Predicting malignity of mammographic masses

Eduardo Calò

1 Problem description

Mammography is considered to be one of, if not the, most effective method for breast cancer screening available today. However, human mammogram interpretation leads to approximately 70% biopsies with benign outcomes. To help doctors in their decisions and reduce this high number of unnecessary breast biopsies, computer-aided diagnosis (CAD) systems have been proposed. These systems are usually created using Supervised Machine Learning (SML) techniques. In this report, I will present the methodology I followed and the results I achieved using SML on the Mammographic Mass Data Set.

In cases like the one we are dealing with, the problem that these SML algorithms have to solve is called *binary classification*. Several instances are collected (in our case, 961 patients), and for each of them, values for several features or attributes (in our case, 5 attributes described later in the report) are recorded. Eventually, a category is manually given to each sample (in our case, only 2 outcomes are possible: benign or malignant), and the algorithms have the task to learn which values correspond to which category, and predict the category of new unseen data. Specifically in our case, the algorithms have to predict the severity (benign or malignant) of a mammographic mass lesion, from BI-RADS (Breast Imaging-Reporting and Data System) attributes and the patient's age.

2 Dataset and features engineering

The dataset used for this experiment is taken from the UCI Machine Learning Repository[1] and consists of 961 cases (516 benign and 445 malignant), which have been collected at the Institute of Radiology of the University Erlangen-Nuremberg during the period of 2003 to 2006. The attributes recorded are the following:

- **BI-RADS assessment**: general assessment given by specialists, ranging from 1 (definitely benign) to 5 (highly suggestive of malignancy). This is considered to be a non-predictive attribute, which means that it is not useful for the algorithms.

- **Age**: age in years of the patient taken into consideration, ranging from 18 to 96.

- **Shape** (BI-RADS attribute #1): shape of the mass. Values are integers from 1 to 4: round=1, oval=2, lobular=3, irregular=4. This is a nominal feature.

- **Margin** (BI-RADS attribute #2): type of margins of the mass. Values are integers from 1 to 5: circumscribed=1, microlobulated=2, obscured=3, ill-defined=4, spiculated=5. This is a nominal feature.

- **Density** (BI-RADS attribute #3): the density of the mass. Values are integers from 1 to 4: high=1, iso=2, low=3, fat-containing=4. This is an ordinal feature.

- **Severity**: benign=0 or malignant=1. This is the outcome, our category to be predicted.

The original dataset was quite dirty, thus some pre-processing was needed. First of all, since **BI-RADS assessment** is considered to be a non-predictive attribute, the relative column was discarded, and the feature not taken into account for prediction. Subsequently, I discovered that the values of this dataset are ill-formatted and some of them are missing. This is a common problem which affects most of datasets containing real-world data. Before going on, I got rid of the instances containing missing values and fixed the rest. After these changes, 831 samples (from the original 961) were kept. Some statistical analysis show that all the values are within the features' expected ranges presented above, and the distribution benign/malignant is balanced (with 428 benign and 403 malignant instances). Moreover, statistics show that an *irregular* **Shape** is synonymous of malignity, while *circumscribed* **Margin** is synonymous of benign. As far as correlations between features are concerned (see Figure 1), we can notice an higher incidence of malignity among higher ages (with a peak around 65 years-old).

After this step, before going for the actual choice of the algorithm, given the different nature of the attributes (quantitative, nominal or ordinal), some other normalization over the data had to be done, in order to allow ML algorithms to work properly.

[1] `https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass`

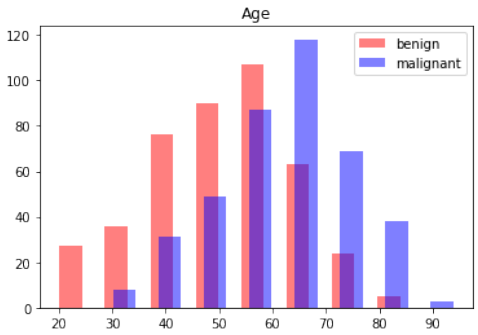

Figure 1: ”Age” and ”Severity” correlation.

3 Choice of the method

As common practice in ML, first of all, I split the dataset in training set and test set, in order to avoid that the same samples be used for both training and testing. I used the training set for choosing the best algorithm and finding the best combination of its parameters, while keeping the test set only for the final results.

In order to select the best algorithm, I compared the performances of the major ML algorithms suitable for our type of problem. From this preliminary comparison, two models slightly outperformed all the others. In order to achieve an even better performance, I focused on adjusting the parameters of these algorithms. It turned out that the best model is **Logistic Regression**. Even though this is one of simplest ML models, it often outperforms more sophisticated ones, like in our case. Finally, I tested this best model using the test data kept aside at the beginning.

4 Interpretation of the results

The problem we had to solve was to reduce the high amount of unnecessary prescribed biopsies, due to inaccurate human interpretation of mammographic results. With human interpretation only, around 70% of biopsies turn out to be unnecessary. This means that cases of false positives (detection of malignity when it is not present) are frequent.

The results that this classifier achieves on this problem are positive. It shows an overall accuracy of 84% on the unseen data of the test set. In particular, as it can be seen from Figure 2, this model only for the 8% of cases would predict a biopsy when indeed it is not necessary. There are indeed only 20/250 false positives. However, it also predicts 19/250 (7.6%) false negatives, considering healthy a mass which is indeed malignant. That is why machines always need to be used as support for diagnosis, never replacing completely the experts' opinions.

This model is also really efficient in predicting true positives (prediction of malignity when mass is malignant) and true negatives (prediction of benign when mass is benign), achieving high scores in these tasks, as well.

All in all, our goal has been accomplished. Using this model, a drastic reduction of the amount of unnecessary biopsies can be achieved. This classifier can be of aid and give great support to doctors and specialists in their job.

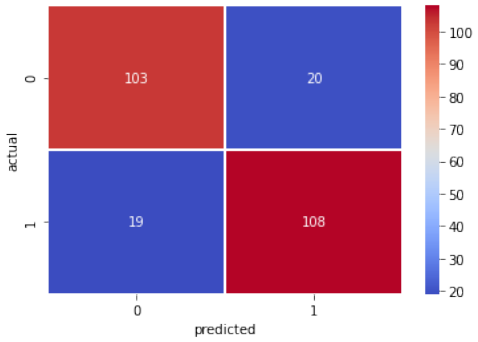

Figure 2: Final test results.