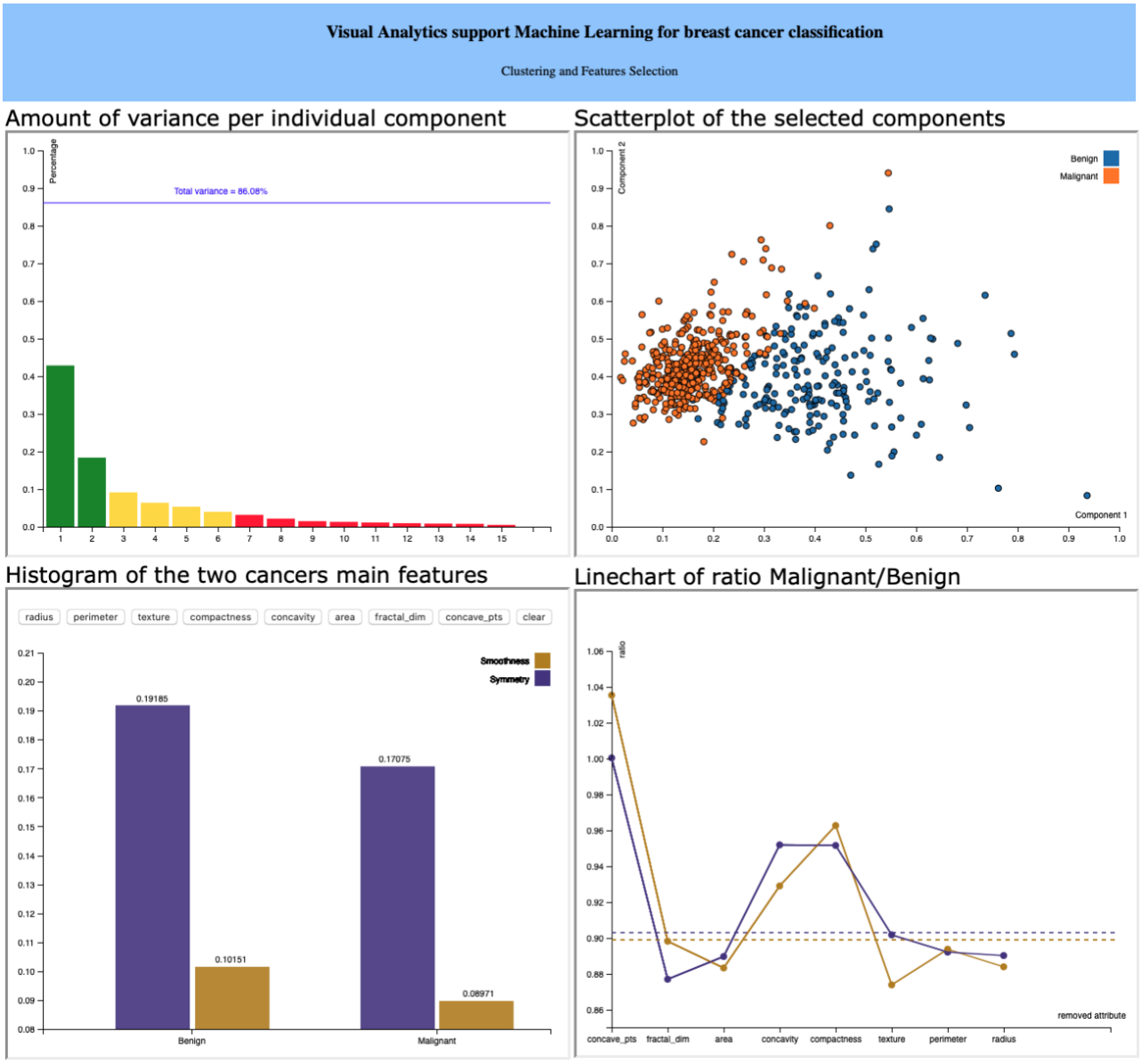
Support "Machine Learning" visual analysis

Luigi R. Zollo 1390138, Michele Sorrentino 1609244

ABSTRACT

The paper describes the environment, design choices, techniques adopted and the results obtained of the Visual Analytics project in order to support machine learning. The pourpose of the analisys is first to verify if the data is well-suited for some kind of classification model and then extract the “best” features for classification.



1. INTRODUCTION

Nowadays the need to classify as soon as possible the nature of breast's cancer, in order to find the right therapies, growed up. In this context, a “Machine Learning” based solution is a powerful tool that helps us efficacely to obtain this purpose. In particular is important to know how to use the algorithms to optimise the results’ accuracy.
Our analysis aims to find graphically some relevant directives that a user should follow to categorize new breasts cancer tempestively.

2. DATASET

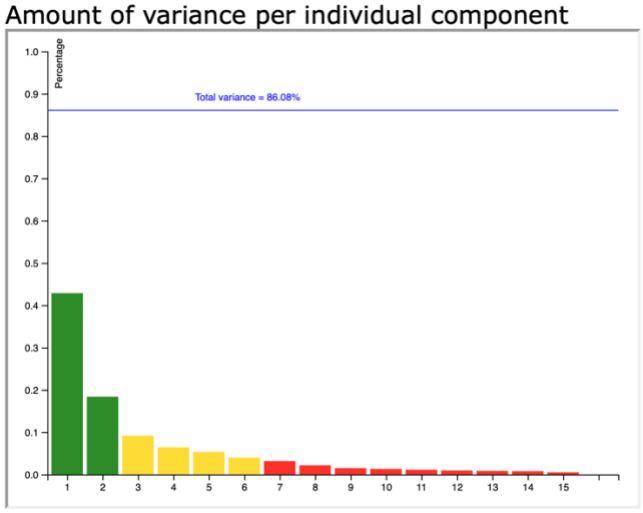
The dataset of the project is composed by two CSV file: in the first one there is the description of the features and in the second all the values. Each value is computed from a digitalized image of a fine needle aspirate (FNA) of a breast mass. The data file contains 569 tuples and 32 attributes structured in this way: the first 2 fields are the **ID** and the **diagnosis** and after, for each feature, is computed the mean, the standard error and the "worst" (e.g. in the column 3 there is the mean radius, the radius SE in 13 and worst radius in 23). The features are: **radius** (mean of distances from center to points on the perimeter), **texture** (standard deviation of gray-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).

3. CLUSTERING ANALYSIS

In this analysis we decided to used PCA, implemented in a python programm, to reduce the 30 components of the dataset in order to visualize them in a scatterplot.
We divided the analysis in 2 charts:

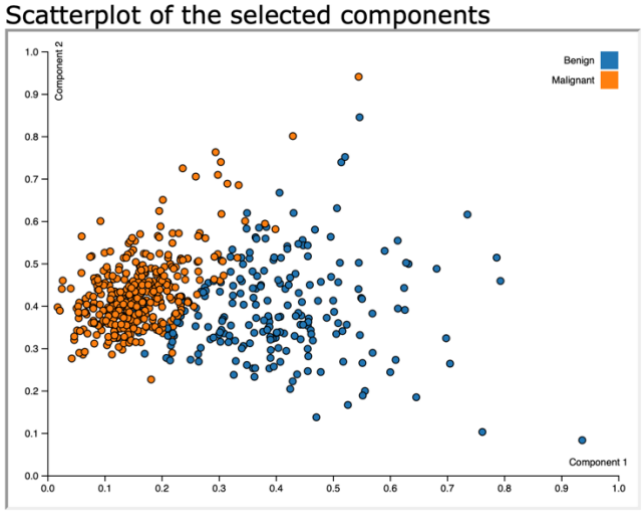
- Barchart: the value of each bar represent the amount of variance per component obtained by PCA. An “onClick” listener is positioned on the bar with 2 purpose: compute the total variance (represented as a blue line in the chart) of the selected components and draw the scatterplot with the first 2 component selected. The components that will be represented on the scatterplot are marked with colour green, the components that are use just to compute the total variance instead are marked with colour yellow.

Amount of variance per individual component



- Scatterplot: the chart is populated when 2 components are chosen in the barchart. Two different colors are used to visualize the clusters obtained (benign and malignant)

Scatterplot of the selected components

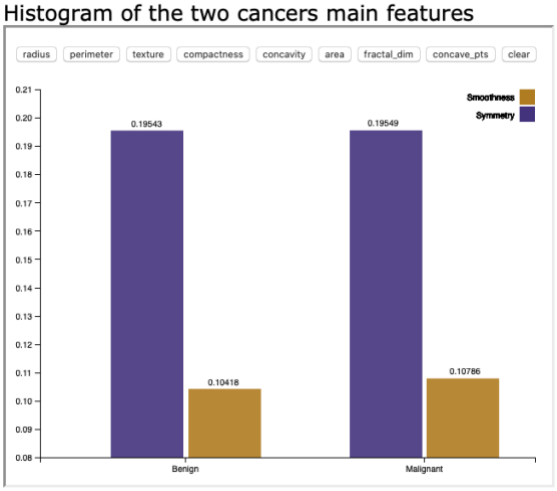


4. FEATURES EXTRACTION

The second analysis aims to find the “best feaatures” to put as input of a "Machine Learning" algorithm.
In order to do that we define 2 “reference features” that are significant for the cancer's class and we studied the behaviours of the other features under certain conditions.
The 2 features are: “Symmetry” and “Smoothness”; the more the cancer is smooth (symmetric) the more is benign.
To carry out this analysis we decided to use again 2 charts:

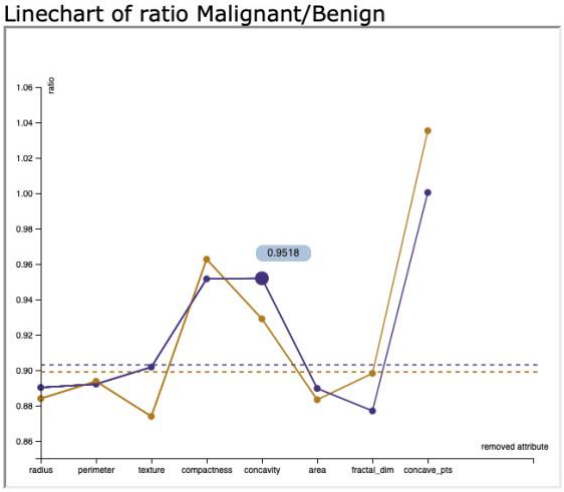
- Histogram: composed by 4 column that represent the mean of all tuples of the attributes “smoothness” and “symmetry” for benignant and malignant cancers. In the top of the chart there are some button that allows us to recompute the mean (of “smoothness” and “symmetry”) without the tuples that have value bigger than a certain threshold for the selected attribute. Such threshold corresponds to the mean of all tuples of the selected attribute.

Histogram of the two cancers main features



- Linechart: contains the informations about the variation of the 2 reference ratios. When a button of the above graph is pushed, a dot with value of the recomputed ratio is added. When there are at least 2 dots, all of them are connected through lines. Each dot has a “mouseOver” that show his exactly value. This procedure is executed for both ratios.

Linechart of ratio Malignant/Benign



5. FINAL CONSIDERATION

From both visual Analysis there are some interesting results that are immediately appreciable:

- The amount of variance cover by the first 6 components cover almost the 90% of the total variance so it possible to used only these components (instead of the 30 total attributes) as input of some classifier algorithm. Moreover, from the scatterplot is visible that the dataset is well-suited for a classification model.
- There are some features that deviate greatly from the reference ratio. Indeed if we run the classifier with all features we obtain an accuracy of the 92.6%

```
Dataset: WDBC
Number of attributes/features: 10
Number of classes: 2 ['Benigno' 'Maligno']
Number of samples: 569


Size of training set: 379
Size of test set: 190


Accuracy 0.926
              precision    recall  f1-score   support

     Benigno      0.924     0.871     0.897        70
     Maligno      0.927     0.958     0.943       120

    accuracy                          0.926       190
   macro avg      0.926     0.915     0.920       190
weighted avg      0.926     0.926     0.926       190

Confusion Matrix
[[ 61   9]
 [  5 115]]
```

Instead if we run the same classifier after the elimination of some “unnecessary” features, the accuracy of the algorithm grows.

```
Dataset: WDBC
Number of attributes/features: 6
Number of classes: 2 ['Benigno' 'Maligno']
Number of samples: 569


Size of training set: 379
Size of test set: 190


Accuracy 0.942
              precision    recall  f1-score   suppo

     Benigno      0.984     0.857     0.916
     Maligno      0.922     0.992     0.956       1

    accuracy                          0.942       1
   macro avg      0.953     0.924     0.936       1
weighted avg      0.945     0.942     0.941       1

Confusion Matrix
[[ 60  10]
 [  1 119]]
```

More precisely the eliminated features are: “compactess”, “concavity”, “concave points” and “area” (area or radius visually are very similar but it is not possible delete both because it would get lost all cancers' size references).