Diagnosis of Breast Cancer using Clustering Data Mining Approach

Jahanvi Joshi
Research Scholar
KSV University
Gujarat, India

Rinal Doshi
Director,
Mahavir Computer Acadmy,
Kalol,Gujarat, India

Jigar Patel, Ph.D
Research Guide
Gujarat Technological
University

ABSTRACT

The main objective of the research is to early diagnosis of the breast cancer patients. Nowadays Brest cancer becomes very major disease in many women not only in India but also in other country. For early diagnosis of the breast cancer patients, clustering data mining algorithm used to detect breast cancer. For the experimental purpose breast cancer dataset carried out form the UCI web data repository. The selection of appropriate clustering data mining technique is a challenge for the diagnosis of breast cancer. To get early result the challenges takes four clustering data mining techniques. This research becomes very helpful to doctor for diagnosis breast cancer and also helpful to patients for early treatment.

Keywords

Clustering, WEKA, Simple K-means, Breast Cancer, Data Mining

1. INTRODUCTION

Breast cancer is a major disease which is found in many women in India. Breast cancer is a disease in which cancer cells form in the tissues of the breast of the woman. The breast is made up of lobes (15 to 20 sections) and ducts. Ducts are thin cube for linking of breast for producing milk. The nipple and areola are outside of the breast and its color is dark than the breast. The most common type of breast cancer is begins in the cells of the ducts. Cancer that begins in the lobes or lobules found in both breasts than are other types of breast cancer. Warm, red, and swollen breast is a mark for breast cancer. Age and health history can affect the risk of developing breast cancer. Breast cancer is caused by gene changes. Chest x-ray, CT scan, BONE scan and PET scan are used to detect the stages of the breast cancer. Recurrent breast cancer is cancer that has come back after it has been treated. The cancer may come back in the breast, in the upper body wall, or any parts of the body. To solve the problem of breast cancer and early diagnosis of the breast cancer this research apply clustering data mining technique for finding the health of the breast cancer patient[1][2].

2. LITERATURE REVIEW

The authors of the paper focused on to improve health awareness. For achieve these they can give the example of self breast cleaning and awareness which will help to diagnosis breast cancer [3]. The author of this paper focus on building a platform on a world health origination for achieve these they need to implement prevention and screening programming to decries a cancer risk sector. They use a simple k-nearest-neighbor algorithm for optimum performance [4]. The author of the paper focus on knows risk factor for Brest cancer. They use a clinical breast care project (CBPC) data file and they analyzing and to develop a prototype for data mining .They use a Bayesian network analysis method for exploration of data interactions and find a caffeine CBCP population[5]. The author of this paper focus on find way to improve a Brest cancer victim's chance of long-term survival is to detect it as early as possible. They use a three statical technique for diagnose of Brest cancer First one is mammography, Second is FNA (fine needle aspirate) and Third is surgical biopsy [6]. The author of this paper focus on analyzing Brest Cancer data and challenging problem they use a clustering method find cancer in Brest cancer dataset. They can apply GHSOM to 24,481 genes of DNA microarray of Brest cancer tumor samples. And results have revealed 17 genes that are likely to be correlated with four breast cancer marker genes [7]. The author of this paper focus try to diagnosis of early Brest cancer of women using SVM, Tree Boost and Tree Forest data mining classification technique[8]. The author of this paper focus on SEER public – use-data to predict Brest Cancer. They use pre-classification method and find a possible solution to discover the information of Brest Cancer. [9]. The author of this paper focus on a transcription process for understanding the co-expressed gene seta under common regulatory mechanism. They have use data preprocessing method and two different association rules mining [10]. The author of the paper used different data mining technique for diagnosis & the prognosis of breast cancer with the main parameter of male and female gene behavior, they take gene expression data set of 311 instance to test and validate model and major the performance. They prove classification data mining algorithm provide more optimum outcome [11]. The author of the paper focus on different data mining classification algorithm for Brest cancer analysis. They use three different algorithms with help of the Waikato environment knowledge analysis open source software. They use a Decision tree, Bayesian Network and K-Nearest Neighbor algorithms in the result they have used different parameter correctly classified instance, incorrectly classified instance, incorrectly classified instances. Time taken kappa statistic, relative absolute error, and root relative squared error display by mining process [12]. The author perform comparative study of well establish and gene expiration programming for diagnosis Brest cancer among the patients [13]. The authors of the paper focus on help a doctor in parent diagnosis and treatment planning procedures for different categories. For process they use classification & clustering algorithm and analyze comparative study of the entire mining algorithm [14]. The author of the paper focus on this type of cancer disease arising from human breast tissue cells, usually from the lobules or the inner lining of the milk ducts that provide the ducts with milk. For the process they use different classification technique apply for barest cancer. They use a machine learning technique for reduce the dimension of dataset they use a two classification technique apply for breast cancer .they use a machine learning technique use for detection of barest cancer. They use principal component analysis technique for reduce component analysis technique for reduce the dimension of data set .They use a two classification technique first MLP using Back propagation NN (MLP BPN) and second Support Vector Machine (SVM) check Accuracy, Precision, Recall, F-measure, Kappa statistic [15].The author of the paper focus on identifying genes that are more correlated with the prognosis of barest cancer. They use a K-means clustering technique for test data. The identify potential bio marks for breast cancer on base on certain attribute and check a different stage by classification technique [16].The author of the paper focus on K-means clustering technique and fuzzy clustering algorithm to optimize the barest cancer [17] .The other of this paper use AI tool and neural network method for diagnosis of barest cancer among woman [18].

3. LITERATURE REVIEW FINDING

Many researcher talks about the main factor which is responsible for breast cancer such as idleness of cleaning, size, shape, color. Many patients have suffering for this problem due to not wearing proper corset. The major issues of younger are not wearing the corset in ruler area. There is a change in the appearance of nipple by touching breast. There are some patients who have a problem of bump on nipples or streaked with blood from the nipple. Patients are suffering from the thick tissue either in breast or in nipple. There are some major rashes on around a nipple. The prime symptom is found in dimpling on the breast. It may be left breast or right breast or both. The first symptom of breast cancer most women notice is a lump or an area of thickened tissue in their breast. Most lumps are not dangerous for the disease of breast cancer. The age factor is more responsible for the breast cancer. By age group, breast cancer is diagnosed in 4 out of 1,000 women in their age of 30 , 14 out of 1,000 women in their age of 40, 26 out of 1,000 women in their age of 50, 37 out of 1,000 women in their age of 60 [19]. To reduce the breast cancer more and more clean water should be drink. Most of the researchers say to reduce the caloric food for reducing breast cancer. Older age, Menstruating at an early age, Older age at first birth or never having given birth, A personal history of breast cancer or benign (non cancer) breast disease, A mother or sister with breast cancer, Treatment with radiation therapy to the breast/chest, Breast tissue that is dense on a mammogram, Taking hormones such as estrogen and progesterone.

Drinking alcoholic beverages, being white for responsible to the breast cancer.

4. EXPERIMENTAL WORK

This research takes the data from UCI for the purpose of to solve the research objective. To perform experimental work this research take WEKA as an open source data mining tool and then apply different data mining algorithm for measure accuracy and performance for the detection of the breast cancer [2].

The dataset attributes descriptions are as under in Table 1:

Table 1: Brest Cancer Dataset Attribute

Attribute Name	Description
Age	Patient’s Age in years
Menopause	the period in a woman's life when menstruation ceases
Tumor-size	Patient’s tumor-size on her breast
inv-nodes	Node size in main portion of the breast.
Node-caps	Node is present or not in cap of the breast
Deg-malig	Stage of breast cancer
Brest	Left breast or Right breast or both breast
Breast-quad	Portion of the breast for example left-up, left-low, right-up, right-low, central.
Irradiate	Present or not (YES/NO)
Class	no-recurrence-events, recurrence-events (Reduce the risk of breast cancer)

Here display the class of the breast cancer dataset in tabulated as under.

Table 2: Brest Cancer Dataset Class

Class name	Description
Diagnosis	sick, healthy or (unpredictable) no class

4.1 Implementation using FF (Farthest First) Method

Farthest First clustering algorithm performs fast analysis rather than other clustering technique. Farthest First algorithm is an option of K means clustering algorithm that seats each cluster center in turn at the peak extreme from the presented cluster centers. This peak must relax contained by the data part because of lesser amount of relocation and modification [1].

This research work used FF algorithm for diagnosis patient’s health. The model evaluations of clustered instance are tabulated in table3.
All *(286) = healthy (219)+sick(67)

Table 3: Cluster Instance

Cluster no	Instances	Percentage	Diagnosis
0	219	77	Healthy
1	67	23	Sick

This FF technique resulted in 2 clusters comprising of healthy and sick diagnosis for breast cancer patient.

The graphical presentation of the outcome is presented in figure 1.

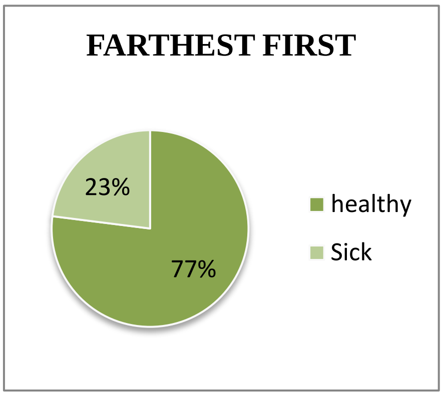

Figure 1: FARTHEST FIRST

4.2 Implementation using EM Method

The EM (Expectation Maximization) algorithm is a one type of clustering data mining technique. It tracks an iterative loom, sub-optimal, which seeks to get the constraints of the probability distribution that can say maximum probability of its characteristics. EM Clustering is basically model based clusters which is nothing but the abstraction of the k-means clustering data mining algorithm [1].

The experimental work by using EM algorithm generate following outcome.
All *(286) = healthy(117)+sick(65)+ No Class(104)

This EM Technique resulted in 3 clusters of NO class, Healthy, and Sick patients which are tabulated as under.

Table 4: EM Cluster Instances

Cluster no	Instances	Percentage	Diagnosis
0	104	36	No Class
1	117	41	Healthy
2	65	23	Sick

The graphical presentation of the outcome is presented in figure 2.

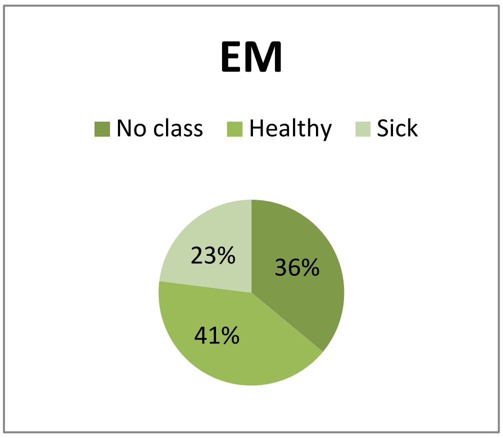

Figure 2: EM

4.3 IMPLEMENTATION USING HCM (HIERARCHICAL CLUSTERER METHOD)

This research work used HCM algorithm for diagnosis the health of the patient using WEKA environment. In this algorithm clusters are created from previously initiated clusters. So this research apply HCM algorithm.
By applying this algorithm the following outcome generated.
All *(286) = healthy (285)+sick(1)

Table 5: HCM Clustering Outcome

Cluster no	Instances	Percentage	Diagnosis
0	285	99.65	Healthy
1	1	0.35	Sick

This HCM technique resulted in 2 clusters comprising of healthy and sick diagnosis for breast cancer among the patient.

The graphical presentation of the outcome is presented in figure 3.

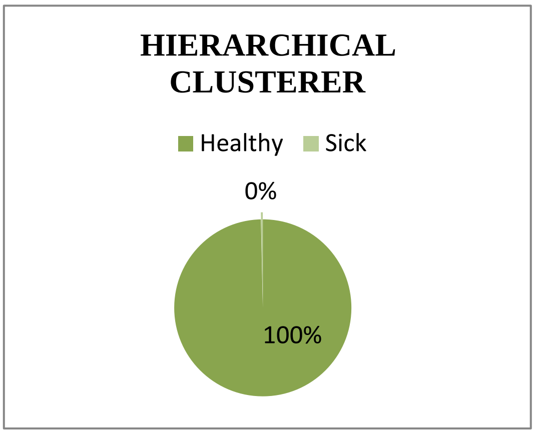

Figure 3: Hierarchical Clusterer

4.4 IMPLEMENTATION USING SIMPLE K-MEANS ALGORITHM:

This research work used k-means algorithm for diagnosis the health of the patient using WEKA environment. K-means clustering works very simple which is explain as under.

1. To divide clusters, items or objects into groups of subsets or sub-subsets.
2. To obtain central or mean values of the clusters.
3. To allocate each cluster, item or object to a appropriate cluster with the nearest mean points.
4. To obtain the central values of the clusters Repeat step 2 to step 4 till the alteration has been achieved.

The Cluster central values are tabulated in Table 6.

By applying this algorithm the following outcome generated.
All *(286) = healthy (236)+sick(50)

Table 6: K-means Clustering Outcome

Cluster no	Instances	Percentage	Diagnosis
0	236	83	Healthy
1	50	17	Sick

The graphical presentation of the outcome is presented in figure 4.

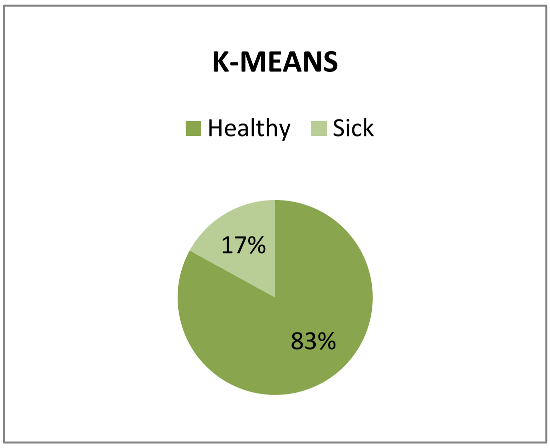

Figure 4: K-MEANS

5. RESULT ANALYSIS

The Comparative predictive results of barest cancer of the 286 patients are tabulated as under.

Table 7: Result Analysis of Breast Cancer Patient

Clustering Technique	Healthy	Sick	NO Class
FARTHEST FIRST	77	23	-
EXPECTATION MAXIMIZATION	41	23	36
HIERARCHICAL CLUSTER METHOD	99.65	0.35	-
K-MEANS	83	17	-

With the help of WEKA open source data mining tool, Breast Cancer dataset is passed through special clustering Data Mining algorithm which is Farthest First data mining algorithm, Expectation Maximization Data Mining algorithm, Hierarchical Clustering Method and lastly Simple K-Means clustering Data Mining algorithm.
In above section this research discuss about different attribute who gives optimum accuracy and performance for reduce the risk factor for breast cancer of patients. Attributes are age, menopause, tumor size, inv nodes, node caps, deg malig, breast, breast quad, irradiate.

After analysis of table 7 this work found k-means algorithm give more optimum outcome. FF algorithm also near to the k-means algorithm. So this research can analyze that 80% patient are healthy and 20% patient are sick.

6. CONCLUSION AND FUTURE WORK

The outcome of the research is justified that k-means clustering algorithm and FF algorithm are helpful to early diagnosis of the breast cancer patients. In HCM algorithm this experiment found high error rate, so it is not convenient for us. In EM technique research cannot able to diagnosis 36% patients.

This research uses four different clustering algorithms. In future this work is extending by applying different classification and association mining algorithm. In this work WEKA open source data mining tool is used for the purpose of the experiment. In future Orange, Tavera, Rapid Miner and other data mining tool and comparison study of their result give more optimum outcome.

7. REFERENCES

[1] Rinal Doshi, "DEVELOPMENT OF PATTERN KNOWLEDGE DISCOVERYFRAMEWORK USING CLUSTERING DATA MINING ALGORITHM", International journal of computer engineering & Technology (IJCET), ISSN 0976 – 6367(Print), ISSN 0976 – 6375(Online), Volume 4, Issue 3, May-June (2013), pp. 101-112

[2] WEKA, "The University of Waikato", machine learning group, weka documentation.

[3] McCready T1, Littlewood D, Jenkinson J, "Breast self-examination and breast awareness: a literature review" access from http://www.ncbi.nlm.nih.gov/pubmed/15840071 on 16/6/2014

[4] Gauthier, E. Inst. Mines-Telecom, Telecom Bretagne, Brest, France Brisson, L. ; Lenca, P. ; Clavel-Chapelon, F. ; Ragusa, S. "Challenges to building a platform for a breast cancer risk score:a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[5] Gauthier, E. Inst. Mines-Telecom, Telecom Bretagne, Brest, France Brisson, L. ; Lenca, P. ; Clavel-Chapelon, F. ; Ragusa, S."Caffeine Intake, Race, and Risk of Invasive Breast Cancer Lessons Learned from Data Mining a Clinical Database a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[6] XiangchunXiong Comput. & Inf. Sci., Towson Univ., MD, USA Yangon Kim ; YuncheolBaek ; Dae Wong Rhee ; Soo-Hong Kim"Analysis of breast cancer using data mining & statistical techniques genetic data a literature review"access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[7] Mansour, Nashat ; Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon ; Zantout, Rouba ; El-Sibai, Mirvat"Mining

breast cancer genetic data a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[8] Abdelaal, M.M.A. ; Stat. & Math. Dept., Ain Shams Univ., Cairo, Egypt ; Farouq, M.W. ; Sena, H.A. ; Salem, A.-B.M. "Using data mining for assessing diagnosis of breast cancer a literature review"access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[9] Qi Fan ; Dept. of Comput. Sci., Huaibei Coal Ind. Teacher Coll., Huaibei, China ; Chang-Jie Zhu ; Liu Yin"Predicting breast cancer recurrence using data mining techniques a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[10] Malpani, R. ; Comput. Sci. Dept., California State Univ., Sacramento, CA, USA ; Lu, M. ; Du Zhang ; Wing Kin Sung"Mining transcriptional association rules from breast cancer profile data a literature review"access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[11] Giarratana, G. ; Dipt.diElettron. e Inf., Politec. di Milano, Milan, Italy ; Pizzera, M. ; Masseroli, M. ; Medico, E. "Data Mining Techniques for the Identification of Genes with Expression Levels Related to Breast Cancer Prognosis a literature review"access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[12] Shah, C. ; Inf. Technol. Dept., ShankersinhVaghelaBapu Inst. of Technol., Gandhinagar, India ; Jivani, A.G."Comparison of data mining classification algorithms for breast cancer prediction a literature review"access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[13] Menolascina, F. ; Clinical & Exp. Oncology Lab., National Cancer Inst., Bari ; Tommasi, S. ; Paradiso, A. ; Cortellino, M. "Novel Data Mining Techniques in aCGH based Breast Cancer Subtypes Profiling: the Biological Perspective a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[14] Palivela, H. ; Dept.Of Comput. Sci. & Eng., East West Inst. of Technol., Bangalore, India ; Yogish, H.K. ; Vijaykumar, S. ; Patil, K."Survey on mining techniques for breast cancer related data a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[15] Shah, C. , ShankersinhVaghelaBapu Inst. of Technol., Gandhinagar, India ; Jivani, A.G.5"A comparative study of breast cancer detection based on SVM and MLP BPN classifier a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[16] Radha, R. ; Dept. of Comput. Sci., S.D.N.B. Vaishnave Coll. of Women, Chennai, India ; Rajendiran, P."Using K-Means Clustering Technique to Study of Breast Cancer a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[17] Vanisri, D. ; Kongu Eng. Coll., Erode, India ; Loganathan, C."Fuzzy pattern cluster scheme for breast cancer datasets a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[18] Voth, D. "Using AI to detect breast cancer a literature review" access from http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Brest+cancer+papre+in+data+mining on 17/6/2014

[19] National Cancer Institute (2006). Probability of breast cancer in American women. National Cancer Institute Fact Sheet. Available online: http://www.cancer.gov/cancertopics/factsheet/Detection/probability-breast-cancer.