BREAST CANCER PREDICTION USING DATA MINING TECHNIQUES

S. Padma Priya[1], P.Sowmiya[2]

[1.] Assistant Professor & Head Department of Information Technology
Sri Adi Chunchanagiri Women's College, Cumbum,(India)
[2.] Research Scholar, Department of Computer Science,
Sri Adi Chunchanagiri Women's College, Cumbum,(India)

ABSTRACT

Data mining (DM) comprises the core algorithms that enable to gain fundamental insights and knowledge from massive data. In fact, data mining is a part of a larger knowledge discovery process. One of the new researches in data mining application involves analyzing Breast cancer, which are the deadliest disease and most common of all cancers in the leading cause of cancer deaths in women worldwide. Among the various DM techniques, classification plays a vital role in DM research. Breast cancer diagnosis and prognosis are two medical applications pose a great challenge to the researchers in medical field. This survey work analyses the various review and technical articles on breast cancer diagnosis. The main goal of this research is to explore the overview of the current research being carried out using the data mining techniques to enhance the breast cancer diagnosis. Particularly, this survey discusses about use of the classification algorithms ID3 and C4.5 in breast cancer analysis.

Keywords: Breast Cancer Analysis, Classification Algorithms, C4.5 Algorithm, Decision tree, ID3 Algorithm.

I. INTRODUCTION

Data mining has become a fundamental methodology for computing applications in the domain area of medicine. Evolution of data mining applications and its implications are manifested in the areas of data management in healthcare administrations, epidemiology, patient care and intensive care systems, significant image analysis to information extraction and automatic identification of unknown subjects. In the recent years, the data from several domains including banking, retail, telecommunications and medical diagnostics contains valuable information and knowledge which is often hidden.DM has various techniques such as Classification, Clustering, Prediction, Association Rules, Decisions Tress, and Neural Networks. Among the various classification algorithms, the very famous algorithms ID3 and C4.5 plays an essential role in breast cancer analysis. Breast cancer has become the primary reason of death in women in developed countries. Numerous researchers have attempted to apply machine learning algorithms for detecting survivability of cancers in human beings and also it has been proved. This paper survived about the various research works carried out using ID3 and C4.5 algorithms, done by different researchers. This will identify to predict general and individual performance of patient. The remaining of this paper is organized as follows. Section II discusses about the data mining techniques, which are used for breast cancer analysis. Section III discusses about literature review of breast cancer. Finally section IV concludes the survey work.

II. DATA MINING TECHNIQUES FOR BREAST CANCER ANALYSIS

Data mining is a powerful and a new field having various techniques to analyses the recent real world problems. It converts the raw data into useful information in various research fields and finds the patterns to decide future trends in medical field. There are various major data mining techniques that have been developed and used in data mining projects recently for knowledge discovery from database [1]. Breast Cancer is the leading cause of death in women in developing countries and a second cause in developed countries as per the statistics of national cancer institute. The breast cancer can occur in both male and female. But the occurrence is high in female throughout the world. Breast cancer is most frequently discovered as an asymptomatic nodule on a mammogram.

III.CLASSIFICATION ALGORITHMS

Classification is a data mining function. The goal of classification is to accurately predict the target class for each case in the data. Classification is used to classify the data items into groups. Derived model can be presented as classification or rules [2]. Many researchers have been applying various algorithms to help health care professionals with improved accuracy in the diagnosis of breast cancer. Divya Tomar andSonali Agarwal were carried out a survey on data mining approaches for healthcare diagnosis [3]. Clustering is an unsupervised learning method is different from classification. It is mainly used for analyzing microarray data because very little details are available for genes. This paper was analyzed the gene expression data with the help of a new hierarchical clustering approach using genetic algorithm comparison of three data mining methods for predicting breast cancer survivability is discussed by Delen et al. in their research paper [4]. They used two popular data mining algorithms namely artificial neural networks and decision trees along with a most commonly used statistical method (logistic regression) to develop the prediction models using a large dataset (more than 200,000 cases). They also used 10-foldcross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The survey of more prediction models for breast cancer continuity using a large dataset along with a 10-fold cross-validation provided us with an insight into the relative prediction ability of different data mining methods. A research paper by Belciug, Smaranda and Florin Gorunescu is discussed about the hybrid neural network genetic algorithm applied to breast cancer detection and recurrence [5].

IV.LITERATURE REVIEW

C4.5 is a notable choice tree acceptance learning system which has been utilized by AbdelghaniBellaachia and ErhanGauven alongside two different strategies i.e. naive Bayes and BackPropagated Neural Network. They exhibited an examination of the expectation of survivability rate of breast growth patients utilizing above information mining systems and utilized the new form of the SEER Breast Cancer Data. The preprocessed information set comprises of 151,886 records, which have all the accessible 16 fields from the SEER database. They have received an alternate approach in the pre-grouping process by including three fields: STR (Sarvival Time Recode), VSR (Vital Status Recode), and COD (Cause Of Death) and utilized the Weka toolbox to explore different avenues regarding these three information mining calculations. A few investigations were led utilizing these calculations. The accomplished forecast exhibitions are practically identical to existing procedures. In any case, they discovered that model created by C4.5 calculation for the given information has a vastly improved execution than the other two methods. [6] Wei-stick Chang, Der-Ming and Liou investigated that the hereditary calculation display yielded preferable results over other information digging models for the examination of the information of breast tumor patients as far as the general exactness of the patient arrangement, the expression and many-sided quality of the classification rule. The artificial neural system, decision tree, calculated relapse, and hereditary calculation were utilized for the similar studies and the precision and positive prescient estimation of every calculation were utilized as the assessment pointers. WBC database was joined for the information investigation took after by the 10 -overlap cross-approval. The outcomes demonstrated that the hereditary algorithm portrayed in the study could deliver exact results in the order of breast disease information and the arrangement run recognized was more adequate and intelligible. [7] Labeed K Abdulgafoor et al wavelet change and K-means clustering calculation have been utilized for intensity based segmentation. [4] Sahar A. Mokhtar et al have examined three diverse classification models for the forecast of the seriousness of breast masses in particular the choice tree, simulated neural system and bolster vector machine.[8] Rajashree Dash et al a hybridized K-mean algorithm has been proposed which consolidates the means of dimensionality decrease through PCA, a novel introduction approach of group focuses and the means of doling out information focuses to proper clusters. [9] Ritu Chauhan et al concentrates on clustering algorithm,, for example, HAC and K-Means in which, HAC is connected on K-means to decide the quantity of bunches. The nature of bunch is enhanced, if HAC is connected on K-means.

V.CONCLUSION

This research work addresses various techniques and review on breast cancer diagnosis and prognosis problems. The prognostic problem is mainly analyzed under C4.5 and its accuracy came higher in comparison to other classification techniques applied for the same. From the various researchers' perspective, it is identified that the causes and the symptoms related to each event will be made in accordance with the record related to each patient and thereby breast cancer can be reduced to a great extent. The classification algorithms ID3and C4.5 are used to identify the various categories of breast cancer. The behavior and performance of both the algorithms are analyzed via its experimental results by many researchers. Applying their own methods and data sets by researchers, they suggest the best method for analyzing breast cancer. From their point of view, this research work concluded that the performance of C4.5 is better than the other algorithms.

REFERENCES

[1] Aarti Sharma, Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava, “Application of Data Mining A Survey Paper”, *Int. Journal of Computer Science and Information Technologies, Vol. 5, Issue 2,2014, pp. 2023-2025*.

[2] Sivagami. P, “Supervised Learning Approach for Breast Cancer Classification”, *Int. Journal of Emerging Trends & Technology in Computer Science, Vol. 1, Issue 4, 2012, pp. 115-129.*

[3]Divya Tomar and Sonali Agarwal, “A survey on Data Mining approaches for Healthcare”, *Int. Journal of Bio-Science and Bio-Technology, 2013, Vol.5, Issues 5, pp. 241-266.*

[4] Delen D, Walker G, and Kadam A., “Predicting breast cancer survivability a comparison of three data mining methods”, Artificialintelligence in medicine, 2005, Vol. 34, Issue 2, pp. 113-27.

[5] 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.

[6] Chang Pin Wei and Liou Ming Der, “Comparision of three Data Mining techniques with Ginetic Algorithm in analysis of Breast Cancer data”.

[7] Labeed K Abdulgafoor “Detection of Brain Tumor using Modified K-Means Algorithm and SVM” *International Journal of Computer Applications (0975 – 8887) National Conference on Recent Trends in Computer Applications NCRTCA 2013.*

[8] A. Sahar “Predicting the Serverity of Breast Masses with Data Mining Methods” *International Journal of Computer Science Issues, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694 -0814| ISSN (Online):1694-0784 www.IJCSI.org.*

[9] Rajashree Dash “A hybridized K-means clustering approach for high dimensional dataset” *International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66.*