Simple Breast Cancer Detector

Khanh Gia Nguyen

Author Note

Correspondence concerning this article should be addressed to Khanh Nguyen Gia, 304 Duong Quang Ham, Go Vap district, Hochiminh city, Vietnam.

Contact: ngiakhanh96@gmail.com Tel: 0986853293

Abstract

This paper focuses on the implementation of 7 different Supervised Learning algorithms (Logistic Regression, K Nearest Neighbors (KNN), linear SVM, kernel SVM, Naïve Bayes, Decision Tree, Random Forest) as well as a Deep Learning approach with a simple Artificial Neural Network (ANN) to classify an unseen cell as a benign cell or a malignant one. This article also compares the performance among those techniques to determine the most appropriate model with accuracy as the only used criterion.

Keywords: classification, multivariate, breast cancer detector, logistic regression, k nearest neighbors, linear SVM, kernel SVM, Naïve Bayes, decision tree, random forest, artificial neural network.

Simple Breast Cancer Detector

Cancer has always been with us since the dawn of mankind, from ancient civilizations to today. It appears to be an incurable disease and remains an unsolvable challenge for many genial scientists and doctors for thousands of years.

Among many types of cancer, breast cancer is the most commonly occurring one in women and the second most common cancer overall which affects 2 million people each year (World Cancer Research Fund, 2019). Therefore, there is an urgent need to build and develop a good model which can help identify breast cancer cells correctly and quickly so that many women and men all over the world can detect early their breast cancer condition which leads to more chance to be completely cured of.

This paper introduces and compares the performance in terms of accuracy among 7 famous Supervised Learning algorithms which are Logistic Regression, K Nearest Neighbors, linear SVM, kernel SVM, Naïve Bayes, Decision Tree, Random Forest and a Deep Learning method using Artificial Neural Network to build a simple breast cancer detector. The breast cancer dataset is taken from UC Irvine Machine Learning Repository (UCI, 1992) containing 699 samples with 2 output classes as benign and malignant cells.

Evaluation Method

Our research took advantages of K-fold Cross-Validation technique to evaluate our final results. From the training set, we continued to split it into 10 folds. 9 of them will be our new training set

while the other is our new test set, hence, we now have 10 different choices for the test set leading to the total of 10 iterations for Cross-Validation.

Preprocessing Dataset

Preprocessing dataset is always the most time-consuming and important step when dealing with very large data and therefore must be done extremely carefully. The performance of a model can be severely affected by a very small mistake in this step.

In our experiment, we downloaded and stored the dataset in a CSV file. We also deleted all "?" representing missing values for the convenient purpose when we fit our model later on.

Firstly, we had to choose how to deal with these missing values. There are 2 common ways: remove the entire row possessing the missing values or replace the missing values with the values we want. They can be the mean, median, some desired fixed values or the value having the most frequency of the corresponding features. Here our strategy is choosing the most frequent values.

Then, in order for our model to know that we have 2 output classes, we encode our dependent variables using One Hot Encoder.

Finally, we split our dataset into Training set (85%) and Test set (15%) and do the feature scaling using Standardization method.

Fitting Supervised Learning Algorithm

We sequentially fit 7 different algorithms which are Logistic Regression, K Nearest Neighbors, linear SVM, kernel SVM, Naïve Bayes, Decision Tree, Random Forest and compared their performance.

Model	Mean Accuracy(%)	Standard deviation(%)
Logistic Regression	96.81	2.14
KNN	96.48	2.60
Linear SVM	96.31	1.91
Kernel SVM	96.65	2.76
Naive Bayes	96.48	2.14
Decision Tree	93.77	2.24
Random Forest	95.98	3.23

Figure 1. Performance comparison of 7 Supervised Learning algorithms.

The figure shows that the top 5 algorithms nearly have the same accuracy which is over 96%. While the worst one is Decision Tree whose accuracy is only 93.77%. Nevertheless, when coming to the standard deviation of the mean accuracy, it is a quite different story. Random Forest, KNN and Kernel SVM stands on top whose error tolerances all equals or higher than 2.60%. Linear SVM surprisingly surpass all other techniques to keep its deviation to be under 2% - only 1.91%. Logistic Regression, Kernel SVM and Decision Tree have 2.14%, 2.76% and 2.24% standard accuracy respectively.

In general, Logistic Regression is clearly the most suitable algorithm due to its high accuracy and decent standard deviation. Besides, it is also worth giving Linear SVM and Naïve Bayes a try.

Fitting Artificial Neural Network

We fit a simple 2-layer ANN with 5 nodes for each layer and chose Rectified Linear Unit (reLU) and Softmax as Activation functions and tested it with different numbers of epochs.

Number of epochs	Mean Accuracy(%)
10	97.81
20	97.98
50	98.15
100	98.48

Figure 2. ANN mean accuracies when varying the number of epochs from 10 to 100.

From the above figure, it can be easily seen that the increase in the number of epochs leads to the improvement in the overall accuracy, from 97.81% for 10 epochs to 98.48% for 100 epochs. However, the dataset is too small and may not cover the most common cases. Hence, the model has a very high chance to be overfitting and bias.

Evaluation and Conclusion

Overall, the ANN approach has a higher accuracy from a very small number of epochs, 97.81% compared to 96.81% - the highest accuracy among Supervised Learning algorithms. However, this approach is worse when talking about the interpretation. Therefore, the choice depends on the our purpose. If the accuracy is the most concern, then ANN is the best option. Otherwise, if the interpretation is a preferred factor, Linear Logistic, Linear SVM or Naïve Bayes should on top of the chosen list.

References

World Cancer Research Fund. (2019). *Breast cancer statistics*. [online] Available at: https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics [Accessed 20 Mar. 2019].

O. L. Mangasarian and W. H. Wolberg. *Cancer diagnosis via linear programming*, SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.