ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

Diagnosing Benign and Malignant Breast Cancer Using Data Mining Classification

A.Suganya[1], K.K.Kavitha[2]

Research Scholar, Dept. of Computer Science, Selvamm Arts & Science College (Autonomous), Tamilnadu, India [1]

HOD & Vice Principal, Dept. of Computer Science, Selvamm Arts & Science College (Autonomous), Tamilnadu, India [2]

ABSTRACT: Breast Cancer Diagnosis and Prognosis are two restorative applications represent an awesome test to the analysts. The utilization of machine learning and information mining procedures has upset the entire procedure of bosom tumor Diagnosis. Breast Cancer Diagnosis recognizes favorable from harmful bosom knots and Breast Cancer Prognosis predicts when Breast Cancer is probably going to repeat in patients that have had their malignancies extracted. Consequently, these issues are chiefly in the extent of the order issues. This study paper compresses different survey and specialized articles on breast malignancy conclusion. This paper introduce a review of the ebb and flow research being done utilizing the information mining procedures to upgrade the breast disease determination.

KEYWORD: Breast cancer; Diagnosis; Prognosis; Data Mining; Classification.

I. INTRODUCTION

A huge number of grandmas, moms, little girls succumb to bosom growth consistently. The human body involves a great many cells each with its own novel capacity. At the point when there is unregulated development of any of these cells it is named as tumor. In this, cells partition and become wildly, shaping an irregular mass of tissue called as tumor. Tumor cells develop what's more, attack stomach related, apprehensive and ci rculatory frameworks disturbing the bodies' ordinary working. In spite of the fact that each single tumor is not carcinogenic. Disease is arranged by the kind of cell that is influenced and more than 200 sorts of diseases are known. This paper is concentrated on Breast disease. Bosom disease is the most well-known sort of malignancy among females over the world [2]. Later a long time have seen an exceptional change in survival rates for ladies with bosom tumor, which can be primarily ascribed to a broad screening and upgraded treatment. The late advances in information accumulation and capacity methods have made it workable for different restorative organizations and healing facilities to keep immeasurable measures of information identifying with their restorative records relating to pharmaceutical and side effects of an illness. Formally, information mining is the way toward running effective calculations on information to remove helpful data. The utilizations and possibilities of these techniques have discovered its extension in medicinal information. Foreseeing result of an infection is a testing errand. Information mining systems has a tendency to rearrange the forecast fragment. Mechanized devices have made it conceivable to gather vast volumes of restorative information, which are made accessible to the restorative research bunches. The outcomes being an expanding ubiquity of information mining methods to identify designs and relationship among expansive number of factors, which make it conceivable to anticipate the result of the illness utilizing pre-existential datasets. This paper exhibits the potential cooperative energies between information mining systems and bosom malignancy diagnosis.

II. LITERATURE SURVEY

Orlando Anunciacao et al.[7] explored the pertinence of call trees for detection of high risk breast cancer teams over the dataset created by Department of biological science of college of Medical Sciences of Universidade star Delaware port with 164 controls and ninety four cases in rail machine learning tool. To statistically validate the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

association found, permutation tests were used. They found a speculative carcinoma cluster composed of thirteen cases and only one management, with a Fisher precise Test(for validation) worth of nine.7*10-6 and a p-value of zero.017. These results showed that it's doable to search out statistically important associations with carcinoma by deriving a call tree and choosing the most effective leaf.

A. Soltani Sarvestani et al.[8] provided a comparison among the capabilities of assorted neural networks like Multilayer Perceptron (MLP), Self Organizing Map(SOM), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN) that square measure wont to classify leucocyte and NHBCD data. The performance of those neural network structures was investigated for breast cancer diagnosing downside. RBF and PNN were evidenced because the best classifiers within the coaching set. But the PNN gave the most effective classification accuracy once the take a look at set is taken into account. This work showed that applied mathematics neural networks may be effectively used for carcinoma diagnosing as by applying many neural network structures a diagnostic system was made that performed quite well.

Dr. Medhat Mohamed Ahmed Abdelaal et al.[9] investigated the aptitude of the classification SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass options along side age that discriminates true and false cases. Here, SVM techniques show promising results for increasing diagnostic accuracy of classifying the cases witnessed by the most important space beneath the mythical creature curve adore values for tree boost and tree forest.

K. Rajiv Gandhi et al.[10] made classification rules victimisation the Particle Swarm Optimization algorithmic program for carcinoma datasets. during this study to deal with serious computational efforts, the matter of feature set choice as a pre-processing step was used which learns fuzzy rules bases victimisation GA implementing the Pittsburgh approach. it had been wont to produce a smaller fuzzy rule bases system with higher accuracy . The resulted datasets once feature choice were used for classification victimisation particle swarm optimisation algorithmic program. The rules developed were with rate of accuracy shaping the underlying attributes effectively. Manaswini

Pradhan et al.[11] urged a synthetic Neural Network (ANN) primarily based classification model mutually of the powerful technique in intelligent field for classifying diabetic patients. The neural network, employed in back propagation algorithmic program, is m-n-1 kind network. The GA is employed for optimally searching for the quantity of neurons within the single hidden bedded model. For coaching and testing 10-fold cross validation technique was adopted for Pima Indian polygenic disease. For Pima dataset the ANN provides the most effective accuracy with five neurons within the hidden layer. Best accuracy being seventy two with average accuracy of seventy two.2%. The designed model was compared with the purposeful Link ANN (FLANN) and several other classification systems like NN (nearest neighbor), kNN(k-nearest neighbor), BSS(nearest neighbor with backward sequent choice of feature, MFS1(multiple feature subset) , MFS2(multiple feature subset) for knowledge classification accuracies. it had been discovered from the simulation that their urged model performed higher than compared to all or any of the collaborating techniques for comparison.

III. EXISTING SYSTEM

Data mining techniques are often accustomed predict cancer during a patient exploitation numerous symptoms knowledge from previous results. Valuable data are often discovered through this data processing techniques. Association Rule Mining (ARM) and Classification for designation and prognosis of cancer is employed. Below ARM FP growth algorithmic program that is applied on attributes of patient knowledge to predict benign or malignant is employed. Call tree are often accustomed categoryify associate degree unknown class knowledge instance.

IV. PROPOSED SYSTEM

The data mining consists of varied ways. totally ways serve different functions, every technique providing its own benefits and drawbacks. However, most data processing ways usually used for this review are of classification class because the applied prediction techniques assign patients to either a ”benign” cluster that's non- cancerous or a ”malignant” cluster that's cancerous and generate rules for an equivalent. Hence, the breast cancer diagnostic issues are essentially within the scope of the wide mentioned classification issues. In data processing, classification is one in every of the foremost vital task. It maps the information in to predefined targets. It is a supervised learning as targets ar

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

predefined. The aim of the classification is to create a classifier supported some cases with some attributes to explain the objects or one attribute to explain the cluster of the objects. Then, the classifier is employed to predict the cluster attributes of recent cases from the domain supported the values of different attributes. The usually used ways for data processing classification tasks may be classified into the subsequent groups

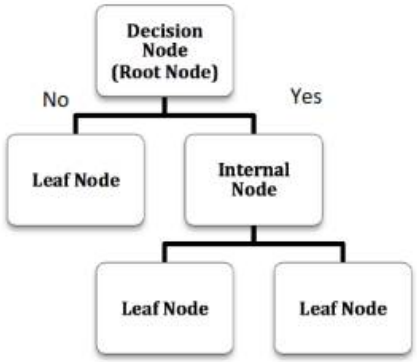

Fig 1: Decision tree architecture

V. IMPLEMENTATION PRE-PROCESSING

In pre-processing step we have a tendency to improve the standard of the image by removing noise. The pre-process is vital to eliminate some components that aren't connecting to the region of interest. So, the second step of pre-processing is removing the background space, removing the pectoral and rib portion from the breast space.

FEATURE EXTRACTION

Feature Extraction is most vital step, uses for extracting price of different-different options and plays a key role in pattern classification. it's a way of capturing visual content of pictures for compartmentalisation & retrieval. Image options may be either general options, like extraction of color, texture and form or domain specific options. X-ray photograph pictures contains heterogeneous info that shows differing kinds of tissues, blood vessels, organ ducts, breast edges and X-ray photograph machine characteristics. thus to possess higher approach for sleuthing traditional and abnormal tissues, we've to decide on such a system which provides higher result with a lot of accuracy. Here we have a tendency to area unit victimisation data processing technique for extracting the textural options.

BREAST CANCER IDENTIFICATION

Clinical identification of carcinoma helps in predicting the malignant cases. A lump felt throughout the examination roughly offer clues on the scale of growth and its texture. the varied common ways used for carcinoma identification area unit diagnostic procedure, Biopsy, antielectron Emission pictorial representation and resonance Imaging. The results obtained from these ways area unit wont to recognise the patterns that area unit reaching to facilitate the doctors for classifying the malignant and benign cases. There area unit varied data processing techniques, applied mathematics ways and machine learning algorithms that area unit applied for this purpose.

BREAST CANCER PROGNOSIS

Prognosis is vital as a result of the kind and intensity of the medications area unit supported it. Prognosis drawback is additionally referred to as as “analysis of survival or life data”. It poses a tougher drawback than that of identification since the info is expurgated. That is, there area unit solely a couple of cases wherever we've AN ascertained repeat of the sickness. during this case, we will classify the patient as recur and that we recognize the time to recur (TTR). On the opposite hand, we have a tendency to don't observe repeat in most patients. For these, there's no real purpose at that we will think about the patient a non continual case. So, the info is taken into account expurgated since we have a tendency to don't recognize the time of repeat. For such patients, all notable is simply the time of their

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

last check-up. we have a tendency to decision this the disease-free survival time (DFS). Prognosis helps in establishing a treatment set up by predicting the result of a sickness. There area unit 3 prophetical foci of cancer prognosis: 1) prediction of cancer status (risk assessment), 2) prediction of cancer repeat and 3) prediction of cancer survivability. the target of prognostic predictions is to handle cases that cancer has not recurred (censored data) further as case that cancer has recurred at a particular time. Thus, carcinoma prognostic issues area unit primarily within the scope of the wide mentioned classification issues. This section consists of the review of assorted technical and review articles on data processing techniques applied in carcinoma prognosis.

CLASSIFICATION

The info mining consists of assorted ways. Totally ways serve different functions, every technique providing its own benefits and drawbacks. However, most data processing ways usually used for this review area unit of classification class because the applied prediction techniques assign patients to either a ”benign” cluster that's non-cancerous or a ”malignant” cluster that's cancerous and generate rules for constant. Hence, the carcinoma diagnostic issues area unit essentially within the scope of the wide mentioned classification issues. In data processing, classification is one amongst the foremost necessary task. It maps the info in to predefined targets. it's a supervised learning as targets area unit predefined. The aim of the classification is to make a classifier supported some cases with some attributes to explain the objects or one attribute to explain the cluster of the objects. Then, the classifier is employed to predict the cluster attributes of latest cases from the domain supported the values of alternative attributes.

VI. RESULT AND DISCUSSION

The basic phenomenon used to classify the Brest Cancer disease classification using classifier is its performance and accuracy. The performance of a chosen classifier is validated based on error rate and computation time. The classification accuracy is predicted in terms of Sensitivity and Specificity. The computation time is noted for each classifier is taken in to account. Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model.

After the preprocessing step, a common analysis would be determining the effect of the attributes on the prediction, or attribute selection. We used the information gain measure to rank the attributes due to the fact that it is a common method and the C4.5 decision tree technique utilizes this measure. Information gain (IG) is measured as the amount of the entropy (H) difference when an attribute contributes the additional information about the class. The following is the information gain and the entropy before and after observing the attribute Xi for the class. In this study, the accuracy of three data mining techniques is compared. The goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics such as specificity and sensitivity. These metrics can be derived from the confusion Ranked Survivability Attributes Extention of tumor (EOD) Stage of cancer Lymph node involve (EOD) Site Specific Surgery No of pos nodes (EOD) Tumor size (EOD) Hostologic type Age Behavior code Number of nodes (EOD) Grade Marital status Primary site Radiation Race Number of primaries 4 matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

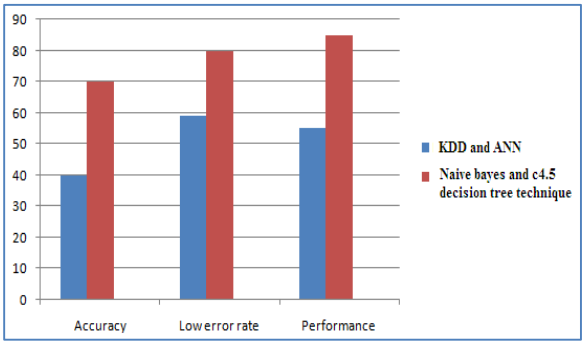

Fig 2: Performance of Computing Time

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

VII. CONCLUSION

This paper provides a study of varied technical and review papers on carcinoma diagnosing and prognosis issues and explores that data processing techniques provide nice promise to uncover patterns hidden within the knowledge that may facilitate the clinicians in deciding. From the on top of study it's determined that the accuracy for the diagnosing analysis of varied applied data processing classification techniques is very acceptable and may facilitate the medical professionals in deciding for early diagnosing and to avoid diagnostic test. The prognostic downside is principally analyzed beneath ANNs and its accuracy came higher compared to alternative classification techniques applied for a similar. However additional economical models can even be provided for prognosis downside like by inheritable the most effective options of outlined models. In each cases we are able to say that the most effective model is obtained when building many differing types of models, or by attempting totally different technologies and algorithms.

REFERENCES

[1] El-Sebakhy A. Emad, Faisal Abed Kanaan, Helmy T., Azzedin F. and Al-Suhaim F., "Evaluation of breast cancer tumor classification with unconstrained functional networks classifier," Computer Systems and Applications, IEEE International Conference, 2006, pp. 281 – 287.
[2] Osmar R. Zaïane, Principles of Knowledge Discovery in Databases. [Online]. Available: webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf
[3] Han J. and Kamber M., Data Mining: Concepts and Techniques, 2 nd ed., San Francisco, Morgan Kauffmann Publishers, 2001
[4] Sarvestan Soltani A. , Safavi A. A., Parandeh M. N. and Salehi M., "Predicting Breast Cancer Survivability using data mining techniques," Software Technology and Engineering (ICSTE), 2nd International Conference, 2010, vol.2, pp.227-231.
[5] Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo and Rueff Jose, "A Data Mining approach for detection of high-risk Breast Cancer groups," Advances in Soft Computing, vol. 74, pp. 43-51, 2010.
[6] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed, "Using data mining for assessing diagnosis of breast cnacer," in Proc. International multiconfrence on computer science and information Technology, 2010, pp. 11-17.
[7] Chang Pin Wei and Liou Ming Der, "Comparision of three Data Mining techniques with Ginetic Algorithm in analysis of Breast Cancer data".[Online]. Available:http://www.ym.edu.tw/~dmliou/Paper/compar_threedata.pdf
[8] Gandhi Rajiv K., Karnan Marcus and Kannan S., "Classification rule construction using particle swarm optimization algorithm for breast cancer datasets," Signal Acquisition and Processing. ICSAP, International Conference, 2010, pp. 233 – 237.
[9] Padmavati J., "A Comparative study on Breast Cancer Prediction Using RBF and MLP," International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011.
[10] Lee Heui Chul, Seo Hak Seon and Choi Chul Sang, "Rule discovery using hierarchial classification structure with rough sets," IFSA World Congress and 20th NAFIPS International Conference, 2001, vol.1 , pp. 447-452.
[11] Hassanien Ella Aboul and Ali H.M. Jafar, "Rough set approach for generation of classsification rules of Breast cnacer data," Journal Informatica, 2004, vol. 15, pp. 23–38.
[12] Sudhir D., Ghatol Ashok A., Pande Amol P., "Neural Network aided Breast Cancer Detection and Diagnosis",7th WSEAS International Conference on Neural Networks, 2006.
[13] Jamarani S. M. h., Behnam H. and Rezairad G. A., "Multiwavelet Based Neural Network for Breast Cancer Diagnosis", GVIP 05 Conference, 2005, pp. 19-21.
[14] Pantel Patrick **,** Breast Cancer Diagnosis and Prognosis.[Online]. Available: http://citeseer.nj.nec.com/pantel98breast.html.
[15] Choi J.P., Han T.H. and Park R.W., " A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", J Korean Soc Med Inform, 2009, pp. 49-57.
[16] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques," Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining," 2006.
[17] 'Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H. , "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.
[18] Street W.N., "A Neural Network Model for Prognostic Prediction", Fifteenth International Conference on Machine Learning, Madison, Wisconsin, Morgan Kaufmann, 1998.
[19] Chi C.L., Street W.H. and Wolberg W.H., "Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets", Annual Symposium Proceedings / AMIA Symposium, 2007.