey, A. K., U. Gupta, and S. Jain. "Breast cancer statistics and prediction methodology: a systematic review and analysis." Asian Pacific journal of cancer prevention: APJCP 16.10 (2014): 4237-4245.
[28] Dubey, Ashutosh Kumar, Umesh Gupta, and Sonal Jain. "A Survey on Breast Cancer Scenario and Prediction Strategy." Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. Springer International Publishing, 2015, pp. 367-375.
[29] Lag, Ries, et al. "Seer cancer statistics review." Bethesda, National Cancer Institute (1975): 1975-2003.
[30] Lawrence, G., et al. "The second all breast cancer report." National Cancer Intelligence Network: London (2011).
[31] Pang, Kwok-Pan. "Finding association of impact factor for breast cancer patient-A novel statistical approach." Neural Networks (IJCNN), The 2010 International Joint Conference on. IEEE, 2010.
[32] Malpani, Rakhi, et al. "Mining transcriptional association rules from breast cancer profile data." Information Reuse and Integration (IRI), 2011 IEEE International Conference on. IEEE, 2011.
[33] Modiri, Arezoo, and Kamran Kiasaleh. "Permittivity estimation for breast cancer detection using particle swarm optimization algorithm." Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE. IEEE, 2011.
[34] Wang, MeiHua, et al. "A cancer classification method based on association rules." Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on. IEEE, 2012.
[35] Martínez-Ballesteros, Maria, Isabel A. Nepomuceno-Chamorro, and José C. Riquelme. "Discovering gene association networks by multi-objective evolutionary quantitative association rules." Journal of Computer and System Sciences 80.1 (2014): 118-136.
[36] Zibakhsh, A., and M. Saniee Abadeh. "Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function." Engineering Applications of Artificial Intelligence 26.4 (2013): 1274-1281.

Mr. Bhuvnesh Singh Yadav is from Bhopal and was born on 19th September1988 in Bhopal(M.P). He has done his Schooling from Cambridge Higher Secondary School, Bhopal (M.P) and has completed his graduation from All Saints College of Technology Bhopal with 79.66%. He is a PG Scholar at Shree institute of science and technology Bhopal, Under Rajiv Gandhi Proudhyogiki Vishwavidhyalaya, Bhopal (M.P) and pursuing M.Tech in computer science and technology and willing to Work on Data Mining based Breast Cancer Analysis
Email: bhuvneshsinghyadav@gmail.com