A survey on breast cancer analysis using data mining techniques

Manjuladevi R [1], Saranya C P [2]

Coimbatore Institute of Engineering and Technology,manjurajkumar28@gmail.com

[1]Student, Department of CSE,

Coimbatore Institute of Engineering and Technology, sarancp@gmail.com

[2]Assistant Professor, Department of CSE,

ABSTRACT:

Cancer is the second most deadliest disease among different disease with the estimation of 7.9 million death in 2007. This may rise and reach 12 million deaths in 2030 and become the major public health issue. There are 100 many types of cancer, depends on the cells that are initially affected the cancer can be categorized. Breast cancer is one of the leading cause of cancer led to death in women worldwide. The rate of breast cancer is all-time high. The intelligent systems are implemented to detect and diagnose such a type of cancer. Nowadays data mining are used in health care industry. Mining techniques are applied to analyze the breast cancer. Classification techniques plays major role in data mining research . Classification is implemented through different algorithms. In Particular here we discussed some of the classification algorithms such decision tree (C4.5) , naive bayes, Support Vector Machine, K Nearest Neighbor, Random Forest . In this survey the enhancement of decision tree algorithm for breast cancer analysis and predict using data mining techniques is involved to overcome the drawback of existing decision tree algorithm.

KEYWORDS

Data mining, Classification, Decision tree , naïve bayes, analysis of breast cancer,C4.5.

1.INTRODUCTION

Mining is a process of extracting the useful information stored in the large database and also discovering new patterns to decide future trends in medical field. There are various data mining techniques are used in various data mining projects for knowledge discovery from database. Through this discovery of knowledge from different perspective the summary of useful information is obtained. Data mining involves several functions such as association rules, concept description, clustering classification, prediction and sequence discovering to acquire useful patterns. As per the report of cancer institute the leading cause of death in women by the breast cancer [2]. Among different cancer breast cancer is most important cancer that affecting females 25% among 1.67 million cases of cancer in 2017 [1]. When compared to United State (95 per 100000) with India (28.8 rate 100000), India rate lower in breast cancer. Breast cancer is not only faced by women worldwide but men too. Through the cells of breast the formation of breast cancer takes place. After skin cancer breast cancer is the most common cancer diagnosed in women at united state. The most common types of cancer is also called as tumor. Nowadays breast cancer is the major health problem faced by women significantly. Detection of breast cancer in the earlier stage reduces the life losses. Through this detection the breast cancer can be cured at the earlier stage. Accurate and reliable diagnosis procedure at the earlier stage slows the physician to distinguish benign breast tumor from malignant ones. These are two different tumors which are diagnosed so far [4]. Benign tumor is not dangerous but it gradually affect the tissues. Malignant tumor which cause hazard to life, permeate muscles and adjacent tissues [4]. Through the classification of breast cancer the prediction of outcomes or genetic behavior of tumor is analyzed. The different test performed upon the patient to diagnosis disease. The use of classifier is massively increasing through this reason [7]. The most interesting and challenging part is prediction of outcomes where to develop a data mining application. The large volume of medical data collection and use of computers with automated tools are made available to the medical research groups. By using the historical datasets can be able to predict the outcomes of disease, exploit patterns and relationship among the large number of variables. Computer Aided Design (CAD) is used to diagnosis of breast cancer [5].

2. LITERATURE SURVEY

Dr. A. Kumarkombaiya and el at [1] have discussed Data mining based detection of breast cancer by using Support Vector Machine ,back propagation network and C5.0 decision algorithm to investigate the rate of survivability and predictability using WCB and SEER datasets. These three algorithms are also used to the prediction of survivability rate. In this survey they include STR, VSR and COD for pre classification process. Through the different investigation they came to know that C 5. 0 and SVM algorithms works better to predict breast cancer accurately.

Dr. C Nalini and el at [2] have discussed data mining based prediction of breast cancer by using Decision tree (J48) for the higher accuracy with minimum execution time. Here they used WEKA tools for implementation. Here the comparison of Naïve Bayes and Decision Tree (J48) is carried to choose better performance and accuracy while predicting breast cancer. Here the J48 follows Divide and conquer method. Through the implementation of research Naïve Bayes considered to be the best algorithm because of the accuracy in

prediction and less execution time

Eng.Ahmed Abd El Hafeez Ibrahim and el at [4] have implemented data mining classification for breast cancer diagnosis. They firstly analyzed benign or malignant tumor for diagnosis then the recurrence or no- recurrence of tumor is analyzed for prognosis. Here the four types tols are used for implementation such as WEKA, ORANGE, KNIME. Among these tools WEKA followed by ORANGE Tool gives accurate result. In diagnosis WDBC achieved higher accuracy than WBC by 0.71%. In prognosis WPBC achieved higher accuracy than using LBCD by 3.9%

Haifeng Wang and Sang Won Yoon and el at [5] have discussed breast cancer prediction using mining models. In their work, they applied mining techniques in health care to detect breast cancer. They tested eight hybrid models based on two datasets. Here the different types of mining algorithms are carried out such as SVM, Naïve Bayes, ANN, Adaboost ad K-Fold cross validation. Through that they come to know PCA which is a linear model gives more accuracy and efficiency while predicting breast cancer. Furthermore they also Tested K-means using standard datasets (SEER)

Akinsola Adniyi , Sokunbi M.A, Okikiola F.M, onadokun I.O and el at [6] have discussed mining for breast cancer classification. The datasets are portioned into several items for evaluation and analysis. Through the comparison of different algorithms performance, execution time, error occurred we came know the best model. They used C4.5 algorithm with accuracy of 93.9854% at 0.28 sec, followed by multilayer perceptron algorithm with accuracy of 83.8673% at 12.68 sec and bayes With accuracy of 76.503% at 0.03 sec. Among these analysis C4.5 gives more accuracy with efficient time.

Ishola,O.B., Ajibola, A.A. and el at [7] have worked on different mining techniques using WEKA toolkit. They used logistic Regression, Naïve Bayes, Linear SVC, SVM with linear kernel and random forest algorithms with wisconsin dataset from UCI. The ARFF file format is used by WEKA tool for data file . The different attributes are used by the datasets. The comparison of different algorithms are used above is carried out. Through the comparison The higher accuracy is achieved through Random Forest algorithm with 97.8%.

R.Preetha and **S.Vinila jinny** el at [14] have discussed data mining based research on breast cancer prediction with suitable diagnosis treatment model gives accurate and dependable prediction effects to the scientific practitioners. Many of the tools are involved for implementation they such NLTK, KNIME, WEKA and R- programming. Among these WEKA tool give more accuracy in prediction. In this different models are compared such DNN , K-means , NN, Fuzzy logic etc.., and found the most accuracy in outcome. Among the ones fuzzy logic seems to better accuracy and effective in prediction of breast cancer.

3.METHODOLOGY AND ALGORITHMS

NAIVE BAYES

Naïve Bayes classifier assumes that the presence of particular feature in a class is unrelated to the presence of any other features. Therefore attributes are independent of each other. The high dimensional datasets are classified using this classification techniques. The probability can be computed as

$$P(c|x) = P(x|c)P(c)/P(x)$$

P (c| X) is the posterior probability of class(c, target) given predictor(x, attribute)

P(c) class of prior probability

P (x| c) is the likelihood

P(x) is the prior probability of predictor

DECISION TREE

The decision tree comes under the family of supervised learning. Decision tree extract predictive information in the form of human-understandable tree-rules. This algorithm is useful for many classification problems. Decision tree is one of the classification algorithm which build tree structure data format. The whole datasets separated into individual node which is known as sub node or sub leaf. The sub nodes are actually represent the instances of the datasets. Leaf node is otherwise called as class label. In WEKA application this algorithm is implemented as J48. In This algorithm some datasets have a missing values that are calculated by this algorithm. Decision tree algorithm gives the accurate result still having the missed values. It is more reliable and robust algorithm.

SUPPORT VECTOR MACHINE

Support vector machine is simply called as SVM. This algorithm comes under the family of supervised machine learning algorithm which is used for regression and classification. Support vector machine is a line model. It solve nonlinear and linear problems and it gives accurate result to practical problems. This algorithms creates hyper plane or line which separate data into classes.

In this algorithm the datasets are plotted in n-dimensional. n represents the instances in datasets. firstly differentiation of the datasets are takes place then comparison of the datasets are done . This SVM algorithm significantly used in face detection, image recognition, bioinformatics, hand written digital recognition and many more.

K-NEAREST NEIGHBOR

KNN is used for classification and regression which is non-parametric in nature. KNN a type of instance based learning model. Based on this instance based learning it learn in testing phase but does not learn anything during training phase. In this algorithm input involve k closest training examples of the feature space. The output depends on whether this algorithm follows classification or regression. The outcome of this algorithm are in the form of 1, -1, and 0. In KNN classification, output is a class membership. In KNN regression, the output is a property of value for object. This algorithm is used in protein-protein prediction also measure the document similarities and many more. It has no assumption about data and used to nonlinear data, easily understandable, high accuracy in output, versatile. This algorithm follows the steps

- K is a positive sample which is specified along with the new sample.
- We select the closest K entries to the new sample in our database.
- We find the most classification of these entries
- We give this most classification to the new sample.

RANDOM FOREST

Random is a individual collection of decision tree which are ordered to build a forest which is basically known as a random forest. It is based on the bagging principle and random subset space method. It is independent on missing values and noise present in a input node. Each decision tree is constructed by random subset of the training data. Every node is separated by the best node among other nodes which is present In forest. The missed values does not affect the outcome. The instances are taken randomly from the datasets through the recursive approach. Random forest is a fastest, flexible and robust approach to mining high dimensional and valid data.

They are the extension of classification and regression trees. In This random forest the models have low correlation is the key. Through the low correlation between the models the prediction can be made by the individual trees.

PAPER REF NO	ALGORITHMS USED	ATTRIBUTES	DATASETS	CONCLUSION
3	Naïve Bayes, J48	Date of birth, menopause age, history of family, intake of alcohol, smoking frequency, breast feeding	From LASUTH information of patients Datasets is used	J48 is a better decision tree model.
8	Artificial neural network, Bayesian network and	Marital status, surgery, radiation, age, tumor size	SEER Dataset	The accuracy of ANN and Hybrid is almost similar and

	hybrid			better than Bayesian network
9	Naïve bayes, back propogation,NN and C4.5	Marital status, surgery, age, tumor size	SEER Dataset	C4.5 results higher accuracy of 86.7%
10	J48,CART,AD Tree, and BF Tree	Sex, age, problem faced, past history, medical history, weight, height, food and habit	Swamy Vivekananda hospital – Chennai	Among these four algorithms J48 classifier had higher accuracy.
11	Logistic regression, SVM, Random forest, Naïve bayes	Sex, age, food habit, medical details, weight and height	Wiscousin dataset	Random forest got higher accuracy of 97.8%
12	J48, Random Forest, Random Tree, REP Tree, decision tree	Sex, age, height, weight, present problem, food habit, medical	Wiscousin dataset	Priority based tree produce higher accuracy of 93.63%

	and priority based tree	diagnosis, past medical problems		
13	ANN, CART, Logistic regression, BBN	Marital status, extension of tumor, radiation, stage of cancer, age , number of positive nodes	SEER dataset	Bayesian network acquired higher accuracy.

CONCLUSION

In this survey the different data mining algorithms are studied to analysis the breast cancer and predict the tumor using datasets. The detection of breast cancer at the earlier stage helps to prevent from breast cancer. Detection of tumor at the starting stage can be cured easily. Through the Data mining algorithms the breast cancer can analyzed and predicted. Some of the data mining algorithms like Support Vector Machine, Naïve Bayes, decision tree, k-Nearest Neighbor, and Random forest are studied. Among these algorithms the random forest provide higher accuracy in detection. Depends on the comparison of above different algorithms the most accuracy will be the best model. The future work can be done by using data mining algorithms to analyze breast cancer.

REFERENCES

[1] F.Leenavinmalar and Dr.kumarkombaiya, “Application of mining techniques in detection of breast cancer”, International Journal of Engineering trends and technology (IJETT)-vol.56 No 1-Feb 2018.

[2] Dr.C.Nalini and D.Meera, “Breast cancer Prediction using Data Mining Methods”, international Journal of pure and applied Mathematics vol.119 No.12 2018, 10901-10911.

[3] K.Williams, J.A.Balogun and P.A Idowu ,” Breast cancer Risk prediction By using data mining techniques”, transaction of communication and neural networks, 3(2),01,2015.

[4] Eng.Ahmed Abd Ei hafeez Ibrahim, ”Implementation of data mining classification in breast cancer diagnosis”, http://www.researchgate.net/publication/316464904.

[5] Haifeng wang and sang won yoon, “Breast cancer prediction using Data mining methods”, https://www.researchgate.net/publication/319688741.

[6] Akinsola Adniyi , Sokunbi M.A, Okikiola F.M and onadokun I.O ,”Data mining for breast cancer classification”, International Journal of Engineering And Computer science ISSN:2319 -7242 vol.6 Issue 8 august 2017,page no. 22250-22258, index Copernicus value (2015):58.10 DOI:10.18535/ijecs/v6i8.06.

[7] Ishola,O.B., Ajibola, A.A.,” Mining techniques used for breast cancer prediction”, https.//dx.doi.org/10.26808/rs.re.v3i5.04.

[8] J.P.Choi, T.H.Han and R.W.Park, “Breast Cancer Prognosis using Hybrid Bayesian Network Model”, Journal of Korean Society Med. Informatics, 15 (1) ,pp.49-57, 2009.

[9] A.Bellaachia and E.guven, ”Surivablity of Breast Cancer Prediction using Mining Techniques”, Ninth workshop on mining with SIAM International Conference on Data Mining Age, 58(13),pp.10-110,2006.

[10] E.Venkatesan and T. Velmurugan, “Analyzing the performance of decision tree algorithm for classification of breast cancer”, International Journal of Science and Technology, 8(29), 2015.

[11] Ruolan Xu and Qiongjia Xu, “Predicting Breast Cancer risk using different Machine Learning Models”.

[12] P.Hamasagayathri and P.Sampath, “Performance analysis classification using decision tree classifiers”, International Journal curr pharm res,9(2), 19-25, 2017.

[13] E.Y.Kibis, I.E.Buyuktahtakin and Ali dag, ”Analytics approach for survivability: comparison of mining models, IISE Annual Conference, Proceeding, pp. 591-596, 2017.

[14] Sonali Agarwal, Divya Tomar , ”Data mining approaches for health” , Int. journal of Bio-Science and Bio-technology, 2013 Vol. 5, Issues 5, pp. 241-266.

[15] Schapire, R. E., and Freund, Y., 2012, Boosting ; Foundation and algorithms, MIT press, Cambridge.

A survey on breast cancer analysis using data mining techniques