OPEN ACCESS
International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Attribute Reduction and Cost Optimization using Machine Learning methods to Predict Breast Cancer

Magesh G, Swarnalatha P

Abstract*: *In this paper, Wisconsin breast cancer dataset is taken from UCI to minimize its features. It has thirty input variables and one output variable. In earlier, the prediction of breast cancer is made by machine learning algorithms like linear regression, neural network, decision tree, SVM and so on. Here, the features or input variables are reduced to eleven input features from thirty-two through similarity measure and optimization method. For this, first Pearson correlation is applied between the variables and the attributes are reduced when its pair has a 90% correlation. Then, Cost Optimization based Machine Learning algorithm is applied to the constraint pairs. From this result, it has observed that we can predict breast cancer with only two input features. The error rate and accuracy of various classifiers are also presented here.

***Index Terms*: Accuracy, Classification, Cost optimization, Machine learning.**

I. INTRODUCTION

The second most deaths causing cancer in women referred to breast cancer. This type of cancer arises at a rate of 1 in 37 people or 2.7%. After advanced screening and treatment, the survival rates have gradually increased. There are more than five million women in the world who survived breast cancer. The awareness of the symptoms, the need for screening in advances is essential in predicting breast cancer. Here, the diagnostic Wisconsin breast cancer from Olvi L. Mangasarian [1] has 569 instances with 32 attributes as features (ID, diagnosis, 30 real-valued input features). The diagnosis has two unique class distribution stating of 357 benign, 212 malignant. By considering this dataset, the machine learning algorithms [2 – 10] uses the feature attributes, to predict the diagnosis of breast cancer. Before ML and mining method, the researcher or data scientist applies statistics and similarity measure [11 – 13] to identify the primary features. This paper aims to deal with the prediction by reducing 30 to 2 input attributes (only one input from the actual dataset and the other is from machine learning output y as input). The efficiency for predicting is nearby to actual efficiency as shown in the experimental results section at step 5.

Revised Manuscript Received on 30 March 2019.
* Correspondence Author
Magesh G*, Professor, Department of SSE, SITE School, Vellore Institute of Technology Vellore University (Tamil Nadu), India
Swarnalatha P, Department of SSE, SITE School, Vellore Institute of Technology Vellore University (Tamil Nadu), India

II. RELATED WORK

A. Pearson correlation

Here the strength of association or relation r among the attributes x and y is measured.

$$r = \frac{m(\sum xy) - (\sum x)(\sum y)}{\sqrt{[m\sum x^2 - (\sum x)^2][m\sum y^2 - (\sum y)^2]}}$$

B. Machine learning algorithm

Broadly, ML [14] is classified into supervised, unsupervised and reinforcement. Mostly, supervised is smeared to classification or regression-based predictions.

III. PROPOSED METHOD

1. Firstly, we will reduce the thirty input data or features affecting Breast Cancer into eleven input variables as they have a high correlation between them.
2. Again, we reduce eleven input variables into two input variables by using the Pearson correlation coefficient method.
3. By using cost optimization Machine Learning algorithm, we can predict the two more input variables as y_1 and y_2. These are learned machine inputs.
4. From these four input variables, we can predict breast cancer, which is two from dataset features and two from machine learned inputs.
5. The error rate is slightly increased as 3% more for these four input features when compared with the thirty input features.
6. Now, we can predict the Breast Cancer within two input features from their body measurements.

IV. EXPERIMENTAL RESULTS

1. The error rate for the thirty input features is mentioned it the matrix form as Decision Tree Model on this dataset while taking all thirty input attributes as shown in Table The overall error is 7%, and averaged classification error is 8.75%.

Table I. Evaluation of Decision Tree Model using all thirty input attributes

Actual/ Predicted	B	M	Error Rate
B	5	1	1.9
M	5	2	15.6

2. After analyzing the highly effective inputs, we can find the Pearson's correlated matrix for 11 input variables as in the following Fig. 1.

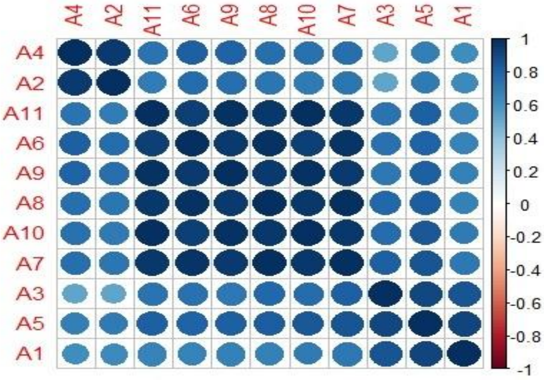

Fig. 1. Correlation Matrix

3. By using Pearson correlation, we can reduce the eleven input feature attributes into two input feature attributes by this analysis, which provides insights into the independence of the numeric input variables. They are a. Concavity_Mean and b. Texture_Mean. By Decision Tree model we can plot the graph with valid statements with these two input attributes for prediction as malignant or benign.

For Malignant

i. if Concavity_Mean is < 0.12 and Texture_Mean > 17 then it is Malignant.

ii. If Concavity_Mean > 0.12 then it is Malignant.

For Benign

i. if Concavity_Mean < 0.072 then it is Benign.

ii. if Concavity_Mean > 0.072 then it is Benign.

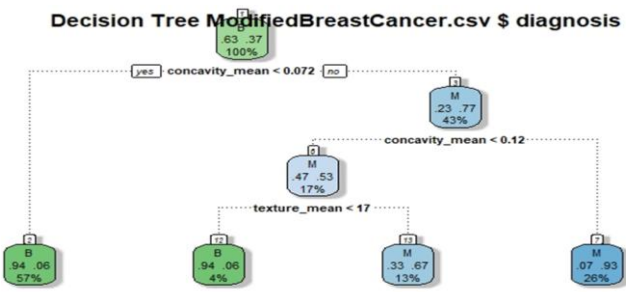

Fig. 2. Decision Tree

Table II. Cost Optimization of A_1 and A_2.

Attributes	A_1 Concavity_mean vs. Area_worst	A_2 Concavity_mean vs. Smoothness_mean
Pearson r	0.66	0.54
Optimized, θ_0 , θ_1	0.4519 , 4.8279	0.0882 , 0.921
$b = r.\frac{Std\ y}{Std\ x}$	384.959	0.1183
$a = \bar{y} - b\bar{x}$	234.416	0.0421

4. By using the Cost Optimization Machine Learning algorithm, we can make the machine to learn two more input variables; they are y_1 and y_2 as shown in Table III. Here the optimized θ_0, θ_1 values are obtained by Linear Regression as shown in Table II.

Table III. Linear Regression of y_1 and y_2 from the Concavity Mean

Attribute	y=a + bx
Concavity_mean vs. Area_worst	y_1=234.416 + 384.959 x
Concavity_mean vs. Smoothness_mean	y_2=0.0421 + 0.1183x

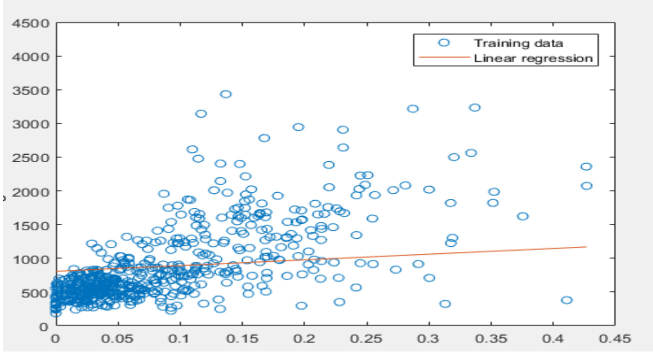

Fig. 3. $Y_{1's}$, θ_0, θ_1

In the above Table III, the equation y1 and y2 are obtained from the Table II variables such as Pearson r, θ_0, θ_1, b and a. They are y_1, and y_2 which are machine learned inputs by Cost Optimization Machine Learning Algorithm [3]. Where x is the input numerical taken from Concavity_Mean and θ_0, θ_1 are theta values obtained by Linear Regression.

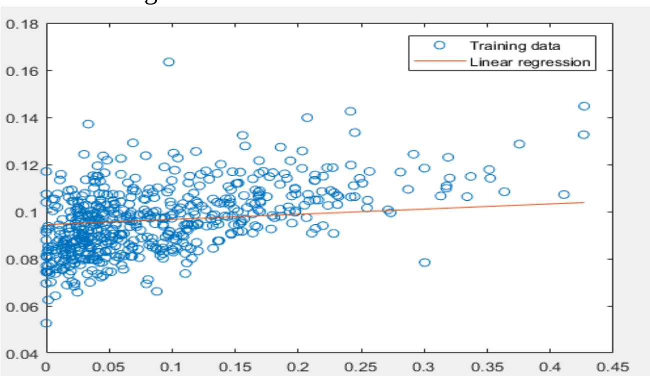

Fig. 4. $Y_{2's}$, θ_0, θ_1

5. Now, the Error Matrix is calculated between two input variables, from all the four input feature variables in which two of them are real input features, and the remaining two are machine learned input features.

Table 4. Error Rate using Concavity_Mean and y_1

Actual/ Predicted	B	M	Error Rate
B	4	7	13.2
M	2	3	6.2

The overall error is 10.6%, and an Averaged class error is 9.7%.

Table 5. Error Rate using Concavity_Mean and y_2

Actual/ Predicted	B	M	Error Rate
B	4	7	13.2
M	2	3	6.2

The overall error is 10.6%, and the Averaged class error is 9.7%.

Table 6. Error Rate using y_1 and y_2

Actual/ Predicted	B	M	Error Rate
B	4	7	13.2
M	2	3	6.2

The overall error is 10.6%, and the Averaged class error is 9.7%.

Table 7. Error rate using Concavity_Mean and Texture_mean

Actual/ Predicted	B	M	Error Rate
B	4	8	15.1
M	2	3	6.2

The overall error is 11.8%, and the Averaged class error is 10.65%.

Table 8. Error rate using Texture_mean and y_1

Actual/ Predicted	B	M	Error Rate
B	4	8	15.1
M	2	3	6.2

The overall error is 11.8%, and an Averaged class error is 10.65%.

Table 9. Error rate using Texture_Mean and y_2

Actual/ Predicted	B	M	Error Rate
B	4	8	15.1
M	2	3	6.2

The overall error is 11.8%, and the Averaged class error is 10.65%.

From the above results, we can find the optimized result for prediction of breast cancer by one of the two input variables from y_1, y_2, and Concavity_mean. Since the first three cases of above Step 5 has been holding with less error rate, i.e., 9.7%. Whereas the other cases of Step 5 hold with 10.65% error rate only.

V. Conclusion

From this analysis, we can able to predict breast cancer by reducing the input attributes. While considering all 30 input feature variables we got an overall error rate is 7% in which benign 1.9% and Malignant is 15.6 %. When we reduced the attributes to four and compared the Overall error rate is 10.6% where benign 13.2% and Malignant is 6.2 %. These error rate results are obtained between y1 (or) y2, Concavity_mean or y1 (or) y2. Hence we can predict the Breast Cancer Result from two input feature variable, as the error rate increases by just 3%. Now, Breast cancer prediction can be made with two input feature variables which reduce from 30 to 2 input features.

REFERENCE

1. O. L. Mangasarian, Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706.
2. C. Thirumalai and R. Manzoor, "Investigating the breast cancer tissue utilizing semi-supervised learning and similarity measure," 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 269-274. doi: 10.1109/ICECA.2017.8212814
3. Wang, Z. & Xu, X. "A sharing-oriented service selection and scheduling approach for the optimization of resource utilization," SOCA, 2012, 6: 15. https://doi.org/10.1007/s11761-011-0096-5
4. Rodríguez-cristerna, A., Gómez-flores, W., & Albuquerque, W. C. De. "A computer-aided diagnosis system for breast ultrasound based on weighted BI-RADS classes," Computer Methods and Programs in Biomedicine. Vol. 153, pp. 33–40. doi: 10.1016/j.cmpb.2017.10.004
5. Wen Jiang, Xianjun Xing, Shan Li, Xianwen Zhang, Wenquan Wang, Synthesis, characterization and machine learning based performance prediction of straw activated carbon, Journal of Cleaner Production, Volume 212, 2019, Pages 1210-1223.
6. C. Thirumalai and R. Manzoor, "Cost optimization using normal linear regression method for breast cancer Type I skin," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 264-268. doi: 10.1109/ICECA.2017.8212813
7. Zeng T LJ. Mixture classification model based on clinical markers for breast cancer prognosis. Artificial Intelligence in Medicine. 2010.
8. Peng, L., Chen, W., Zhou, W., Li, F., & Yang, J. (2016). An immune-inspired semi-supervised algorithm for breast cancer diagnosis. In Computer Methods and Programs in Biomedicine, Vol. 134, pp. 259-265, Elsevier Ireland Ltd.
9. Wen Jiang, Xianjun Xing, Xianwen Zhang, Mengxing Mi. Prediction of combustion activation energy of NaOH/KOH catalyzed straw pyrolytic carbon based on machine learning, Renewable Energy, Vol. 130, 2019, pp. 1216-1225.
10. Lu, H., Wang, H., & Yoon, S. W. (2019). A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. In Expert Systems with Applications (Vol. 116, pp. 340–350). Elsevier Ltd.
11. C. Thirumalai, M. Vignesh and R. Balaji, "Data analysis using box and whisker plot for lung cancer," 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, 2017, pp. 1-6.
12. K. R. Mujahid & C. Thirumalai. (2017). Pearson Correlation Coefficient Analysis (PCCA) on Adenoma carcinoma cancer. *ICEI*, pp. 492-495.
13. K. Sharma, B. Muktha, A. Rani & C. Thirumalai, (2017). Prediction of benign and malignant tumor. *ICEI*, pp. 1057-1060.
14. López, J., & Maldonado, S. "Robust twin support vector regression via second-order cone programming," Knowledge Based Systems, 2018, Vol. 152, pp. 83–93.