Comparing machine learning techniques to diagnose breast cancer

Alejandro Giovanni Navarro Elenes - Take Home Final Exam – 2020-12-08

Introduction

Breast cancer forms in the cells of the breasts, and although it can occur in both men and women, it is far more common in women, and it is the second most common cancer in women in the United States, only after skin cancer. Breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more rapidly than healthy cells do and continue to accumulate, forming a lump or mass. Cells may spread through the breast to the lymph nodes or to other parts of the body.

The process of testing and diagnosing breast cancer requires a combination of individuals with different expertise, but the final diagnosis is given by a physician. If a machine learning algorithm could automate the process of breast cancer diagnosis, physicians could spend more time leveraging treatments for the disease. Many machine learning techniques are used as tools for classification of new or future data. These are often evaluated by their rate of accuracy, for example, a model with 97% of successfully diagnosed cases is better than a model with 93%.

However, an alternative evaluation of machine learning models for breast cancer diagnosis may be beneficial, where one weighs more heavily those models with fewer cases of false negatives. A positive case of breast cancer mistakenly diagnosed as negative would have disastrous consequences for the patient, who would think they are cancer-free, but the disease may continue to spread. In this study, I explore different machine learning models and compare how they perform regarding false negative counts and overall accuracy, looking for diverging trends between the two criteria.

Methods

We compare the following methods using 5-fold cross-validation.

1) K nearest-neighbor
2) Naïve-Bayes
3) Linear discriminant analysis
4) Quadratic discriminant analysis
5) Linear kernel Support Vector Machine (SVM)
6) Quadratic kernel SVM
7) Radial kernel SVM
8) Tuned linear kernel SVM
9) Tuned quadratic kernel SVM
10) Tuned radial kernel SVM
11) Neural networks
12) Random forest
13) Booster gradient

The data is hosted in the University of California Irvine Machine Learning Repository, where they provide information about the data sources and creators, a description of the dataset, relevant papers from the creators and other papers citing the dataset (Dua & Graff, 2019). Features are computed from a digitalized image of a fine needle aspirate of a breast mass, and they describe characteristics of the cell nuclei present in the image. The process of generating the images that originated the data measurements is thoroughly explained in Barford (n.d.).

There are 32 variables in the dataset; the attribute information includes: 1) ID number, 2) Diagnosis (M = malignant, B = Benign) which is the variable to be predicted; the remaining 30 variables are the mean, the standard deviation and the worst measurement (mean of the 3 largest values) of ten real-valued features computed for each cell nucleus, all recoded with four significant figures:

a) Radius (mean of distances from center to points on the perimeter)
b) Texture (standard deviation of gray-scale values)
c) Perimeter
d) Area
e) Smoothness (local variation in radius lengths)
f) Compactness (perimeter^2 / area - 1.0)
g) Concavity (severity of concave portions of the contour)
h) Concave points (number of concave portions of the contour)
i) Symmetry
j) Fractal dimension (“coastline approximation” – 1)

Results

A first exploratory data analysis is summarized in the correlogram of all the 30 variables, where the diagnosis variable is used to show class membership. Some trends can be seen by visual inspection of the correlogram in Figure 1.

In the diagonal, I show the distribution of the data for malignant (blue) and benign (red) tumors; the more separated these red and blue distributions are, the better the data will help in predicting the malignant or benign property of the tumor. If these distributions overlap too much, data will not be as useful for classification.

The overlap seems to be larger when comparing the standard deviation of the measurements, which means the variability of the data is similar in both cases. If we compare the overlap in mean values vs that of worst measurements, they seem to mirror each other, however, the overlap slightly decreases in worst measurements, indicating that these could be more useful for classification than mean and standard deviation values.

Below the diagonal we see the specific correlations. Some strong correlations show that features are connected, for example, radius, perimeter and area are all measures of size, and thus they are strongly correlated. Texture and symmetry do not correlate with any other feature. Another group of features with a smaller and more dispersive correlation is that of smoothness, compactness, concavity and concave points. Fractal dimension does not correlate with any other feature except for smoothness, compactness and concavity, but it is a small correlation and can only be observed in the set of worst measurements.

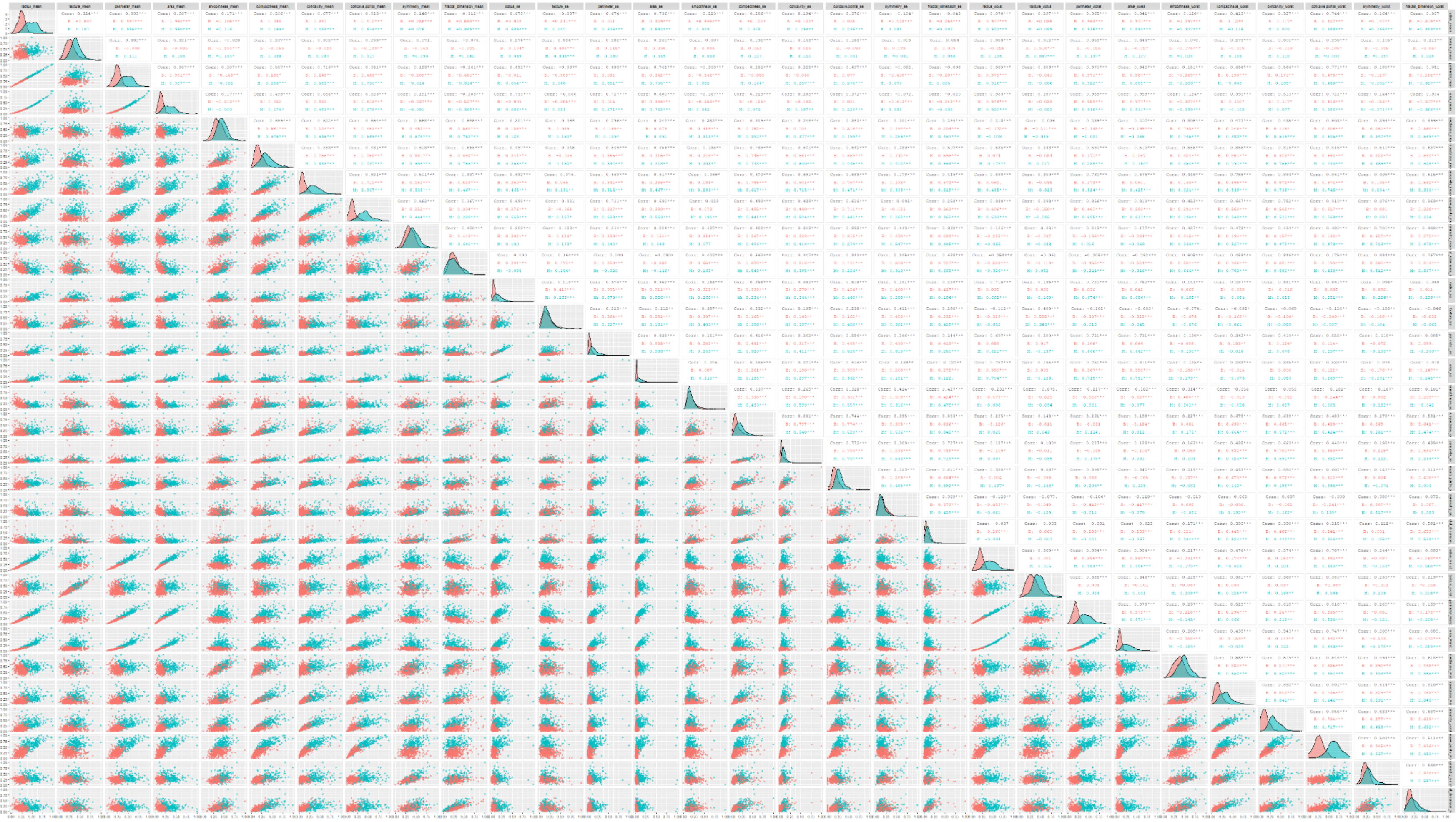
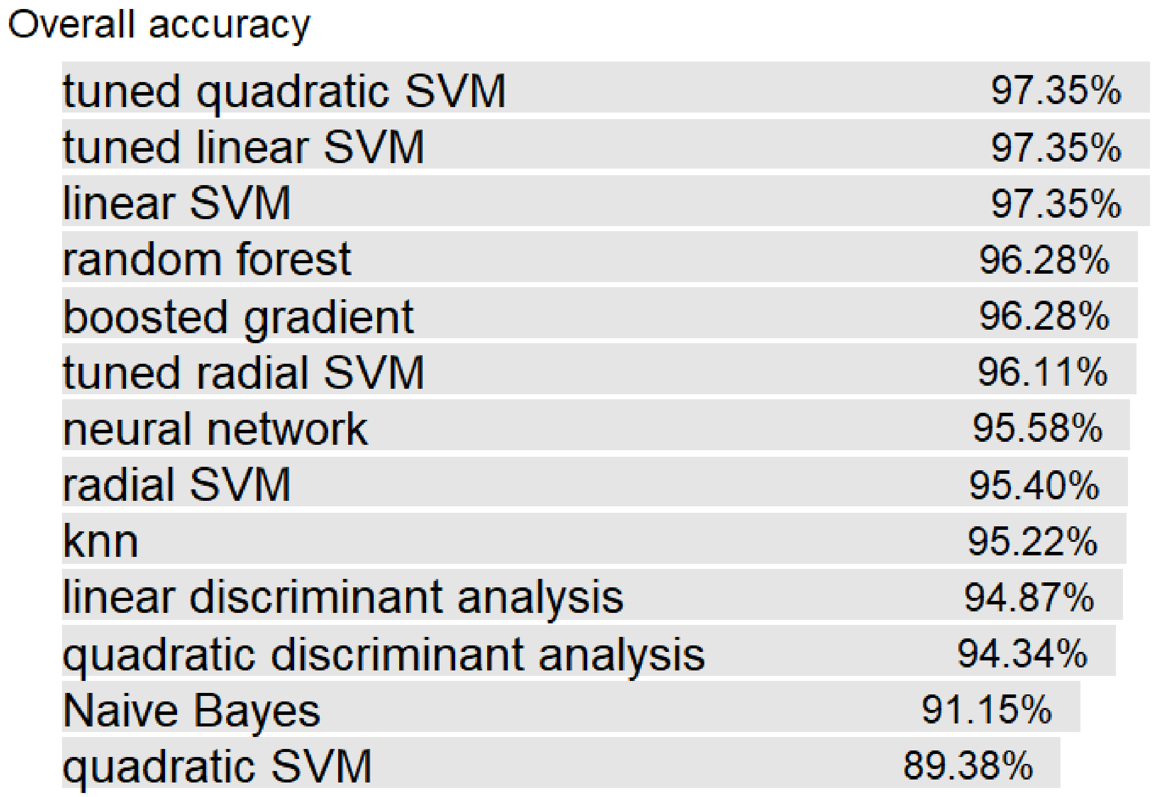


Figure 1. Correlogram of all 30 variables using diagnosis to show class membership. Blue = malignant; Red = benign. A higher resolution version is generated with the accompanying R script.

Given these observations, for the scatterplots showing output results I decided to use the worst measurements for one feature from each of the dominant groups: worst measurement of area, from the group of features related to size, and worst measurement of concave points, from the group of features related to shape.



Overall accuracy

tuned quadratic SVM	97.35%
tuned linear SVM	97.35%
linear SVM	97.35%
random forest	96.28%
boosted gradient	96.28%
tuned radial SVM	96.11%
neural network	95.58%
radial SVM	95.40%
knn	95.22%
linear discriminant analysis	94.87%
quadratic discriminant analysis	94.34%
Naive Bayes	91.15%
quadratic SVM	89.38%

Figure 2. Overall accuracy of all 13 models tested. SVM = support vector machine; knn = k nearest-neighbors.

Figure 2 shows a typical scoreboard of machine learning algorithms, which plot overall accuracy. Based on this figure, one could choose any of the SVM methods at the top as the winner. However, for the case of breast cancer diagnosis, misdiagnosing a malignant tumor as benign may incur high expenses for the patient who would leave their illness unattended, leading to a worsened condition that in time will require more expensive treatments and disable the patient to work for longer. Compared to the costs of misdiagnosing a patient as having a malignant tumor when theirs is benign, it would be convenient to put more weight on those models with less positive misdiagnoses.

In figure 3 I show a scoreboard of the accuracy of malignant cases misdiagnosed as benign. The results are surprisingly different, with many models on the lower half in figure 2 occupying the top of figure 3. The quadratic SVM model, which scored last in overall accuracy, did not misdiagnosed any malignant tumor as benign. Simpler methods like linear discriminant analysis and k nearest-neighbors got second and third place.

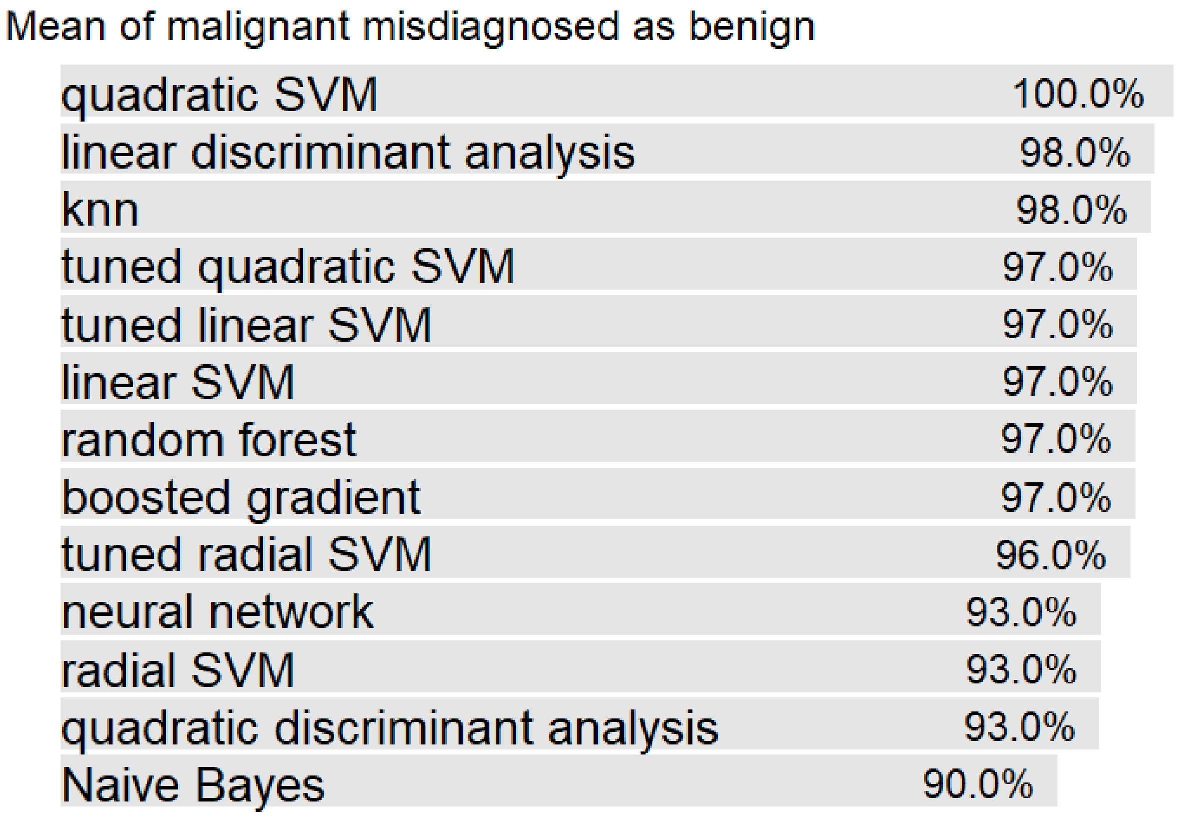


Figure 3. Mean accuracy of malignant cases. SVM = support vector machine, knn = k nearest-neighbors.

Discussion

The stark contrast between overall accuracy and proportion of false negatives for many machine learning models applied to the same case study was briefly shown in this report. This difference in performance when changing evaluation criteria suggests that a more thoughtful approach must be taken when scoring machine learning models. Overall accuracy is not necessarily the best criteria to choose one model over the others, and the specific needs of each case study must be considered when making a decision. This careful procedure is particularly relevant for interdisciplinary applications where tasks are separated, e.g., when the mathematical tasks are left with the statistician and the clinical assessments with the medic.

The point of using machine learning algorithms to assist in breast cancer diagnosis would be to automatize the diagnosis procedure, or most of it, making the work of the physician more efficient. Recurrent cases of misdiagnosis with expensive consequences must then be avoided. One way could be to use these algorithms to separate the most certain cases from the most doubtful.

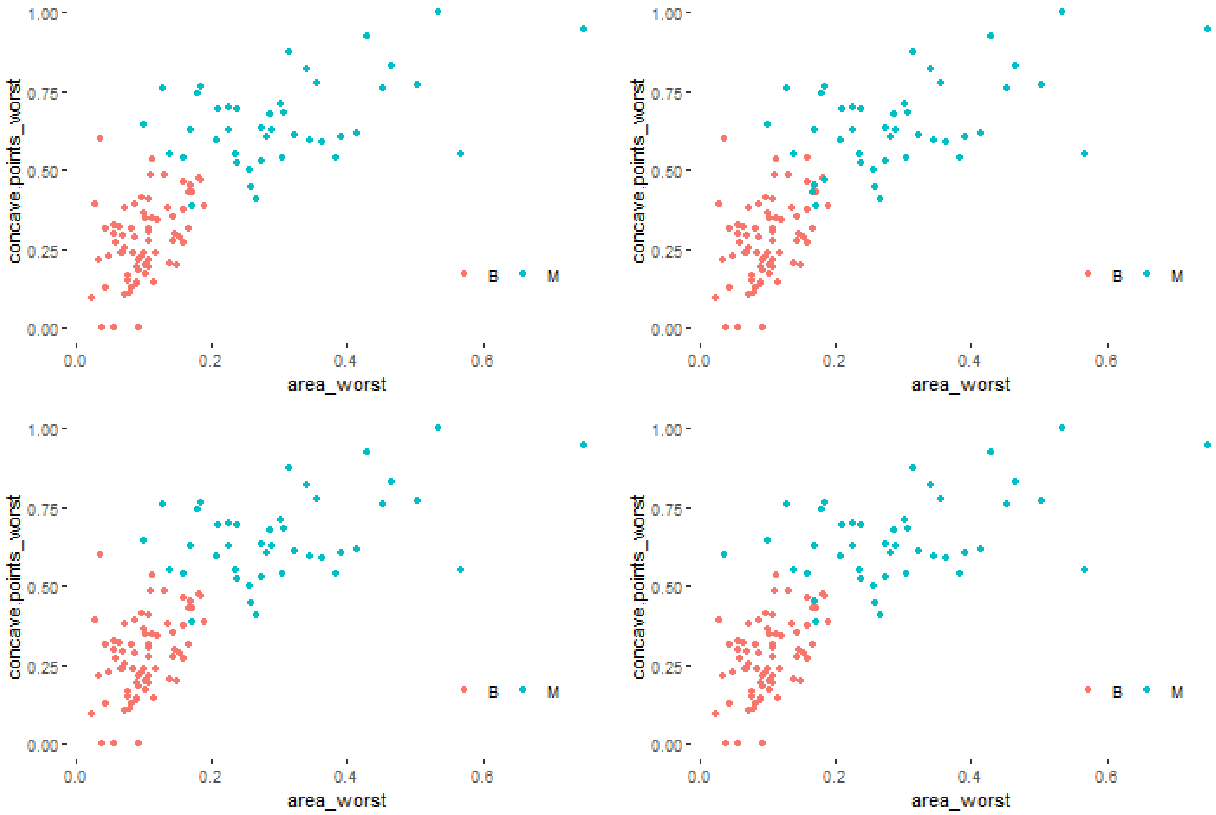


Figure 4. Scatter plots of actual diagnosis (left) and predictions made with quadratic discriminant analysis (top right) and linear support vector machine (bottom right). B = Benign (red), M = Malignant (blue).

As shown in figure 4, most of the data clouds can be easily separated into typical benign and malignant regions. By using probabilities, one can automatize the process to flag those observations that fall within the doubtful region, in other words, those observations that cannot be predicted right 100% of the time. This would limit the time required from the physician to the difficult cases and reduce the risk of misdiagnosis with a net efficiency gain in the diagnosis process.

Additionally, further research can estimate the potential costs associated with blind trusting of false negatives and false positives to find the ideal balance of accuracies expected from a machine learning algorithm i.e., the balance that minimizes the cost of the process

References

Barford, p. (no date). Machine Learning for Cancer Diagnosis and Prognosis [http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html]. Madison, WI: University of Wisconsin.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.