IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-10, May 2019 E-ISSN: 2347-2693

Research on Naive Bayes Algorithm of Breast Cancer Diagnose Data by Machine Learning

Pushpraj Saket[1*], Anshul khurana[2]

[1,2] Department of Computer Science, Shri Ram Institute of Technology Jabalpur, RGPV, Bhopal, India

[*]*Corresponding Author: saket.pushpraj@gmail.com, Tel.:* 0761-4001933

DOI: https://doi.org/10.26438/ijcse/v7si10.149151 | Available online at: www.ijcseonline.org

Abstract— Breast cancer is one amongst the leading cancers for ladies in developed countries including Asian nation .It is the second most typical explanation for cancer death in women. The high incidence of breast cancer in women has redoubled considerably within the last years. Naïve Bayes algorithm is used for carcinoma identification Prognosis and diagnosis. Carcinoma Diagnosis is identifying of benign from malignant breast lumps and carcinoma Prognosis predicts once Breast Cancer is to recur in patients that have had their cancers excised. In this paper Naïve Bayes Algorithm is used to classify the Datasets of Breast Cancer (Diagnosis). The classification results show that when two features of maximum radius and maximum texture is selected, the classification improved accuracy is 98.6%, which is improved compared with previous method.

Keywords—Breast Cancer Dataset, NaïveBayes classification Algorithm.

I. INTRODUCTION

Early diagnosing of cancer needs a correct and reliable diagnosis procedure that permits physicians to differentiate benign breast tumors from malignant ones while not going for surgical diagnostics. The objective of those predictions is to assign patients to either a “benign” cluster that's noncancerous or a “malignant” group that's cancerous. The prognosis problem is the long-run outlook for the malady for patients whose cancer has been surgically removed. In this problem a patient is assessed as a ‘recur’ if the malady is determined at some future time to growth excision and a patient for whom cancer has not recurred and should ne'er recur. The objective of those predictions is to handle cases that cancer has not recurred (censored data) furthermore as case that cancer has recurred at a selected time. Thus, breast cancer diagnostic and prognostic issues square measure principally within the scope of the wide mentioned classification problems. In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

II. RELATED WORK

Breast cancer is one amongst the leading cancers for ladies in developed countries including Asian nation. It is the second most typical explanation for cancer death in women. The high incidence of breast cancer in women has redoubled considerably within the last years. In this paper we've got discussed varied data processing approaches that are used for carcinoma identification and prognosis. Carcinoma Diagnosis is identifying of benign from malignant breast lumps and carcinoma Prognosis predicts once Breast Cancer is to recur in patients that have had their cancers excised. This study paper summarizes various review and technical articles on carcinoma identification and prognosis additionally we tend to concentrate on current analysis being dole out the info mining techniques to boost the breast cancer identification and prognosis.

III. METHODOLOGY

I. Description of Wisconsin Breast Cancer Dataset

The cancer data used in this article comes from the Wisconsin Data set in the University of California at Irvine’s machine learning data collection ware house. The data set has 569 data points with 30 attributes per data point. The attributes from breast lumps by fine needle aspiration (FNA) are 10 kinds of digital image. The image of nucleus of the maximum value, average value and variance. These 10 kinds of features including radius, texture, perimeter, area, compactness, smoothness, concavity, concave points, symmetry, fractal dimension etc. the specific properties are illustrated as shown in the Table 1.

Table 1. Attributes of the Wisconsin Breast Cancer dataset

Mean Value	**Variance**	**Maximum value**
mean radius	Radius error	Worst radius
Mean texture	Texture error	Worst texture
Mean perimeter	Perimeter error	Worst perimeter
Mean area	Area error	Worst area
Mean smoothness	Smoothness error	Worst smoothness
Mean compactness	Compactness error	Worst compactness
Mean concavity	Concavity error	Worst concavity
Mean concave points	Concave points error	Worst concave points
Mean symmetry	Symmetry error	Worst Symmetry
Mean fractal dimension	Fractal dimension error	Worst fractal dimension

There are two properties in Breast cancer, Data, Target.
Data is a matrix, each column represents a 30 properties, a total of 30 rows, each row represents a measurement of breast lumps of digital image, sampling a total of 569 records.
Target is an array that store every kind of tumor belonging ro every record in data, so that the length of array is 569, and the value of array element is two, so the two values are different, 1 is malignant (M) and 0 is benign (B).

The results of output classification are as follows.

```
[11111111100001000000000000000001111111110000010101
0100000000000000111111111110000101011100001111001
111100001111000101010
………
0000011111000111111111000000001111010101010101101
1001010011100000111111010100111100010101]
```

II. Naïve Bayes Classification Algorithm

Naïve Bayes is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(A|B) from P(A), P(B) and P(B|A).
Look at the equation below:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{P(B|A)\,P(A)}{P(B)}$$

Above,

- *P*(*A*|*B*) is the posterior probability of *class* (A, *target*) given *predictor* (B, *attributes*).
- *P*(*A*) is the prior probability of *class*.
- *P*(*B*|*A*) is the likelihood which is the probability of *predictor* given *class*.
  P(*B*) is the prior probability of *predictor*.

IV. RESULTS AND DISCUSSION

In the study, two sets are designed for training set P with 75% samples and a test set T containing 25% samples. We trained the training dataset for learning afterwards we apply testing for classification dataset by naïve bayes and we got increase accuracy while having maximum radius and maximum texture. Based on the two features mean radius and mean texture used in this paper the classification accuracy rate is 98.6% , recall value is 37.6% and precision value is 100% . It can be seen that the two features selected in this paper are very effective in the diagnosis of cancer tumor.

Accuracy=TP+TN/(TP+TN+FP+FN)*100
Recall Value=TP/TP+TN
Precision Value=TP/TP+FP
Where,
TP= True Positives
TN= True Negatives
FP= False Positives
FN= False Negatives

Scatter diagram of data points is suitable for representing the general trend of dependent variables changing with independent variables, so we can choose suitable functions to fit data points.
Load the breast cancer dataset, then distinguish the malignant sample data and the benign sample data, and store them in data sets Benign and Malignant, respectively. 357 benign samples and 212 malignant samples are obtained.

From benign samples and malignant samples, two column data, mean radius and mean texture are extracted, and the values obtained are assigned to XB, YB, XM, YM variables. Finally, scatter function is called to draw scatter plots. The key code is as follows.

The first 50 malignant samples were drawn and marked with green dots.

XM [50],YM [50], color=’green’, marker=’<’, label= ‘malignant’

The first 50 benign samples were drawn and marked with red dots.

XB [50],YB [50], color=’red’, marker=’’, label= ‘benign’

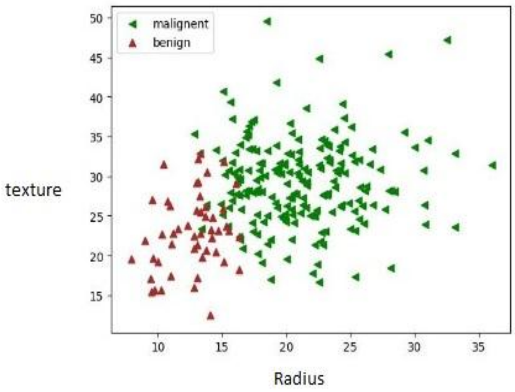

Figure 1. Scatter Plot of Classified Samples

It can be concluded from the results shown in the scatter plot that the discrimination of malignancy is directly related to the size and texture of the tumor, which provides reliable data for this conclusion.

V. CONCLUSION AND FUTURE SCOPE

In the experiment , the classification when choosing the mean radius and mean texture two characteristics , the use of all the training samples, the classification accuracy can reach 97.5% while choosing maximum radius and maximum texture as two characteristics , classification accuracy can reach 98.6% , therefore , choose the better feature combination will improve the classification accuracy . The experiment results show that the Naïve bayes classification algorithm model can be used to diagnose breast cancer quickly, easily and efficiently, and can help diagnose breast cancer. The dataset of the diagnostic test of Wisconsin breast cancer dataset used in this experiment. In the training phase, the tumor features were extracted from 32 original features. The results not only demonstrate the ability of the method to diagnose breast cancer, but also show time saving in the training phase. By better extracting the feature attributes of different types of tumors, we can effectively improve the classification accuracy of the method, and doctors can also benefits from the abstract tumor features.

REFERENCES

[1] Abdelghani Bellaachia, Erhan Guven, “Predicting Breast Cancer Survivability Using Data Mining Techniques”, The George Washington University, Washington DC 20052

[2] Shweta Kharya, “USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[3] G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani, “An Efficient Prediction of Breast Cancer Data using Data Mining Techniques”, International Journal of Innovations in Engineering and Technology (IJIET)

[4] A.PRIYANGA, Dr.S.PRAKASAM, “The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness”, International Journal of Computer Science and Engineering Communications- IJCSEC. Vol.1 Issue.1, December 2013

[5] Shelly Gupta, Dharminder Kumar, Anand Sharma “Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis”, Indian Journal Of Computer Science And Engineering (Ijcse)

[6] 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.

[7] 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.

[8] 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.

AUTHORS PROFILE

Mr. Pushpraj Saket pursed Bachelor of Engineer from RGPV university Bhopal, India in 2012. He is currently pursuing Master of Technology from Computer Application and Technology. And he has worked as a teacher of computer science before that.His main research work focuses on Breast cancer diagnosis by machine learning Naïve Bayes Algorithm , data mining He has 4 years of teaching experience.