BreastCancerDetection

by Team3 Sadhananandhita

Submission date: 04-Feb-2022 10:55AM (UTC+0530)
Submission ID: 1754684901
File name: Team03_P_Sadhana.pdf (471.66K)
Word count: 1888
Character count: 9594

Detection of Breast Cancer using PCA, Logistic Regression and SVM

Nandhitha Ravishankar, P Sadhana, Dr. Sarada Jayan
Department of Computer Science and Engineering (Artificial Intelligence)
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
nandhitharavishankar@gmail.com, psadhana2002@gmail.com, j_sarada@blr.amrita.edu

Abstract— **Breast cancer is one of the biggest issues faced by women nowadays. According to global statistics, it is responsible for the majority of new cancer cases and cancer-related deaths, making it a serious public health issue in modern civilization. Support Vector Machine (SVM), Artificial Neural Network (ANN), logistic regression and Naive Bayes Algorithms are very popular and powerful supervised learning algorithms to classify and detect the presence of cancer. The dataset selected has ten individual attributes, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. The mean, standard error, and worst of these features are computed for each image which results in 30 features.**
Principal Component analysis was performed to reduce the dimensions of the dataset to obtain two columns of patients of those having breast cancer and those who don't. In this paper, we trained the dataset using Logistic Regression and SVM and obtained an accuracy rate of 96.5% and 90.9% respectively.

Keywords **PCA, SVM, Logistic Regression, Breast cancer**

I. Introduction

Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump. Breast cancer can occur in both men and women, but usually appears in women.

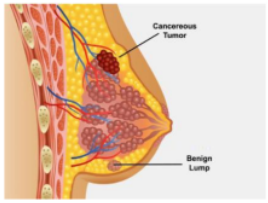


Fig. 1 Presence of a cancerous tumor and benign lump in a breast

II. DATASET

A dataset is a collection of related sets of data that is made up of individual pieces yet may be handled by a computer. Our dataset consists of data from 569 patients who either have cancer or not.

It has two attributes, id and diagnosis in columns one and two respectively. From column three through column thirty-two, we have 10 real-valued properties of the cell nucleus computed. They are,

- Radius: distances from centre to points on the perimeter
- Texture: standard deviation of grey-scale values
- Perimeter
- Area
- Smoothness: local variation in radius lengths
- Compactness: perimeter^2 / area - 1.0
- Concavity: severity of concave portions of the contour
- Concave points: number of concave portions of the contour
- Symmetry
- Fractal dimension: "coastline approximation" – 1

The mean, standard error, and worst of these features are computed for each image, resulting in 30 features.
Here, the mean is basically the mean of all the tests taken by a person, standard error is the standard deviation of the given values and worst is the average of the three largest values of the observation.

III. Diagnosis

When cells divide rapidly and abnormally it results in the formation of tumours. They are of two types:

1. Malignant (cancerous)
2. Benign (non-cancerous)

From the figure, we can see that among 569 patients, 357 patients are labelled benign and 212 as malignant.

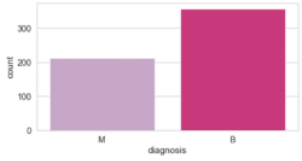


Fig. 2 Graphical representation of patients having Malignant and Benign tumor

For easier understanding of data, we have represented malignant as 1 and benign as 0 in MATLAB.

Now we'll observe patterns from 10 mean columns, here, as we can see, there are some interesting patterns visible. The nearly perfect linear relationships between the radius, perimeter, and area attributes for example, suggest that these variables are multicollinear. Another set of variables that can possibly imply multicollinearity are the concavity, concave_points and compactness.

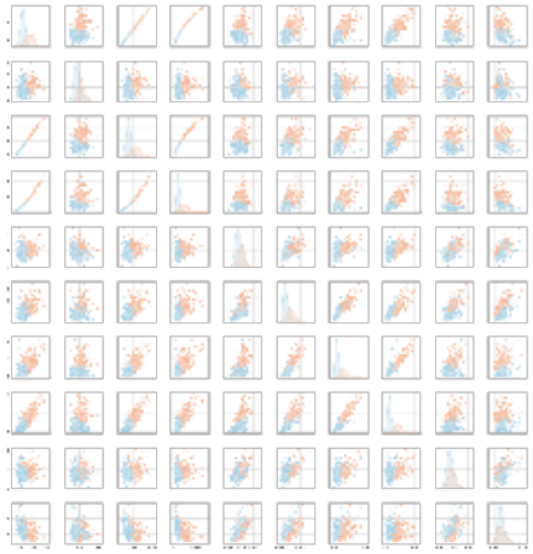


Fig. 3 Scatter plot of all the 10 attributes with each other

IV. CORRELATION MATRIX

A correlation matrix is a table that displays the coefficients of correlation between variables. Each table cell displays the correlation between two variables. Because the identical values will display on the other side when the same variables are compared again, we mask the upper triangular matrix.

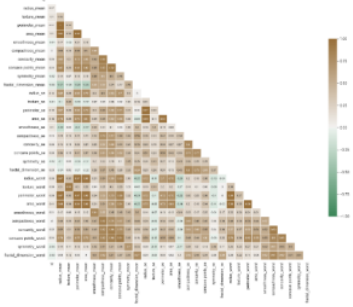


Fig. 4 Correlation matrix of all the 10 attributes with each other

It's worth noting that the worst columns have values that are similar to those in the mean columns. Because worst columns are essentially a subset of mean columns, we delete the worst columns to make things even simpler.

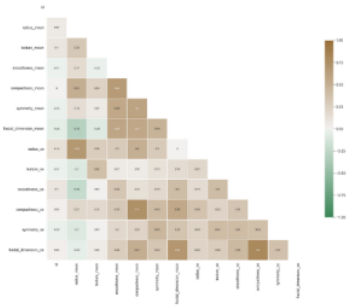


Fig. 5 Correlation matrix after removing the "worst" columns

V. PRINCIPAL COMPONENT ANALYSIS

It is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

PCA can be performed in two ways:

1. By finding mean, covariance and maximizing the variance

Maximizing the variance,

$Z = C1X + C2Y$

Linear combination of X and Y

So, to find the possible value of C1 and C2 where the variance will be the highest, we solve it using the optimization technique

Let us consider C1 as C1 and C2 and U2

$Z = U1X1 + U2X2 + \ldots\ldots\ldots + UDXD$

Now let us consider x and u as a vector that forms this

$$\vec{X} = \begin{bmatrix} X1 \\ X2 \\ \vdots \\ XD \end{bmatrix} \qquad U = \begin{bmatrix} U1 \\ U2 \\ \vdots \\ UD \end{bmatrix}$$

And now we can express z as an inner product

$Z = U.X = UTX = XTU$

Let us assume that we have a sample set and we know that these samples are vectors given in some space

$X = \{ X1, X2, \ldots\ldots\ldots XN \}$

Mean is the average of all the vectors given in some space

Sample mean

$$X = 1/N \sum Xi$$

Covariance forms a matrix of size DxD and is generally positive semi definite matrix i.e., all the eigenvectors are positive or zero.

Covariance of x

$$C = 1/N \sum Xi \, . \, Xi'$$

Variance is $z = utx$, where u is the linear transformation which transforms D dimensional space to a single dimensional space.

$Var(z) = 1/N \sum (Z-Z)^2$

On simplifying, we get

Var(z) = U'CU

To maximize var(z), we can assume that U tends to infinity

To avoid this, we must have limitations on the size of the vectors

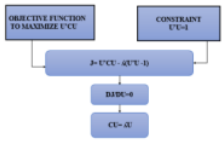
$$\|U\| = 1$$

Or

$$U'U = 1$$

Here, U is the linear transformation which transforms D dimensional space to a s single dimensional space.

The lagrange multiplier is used to augment the constraint into the equation



Finally, to compute the principal components,

T = mean*eigenvectors(c).

2. Using singular value decomposition

$$X = U.\Sigma.V^T$$

Where,

U – orthogonal matrix with unit eigenvectors of AA' in columns

V – orthogonal matrix with unit eigenvectors of A'A in columns

Σ – leading diagonal as square root of eigenvalues of A'A and AA'

$$T = U.\Sigma$$

Singular values in Σ gives us the indication of the amount of variance of the dataset that principal components capture. Here we take 2 principal components as we have to results for the diagnosis, either they have cancer or they don't.

A. *Plotting graphs*

Percentage of plotting singular values and cumulative variance graph:

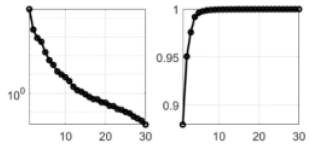


Fig. 6 Plotting singular values on a log scale and cumulative variance

As seen in the 1st graph, the percentage of plotting the singular values on a log scale, x axis - (number of columns (rank)), y axis - magnitude of log of the singular value. We can see that the first few values are quiet high, which indicate that those first few values have the most energy, so they are more likely to have the maximum variance and then it is seen that the graph tapers off.

In the 2nd graph, we plot the cumulative variance, it basically gives us the percentage of variances accounted for by the first n components. Since we have taken two principal values, we can see that nearly 95% of the data has been extracted.

Graph of data after PCA:

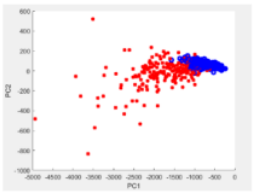


Fig. 7 Formation of two clusters after PCA

We can see the formation of two clusters. The red cluster or color indicates the patients who have cancer and the blue cluster or color indicate the patients who don't have cancer.

Graph using new data:

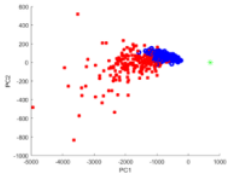


Fig. 8 Formation of two clusters and a point after PCA with the new data

We predict if a person has cancer or not with the data provides. Certain values are given and when the model runs, we observe that the data of the new patient lies on the side where the patients do not suffer from cancer. This indicates that the patient has a higher chance of not having cancer.

VI. LOGISTIC REGRESSION

A. *Model*

The dataset is divided into two sections: training and testing. The training dataset is a set of data used to teach a programme how to employ technologies like neural networks to learn and produce sophisticated results, whereas the test set is a collection of observations used to assess the model's performance using some performance metric.

B. *Confusion matrix*

True negative = 111

False positive = 4

False negative = 2

True positive = 54

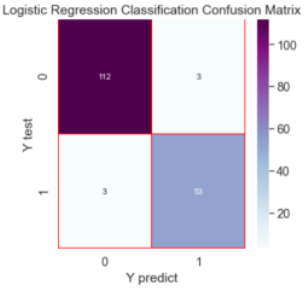


Fig. 8 Confusion matrix obtained after performing Logistic Regression

C. Prediction

After we train the model, testing is done and based on the values that we get, we put it in the matrix (confusion matrix). The output is the correct prediction.

```
              precision    recall  f1-score   support

           B       0.974     0.974     0.974       115
           M       0.946     0.946     0.946        56

    accuracy                          0.965       171
   macro avg       0.960     0.960     0.960       171
weighted avg       0.965     0.965     0.965       171

Confusion Matrix:
 [[112   3]
 [  3  53]]

True Negative: 112
False Positive: 3
False Negative: 3
True Positive: 53
Correct Predictions 96.5 %
```

Fig. 9 Prediction obtained after performing Logistic Regression

VII. SUPPORT VECTOR MACHINE

It is one of the most widely used Supervised Learning algorithms, with applications in both classification and regression. It's a binary linear classification with a deliberately built decision boundary to reduce generalisation error. It can perform linear and nonlinear classification, regression, and even outlier detection.

A. Confusion matrix

True negative = 83
False positive = 7
False negative = 6
True positive = 47

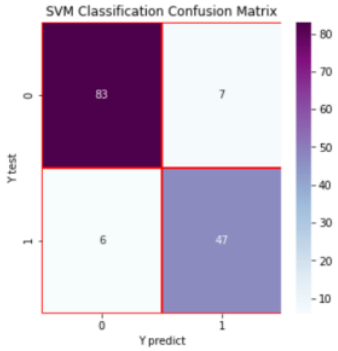


Fig. 10 Confusion matrix obtained after performing SVM

B. Prediction

- Pink ones show that does not have cancer, benign
- Blue ones show that it has cancer, malignant

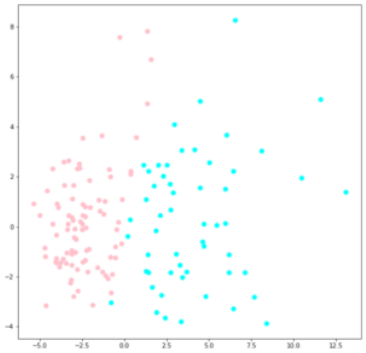


Fig. 10 Prediction obtained after performing SVM

VIII. CONCLUSION

The method proposed in this study, can be used to effectively predict breast cancer incidents in patients.

On analyzing the data, we were able to verify the presence of multicollinearity between some of our variables. It was observed that the radius_mean column has a correlation of 1 and 0.99 with perimeter_mean and area_mean columns, respectively.

Using PCA, we reduced the large data and obtained two clusters each indicating patients either having cancer or not respectively thus making analysis easier. The cumulative proportion of the top two major components was 95% Now when we plotted the data of the new patient, we were able to predict the possibility whether the patient is suffering from cancer or not. The use of logistic regression helped us to create a model that would predict the presence of cancerous cells in the patient by using training and testing datasets and have achieved an accuracy of 96.5% that is, our model has accurately labeled 96.5% of the test data. Using python, we obtained the accuracy rate to be 90.9% for SVM and in order to achieve a higher accuracy rate, various other algorithms such as k-means and kNN can be implemented.

REFERENCES

[1] Huan-Jung Chiu, Tzuu-Hseng S.Li, and Ping-Huan Kuo, "Breast Cancer–Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and Support Vector Machine", November 10, 2020

[2] Ade Jamala1, Annisa Handayania2, Ali Akbar Septiandria3, Endang Ripmiatina4 , Yunus Effendi –" Dimensionality Reduction using PCA k-Means Clustering for Breast Cancer Prediction", 3 December 2018

[3] Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan- "Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors", December 23, 2017

[4] Jabeen Sultana1, Abdul Khader Jilani2 –" Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers", 2018

BreastCancerDetection

ORIGINALITY REPORT

22%	15%	12%	18%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Exeter Student Paper	3%
2	scholarcommons.scu.edu Internet Source	3%
3	ijariie.com Internet Source	2%
4	Saravanan M. S, Pradnya Patil, K. Venkata Subbaiah. "Analysis of breast cancer event logs using various regression techniques", 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021 Publication	2%
5	Submitted to University of Hertfordshire Student Paper	2%
6	Submitted to Government College of Engineering Aurangabad, Maharashtra State, India Student Paper	2%
7	Submitted to Higher Education Commission Pakistan	1%

	Student Paper	
8	doaj.org Internet Source	1 %
9	Submitted to Sultan Qaboos University Student Paper	1 %
10	Submitted to The Robert Gordon University Student Paper	1 %
11	Submitted to Birkbeck College Student Paper	1 %
12	dmas.lab.mcgill.ca Internet Source	1 %
13	ebin.pub Internet Source	1 %
14	trepo.tuni.fi Internet Source	1 %
15	Submitted to Amrita Vishwa Vidyapeetham Student Paper	<1 %
16	Huan-Jung Chiu, Tzuu-Hseng S. Li, Ping-Huan Kuo. "Breast Cancer–Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and Support Vector Machine", IEEE Access, 2020 Publication	<1 %

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		