Classification Models

5CS037

Abeykoon Mudiyanselage, Nirmal P.

20/21

Contents

[Breast Cancer Classification With SVM 3](#_Toc58835897)

[About the Model 3](#_Toc58835898)

[Why SMV 4](#_Toc58835899)

[Data Set 4](#_Toc58835900)

[Data prepossessing 4](#_Toc58835901)

[Data visualization 6](#_Toc58835902)

[Accuracy and Predictions 9](#_Toc58835903)

[Improvements and Suggestions 10](#_Toc58835904)

[Changing the Kernel 10](#_Toc58835905)

[Changing the training dataset randomness 10](#_Toc58835906)

[Increasing the number of training recodes 12](#_Toc58835907)

[Changing the features 13](#_Toc58835908)

[Breast Cancer Classification with Random Forest 15](#_Toc58835909)

[About Random Forest 15](#_Toc58835910)

[Data set 15](#_Toc58835911)

[Accuracy and Prediction 16](#_Toc58835912)

[Improvements and Suggestions 17](#_Toc58835913)

[Changing the Max depth and random state of the model 17](#_Toc58835914)

[Changing the training dataset randomness 17](#_Toc58835915)

[Increasing the number of training records 18](#_Toc58835916)

[References 20](#_Toc58835917)

# Breast Cancer Classification With SVM

Breast cancer is one of the most common cancer types among women. Breast cancer is caused by when cells in the breast begin to grow without control. The main form of diagnosis for breast cancer is an X-ray or a CT scan.

Early diagnosis of any type of cancer can increase the chance of curability of the patient. A tumor can be malignant or benign, malignant means that the tumor cells can be expanded to other parts of the body, a benign tumor won't spread to other parts of the body like a cancer tumor but it may also cause some health problems. (The American Cancer Society medical and editorial content team, 2019)

With an appropriate machine learning model, a computer can detect breast cancer with an accuracy of 90% or more. In the following section, we are going to discuss how we can implement a suitable machine learning model to detect breast cancer.

## About the Model

For this project, we are going to use support vector machines known as the SVM algorithm to classify if a cell is malignant or benign.

SVM is a binary linear classification model this model is explicitly created to minimize the generalization error. This model can be used for linear classification or regression. According to the sklearn documentation on this model, it is most suitable to classify small to medium-sized data sets.

In this task the classification dataset is linearly separable hence the algorithm uses the largest margin between the closes points to identify if a cancer cell is malignant or benign. This process is called maximum margin hyperplane or MMH. (Jakkula, n.d.)

### Why SMV

Generally, neural networks are far more superior in classification problems but with this problem the regular patterns that appear with attributes and a small number of recodes in the dataset (600) justify the use of SVM. Furthermore, the SVM algorithm can avoid the pitfalls of very high dimensional representations.

Even though SVM is a very reliable and accurate machine learning method there are some challenges when using SVM such as choosing a kernel that impacts the final accuracy of the model massively. We will discuss this issue in the improvement and suggestion topic. (Jakkula, n.d.)

## Data Set

This dataset was downloaded from kaggles.com and it was uploaded by UCI Machine Learning. (Learning, 2016).

This dataset was created by analyzing digitized images of a fine needle aspirate (FNA) of a breast mass. The dataset contains 32 attributes, ID number, and the diagnosis. There are 569 records in the dataset.

### Data prepossessing

There was some prepossessing required in this dataset. The diagnosis was in string format, using sklearn.preprossesing M(malignant) was converted to 1s and B(benign) was turned to 0s.

Furthermore, there was a column named Unnamed:32 this column had Nan values this was dropped before using the dataset.

Head of the dataset before prepossessing.

id diagnosis ... fractal\_dimension\_worst Unnamed: 32

0 842302 M ... 0.11890 NaN

1 842517 M ... 0.08902 NaN

2 84300903 M ... 0.08758 NaN

3 84348301 M ... 0.17300 NaN

4 84358402 M ... 0.07678 NaN

Head of the dataset after prepossessing.

id diagnosis ... symmetry\_worst fractal\_dimension\_worst

0 842302 1 ... 0.4601 0.11890

1 842517 1 ... 0.2750 0.08902

2 84300903 1 ... 0.3613 0.08758

3 84348301 1 ... 0.6638 0.17300

4 84358402 1 ... 0.2364 0.07678

There are no Null values left in the dataset. (Null value percentage )

id 0.0%

diagnosis 0.0%

radius\_mean 0.0%

texture\_mean 0.0%

perimeter\_mean 0.0%

area\_mean 0.0%

smoothness\_mean 0.0%

compactness\_mean 0.0%

concavity\_mean 0.0%

concave points\_mean 0.0%

symmetry\_mean 0.0%

fractal\_dimension\_mean 0.0%

radius\_se 0.0%

texture\_se 0.0%

perimeter\_se 0.0%

area\_se 0.0%

smoothness\_se 0.0%

compactness\_se 0.0%

concavity\_se 0.0%

concave points\_se 0.0%

symmetry\_se 0.0%

fractal\_dimension\_se 0.0%

radius\_worst 0.0%

texture\_worst 0.0%

perimeter\_worst 0.0%

area\_worst 0.0%

smoothness\_worst 0.0%

compactness\_worst 0.0%

concavity\_worst 0.0%

concave points\_worst 0.0%

symmetry\_worst 0.0%

fractal\_dimension\_worst 0.0%

### Data visualization

#### Diagnosis ratio

In the following chart, we can see there is about a 2:3 ratio between malignant and benign recodes. If either of one of these types has more recodes the model accuracy may reduce due to the model randomly classifying data because they have more recodes.

Chart, bar chart

Description automatically generated

Figure :ratio between data recodes

#### Correlation heatmap

This is the correlation heatmap of all the features(before prepossessing).In the following graph, we can see there is a strong correlation between the mean radius and the parameter mean.

Chart

Description automatically generated

Figure : Correlation chart

Considering this chart, we will use the following 10 attributes for classification.A picture containing calendar

Description automatically generated

Figure :Correlation diagram of the used features

By looking at this graph we can see that there are good correlations between all the values that we going to use. The exception is fractal\_dimensio\_mean. This feature is poorly correlating with few other features.

#### Histogram

In this chart, we can see a histogram of all the numerical values.

Chart

Description automatically generated

Figure : histograms of all the numerical values

#### Scatterplot matrix

A picture containing timeline

Description automatically generated

Figure :scatter matrix of used features

By studying this diagram, we can see that a few features show clear upward trends indicating good correlation. Furthermore, we can see that these points are not too dispersed. These well-connected attributes match the result we saw with the correlation heatmap (Figure 3: Correlation diagram of the used features).

## Accuracy and Predictions

Accuracy: 91.22807017543859 %

Actual: B Prediction: B

Actual: B Prediction: B

Actual: B Prediction: B

Actual: B Prediction: B

Actual: B Prediction: B

Actual: B Prediction: B

Actual: B Prediction: B

Actual: M Prediction: M

Actual: B Prediction: B

Actual: M Prediction: M

Actual: M Prediction: M

Actual: B Prediction: B

Actual: B Prediction: B

Actual: B Prediction: B …

Chart

Description automatically generated

When we study the confusion matrix, we can see that the model classified the 0 (benign) with almost 100% accuracy and 1 (malignant ) with about 80% accuracy.

## Improvements and Suggestions

In this section, we will look at what are the improvements we can do to increase the accuracy of the SVM model.

### Changing the Kernel

The kernel of SVM helps the computations of data in more than 2 dimensions. In a nutshell, the kernel allows using original feature space without computing the dimensions. For the task polynomial kernel known as Polly manages to produce the best results.

In the following chart, we can see the accuracy of each kernel.

|  |  |
| --- | --- |
| Kernel | Accuracy |
| Linear | 91.22807017543859% |
| Polly | 92.98245614035088% |
| RBF | 85.96491228070175% |

For this classification task, we cannot use the precomputed kernel. This kernel requires a square matrix as the input dataset E.g. (10,10).

We can see when we use the polynomial kernel, we see improved accuracy compared to the linear kernel that we used in the base model. That is about a 1.7% improvement.

### Changing the training dataset randomness

In this section, we will look at how the random state parameter affects the accuracy of the model.

Random State= 0

Accuracy: 92.98245614035088 %

Random State= 1

Accuracy: 91.22807017543859 %

Random State= 2

Accuracy: 94.73684210526315 %

Random State= 3

Accuracy: 92.98245614035088 %

Random State= 4

Accuracy: 89.47368421052632 %

Random State= 5

Accuracy: 96.49122807017544 %

Random State= 6

Accuracy: 96.49122807017544 %

Random State= 7

Accuracy: 94.73684210526315 %

Random State= 8

Accuracy: 89.47368421052632 %

Random State= 9

Accuracy: 91.22807017543859 %

Figure :Accuracy compared to a random state

When looking at the results we can see that random state 6 and produce the best results with an accuracy of 96%.

### Increasing the number of training recodes

In this section, we will look at how the number of training recodes going to affect the accuracy of the model.

since we cannot increase the number of training recodes, we will reduce the number of training recodes and see how they affect accuracy.

Number of training recodes= 455

Accuracy: 92.10526315789474 %

Number of training recodes= 341

Accuracy: 92.54385964912281

Number of training recodes= 227

Accuracy: 92.10526315789474 %

Number of training recodes= 113

Accuracy: 90.35087719298247 %

Number of training recodes= 56

Accuracy: 89.08382066276803 %

Number of training recodes= 28

Accuracy: 88.53974121996303 %

Number of training recodes= 5

Accuracy: 85.46099290780141 %

Figure : accuracy compared to the number of records

When we study this data, we can see that the accuracy of the model does go down related to the number of training records. But there is not a big drop in accuracy even their training records reduced massively.

### Changing the features

in this section, we will look at how the feature set affects the final accuracy of the model. For the base model and all the tests above, we used the following features.

X = list(zip(data["radius\_mean"], data["texture\_mean"], data["perimeter\_mean"], data["area\_mean"],  
data["smoothness\_mean"], data["compactness\_mean"], data["concavity\_mean"], data["concave points\_mean"],data["symmetry\_mean"], data["fractal\_dimension\_mean"]))

to test this, we will remove most correlating features “area mean”, “radius mean” and “perimeter mean” from the data frame and test the accuracy of the model.

Without these features, the accuracy of the model got reduce to 87.71929824561403 %. Although this looks like not a huge loss of accuracy, considering this model has been used to categorize medical recodes this is a huge loss of accuracy.

Furthermore, if we remove the feature fractal\_dimension\_mean which is not correlating well with other features we will get an accuracy of 92%

# Breast Cancer Classification with Random Forest

## About Random Forest

The random forest which is like the decision tree model uses a tree structure to make predictions or classification.

According to the (Friedman, 2009) inaccuracy is one of the main reasons that decision tree models are not great for predictive learning. This means although the decision tree model works great with the training data the performance levels drop when classifying new data. Random forest uses the simplicity of the diction trees and combines flexibility to improve accuracy.

In a nutshell, Random forest is an extension of the decision tree.

## Data set

Since we are using the same dataset, we used in the SVM model there is no additional preprocessing required other than we did in the SVM model.

## Accuracy and Prediction

Accuracy: 93.85964912280701 %

Actual: B Prediction: B

Actual: B Prediction: B

Actual: B Prediction: B

Actual: B Prediction: B

Actual: M Prediction: M

Actual: M Prediction: M

Actual: M Prediction: M

Actual: B Prediction: B

Actual: B Prediction: B

Actual: M Prediction: M

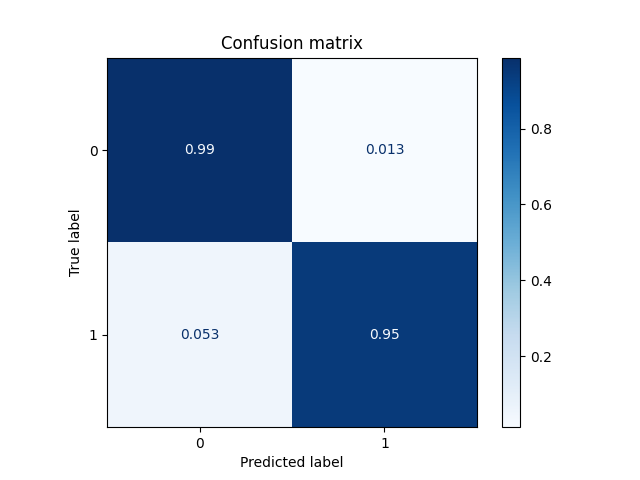


Figure :Model Confusion matrix

## Improvements and Suggestions

### Changing the Max depth

In this section, we will investigate how we can adjust the random state and max depth of our model to increase accuracy.

Max depth represents the depth of the tree. Deeper a tree is more splitting points it has because of that a deeper tree can fit more data.

MaxDepth= 1

Accuracy: 87.71%

MaxDepth= 2

Accuracy: 90.35 %

MaxDepth= 3

Accuracy: 91.22 %

MaxDepth= 4

Accuracy: 92.10 %

MaxDepth= 5

Accuracy: 91.22 %

MaxDepth= 6

Accuracy: 92.10 %

MaxDepth= 7

Accuracy: 92.98 %

MaxDepth= 8

Accuracy: 92.10 %

MaxDepth= 9

Accuracy: 92.10 %

MaxDepth= 10

Accuracy: 92.10 %

MaxDepth= 11

Accuracy: 92.10 %

MaxDepth= 12

Accuracy: 92.10 %

MaxDepth= 13

Accuracy: 92.10%

When studying this graph, we can see that when we increase the max depth pass certain point we don’t see any improvement after that. For this model MAXDEPTH=7 produced the best results.

### Changing the training dataset randomness

Using the method, we used for the SVM model we are going to see how the random state affects the accuracy of our model. We will only change the training data random state parameter other parameters are remain same as the base random forest model.

Random State= 0

Accuracy: 91.22807017543859 %

Random State= 1

Accuracy: 92.98245614035088 %

Random State= 2

Accuracy: 89.47368421052632 %

Random State= 3

Accuracy: 90.35087719298247 %

Random State= 4

Accuracy: 92.10526315789474 %

Random State= 5

Accuracy: 94.73684210526315 %

Random State= 6

Accuracy: 91.22807017543859 %

Random State= 7

Accuracy: 95.6140350877193 %

Random State= 8

Accuracy: 92.98245614035088 %

Random State= 9

Accuracy: 92.98245614035088 %

Figure :Accuracy change with the random state

### Increasing the number of training records

We are going to see how the number of training recodes going to affect the model accuracy. For this, we are going to use the same method we used in the SVM model.

Number of training recodes= 512

Accuracy: 92.98245614035088 %

Number of training recodes= 455

Accuracy: 92.98245614035088 %

Number of training recodes= 398

Accuracy: 91.22807017543859 %

Number of training recodes= 284

Accuracy: 94.03508771929825 %

Number of training recodes= 170

Accuracy: 91.72932330827066 %

We can see that this model does not depend heavily on the number of training records if it has well-correlating features as we have in this case.

Figure :accurasy change with the number of training recodes

# References

Andy Liaw, M. W., 2002. *Classification and Regression by.* [Online]   
Available at: https://www.researchgate.net/profile/Andy\_Liaw/publication/228451484\_Classification\_and\_Regression\_by\_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-by-RandomForest.pdf  
[Accessed 25 11 2020].

Dibike, Y. B., Velickov, S., Solomatine, D. & Abbott, a. M. B., 2001. *Model Induction with Support Vector Machines: Introduction and Applications.* [Online]   
Available at: https://ascelibrary.org/doi/abs/10.1061/(ASCE)0887-3801(2001)15:3(208)  
[Accessed 17 10 2020].

Esmaily, H., 2018. *A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes.* [Online]   
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204421/  
[Accessed 2020 11 20].

Friedman, 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York: Springer.

Géron, A., 2020. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems.* Second ed. Sebastopol: O’Reilly Media.

Jakkula, V., n.d. *Tutorial on Support Vector Machine (SVM.* [Online]   
Available at: https://course.ccs.neu.edu/cs5100f11/resources/jakkula.pdf  
[Accessed 20 10 2020].

Learning, U. M., 2016. *Breast Cancer Wisconsin (Diagnostic) Data Set.* [Online]   
Available at: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data  
[Accessed 10 10 2020].

Russell, S. P. N., 2016. *Artificial Intelligence : a Modern Approach.* Third ed. Harlow, Essex, England: Pearson.

The American Cancer Society medical and editorial content team, 2019. *What Is Breast Cancer.* [Online]   
Available at: https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html#written\_by  
[Accessed 17 10 2020].