RESEARCH

Introduction to the profile areas of data sciences: project 2

Eva Aßmann* and Paul Vogler

*Correspondence:
eva.assmann@fu-berlin.de
Institute of Mathematics and Informatics, Berlin, Takustr. 9, Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Goal of the project: Utilize feature selection and feature extraction to increase the correctness and speed of a breast tumor diagnosis using machine learning classifier.

Main results of the project: The Random Forest and Support Vector Machine classifiers were very accurate at predicting the target groups, with an accuracy of about 96% using feature selection and slightly lower using feature extraction.

Personal key learnings: We learned the difference between feature selection and extraction and applied RF and SVM for the first time.

Estimated working hours: 8

Project evaluation: 1

Number of words: 1234 words

Suggestions for improvements of the project: None

Scientific background of the project

Unlike a full biopsy, fine needle aspirations (FNA) introduced a way to examine small amounts of tissue from a tumor without an invasive surgical procedure. Diagnosis based on FNA required the integration of characteristics of individual cells and important contextual features. However, the diagnosis still depends on the expertise of the physician and thus remains highly subjective. Furthermore, many different features are thought to be correlated with malignancy which increases the complexity and effort of the diagnosis process [1].

Goal of the project

This project aimed at using different feature combinations and machine learning techniques in order to increase speed, correctness and objectivity of the breast tumor diagnosis process.

1 Description of the data

The implemented methods described below were based on the Breast Cancer Wisconsin Diagnosis Data Set. This database contained bio-medical attributes extracted from digitized images of FNA breast tissues for 569 patients. For a selected set of cell nuclei from each image, the shapes were analysed resulting in ten derived nuclei features. The mean value, largest value and standard error of each feature were computed over the range of isolated cells and stored in the database (Figure 1).

1 ID number
2 Diagnosis (M = malignant, B = benign)
3 The mean, standard error and largest value were computed for each feature:
  (a) Radius: The average of the radial lines from the nucleus centroid to the perimeter points
  (b) Texture: The variance in grey scale intensity of the pixels
  (c) Perimeter: The total distance between all perimeter points
  (d) Area: The number of pixels inside the perimeter
  (e) Smoothness: The difference in length between a radial line and the mean length of the lines surrounding it
  (f) Compactness: perimeter^2 / area
  (g) Concavity: Draw lines between non-adjacent perimeter points and measure the cell boundary lying inside the outline
  (h) Concave points: The number of concave portions of the contour
  (i) Symmetry: Find the major axis through the cell and measure the length differences of perpendicular lines to the cell boundary
  (j) Fractal dimension: Measure the perimeter with "coastline approximation"

Figure 1: For each patient a set of cell nuclei was isolated. 32 characteristics features were computed over the range of the nuclei and stored in a database.

2 Summary of the data statistics

Basic statistical information was gathered for the data. The resulting statistics were summarized for three selected attributes (mean radius, concavity and fractal dimension).

Mean, variance and standard deviation were calculated for the complete value range and additionally grouped according to the diagnosis field (M/B). Looking at the variable distributions, it was observed that mean nucleus radius and mean concavity of malign cells were bigger than for benign cells (Figure 2). The histogram plot for mean fractal dimension showed almost the same value range for benign and malign cells.

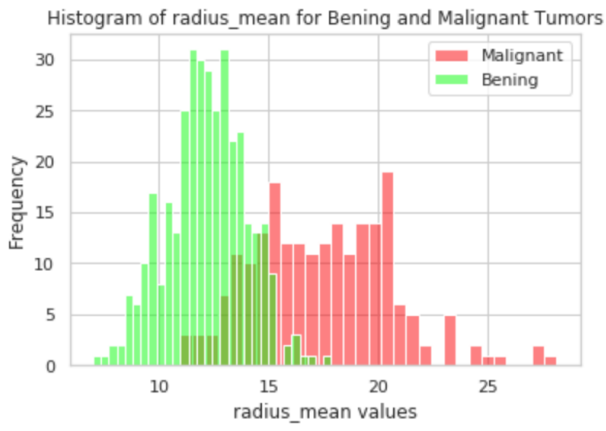


Figure 2: Value distribution for malign and benign radius_mean.

For all three attributes some potential outliers were located around the edges of the variable distributions. None of the outliers were located strikingly outside of the defined quartile.

Next the effect size was analysed for the three attributes. The effect size quantifies the difference between two groups (malign and benign) regarding a feature. According to Cohen, an effect size of d= 0.2 is small, d = 0.5 is a medium effect size and d = 0.8 large effect size. Mean radius (d=2.2) and concavity (d=2.0) showed a very large effect size indicating that malign and benign cells differ distinctly in these features. Meanwhile the effect size of the mean fractal dimension was rather small (d=-0.026).
The pairwise correlation was analysed between all attributes. Overall it was observed that the cell area measures (area, perimeter and radius) strongly correlated, as did the concavity measures (concavity, concave_points). Some of the standard error measures (e.g. smoothness, symmetry and texture) showed almost no correlation with most other variables (Figure 3).

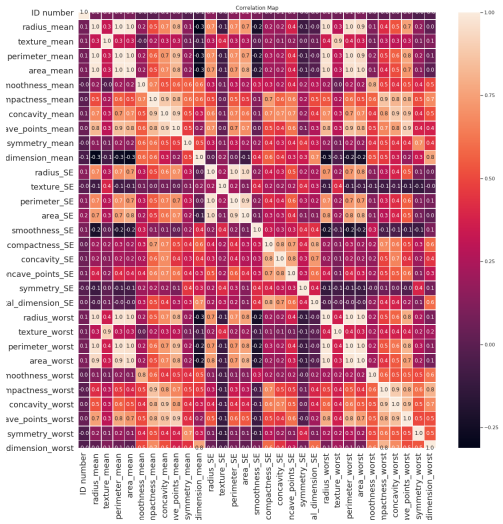


Figure 3: Pairwise correlation plot for all data set features.

To assess how good the values of each attribute could be separated according to diagnosis group, a swarmplot was plotted for each attribute's normalized values. In particular, fractal dimension and symmetry seemed to be difficult to cluster, while the values of the area and concavity attributes were observed to be separable into malignant and benign more clearly.

3 Description of the best features and how they have been determined

Feature selection was performed in order to determine the most important features in the data set that would yield the best classifier for breast tumor diagnosis. Feature selection and removal was decided based on correlation: From a set of 100% correlated features we kept only one feature. The 100% correlated feature sets were (radius_mean, perimeter_mean, area_mean, radius_worst, perimeter_worst) from which area_mean was selected, because the corresponding swarm-plot showed a good group decomposition (Figure 4a).

In the second group of 100% correlated features (radius_worst, area_worst, perimeter_worst) area_worst was selected, but with all three swarm-plots looking similar the choice was empirical (Figure 4b). For the third group (area_SE, radius_SE, perimeter_SE) all three swarm-plots showed a good decomposition and area_SE was selected empirically. For the last group of 100% correlated features (area_mean, area_worst) area_mean was selected since it appeared to be more distinct regarding the two diagnosis groups (Figure 4c). The remaining variables from all four groups were removed from the database ('perimeter_mean', 'radius_mean', 'radius_worst', 'perimeter_worst', 'radius_SE', 'perimeter_SE', 'area_worst') resulting in the reduced data set.

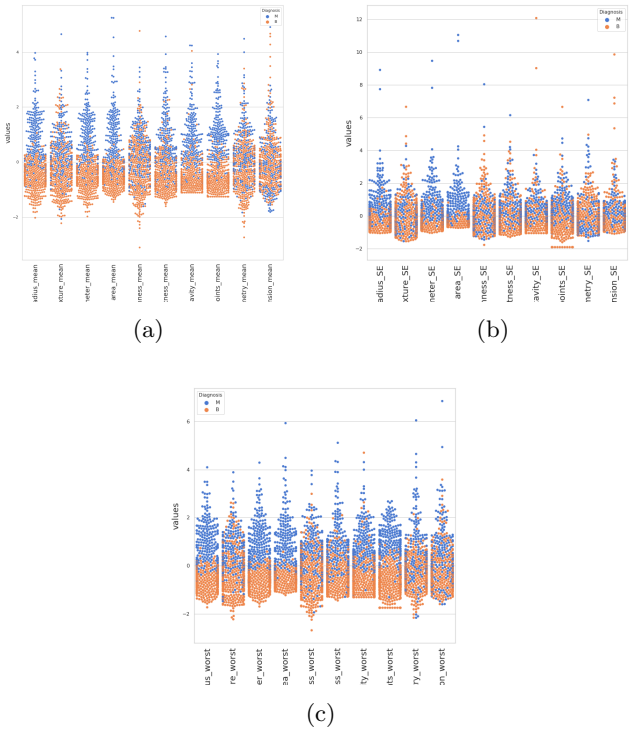


Figure 4: a) Swarm-plot for mean attributes, b) Swarm-plot for worst attributes, (c) Swarm-plot for standard error attributes.

In a second approach, Principal Component Analysis was used for feature extraction. Before performing PCA, the data needed to be normalized for better performance. The complete data set was split into 70% training and 30% test data. The training and test data sets were normalized by the difference of their respective maximum and minimum value. A PCA was performed on the training data with all variables kept in the set. To estimate the number of components needed to describe the data, the explained variance ratio was plotted for all components. It was observed that with already three components only close to 10% of the variance could be retained (Figure 5). Two PCA models were fitted and the data was reduced to 3 and 4 dimensions, respectively.

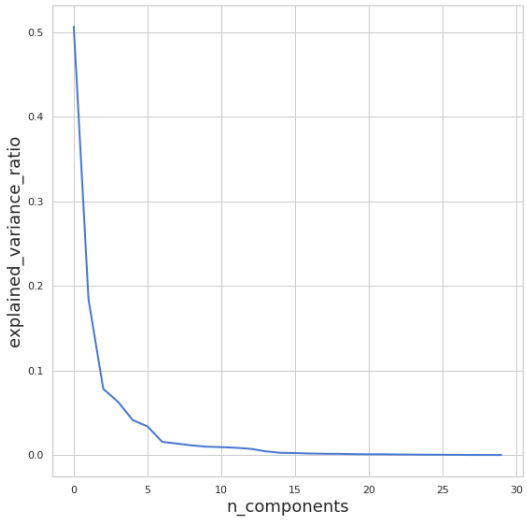


Figure 5: Explained variance ratio for every number of principal components.

4 Description and comparison of the classifier methods (RF and SVM)

For both feature selection and extraction two different classifiers for breast tumor diagnosis were trained respectively.

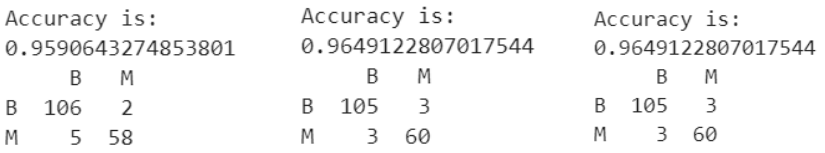
The reduced data set was split into 70% training and 30% test data. A Random Forest (RF) classifier was trained and then evaluated on the test data. The same data was split into 70% training and 30% test data again. A Support Vector Machine (SVM) was trained and then evaluated on the test data.

The complete 3D and 4D PCA-fitted data sets were each split into 70% training and 30% test data. Two RF classifiers were trained using ten estimators and then evaluated on the respective test set. The same two PCA-fitted data sets were split into 70% training and 30% test data again. Two Support Vector Machines were trained and then evaluated on the respective test set.

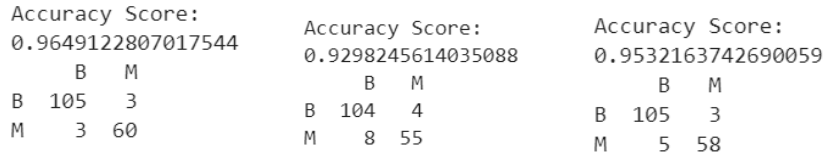
5 Results

The RF classifier and SVM that were trained on the reduced data set yielded 95.9% and 96.49% accuracy, respectively (Fig. 6a6d). The feature selection helped achieving good accuracy values at predicting malignant and benign data sets. The RF classifier and SVM that were trained on the data set extracted by 3D PCA yielded 94.7% and 93% accuracy, respectively (Fig. 6b, 6e). The RF classifier and SVM that were trained on the data set extracted by 4D PCA reached 96.5% and 95.3% accuracy (Fig. 6c, 6f).

The classifiers based on the 3D PCA data showed the worst classification accuracy.



(a) Feature Select RF (b) 3D PCA RF (c) 4D PCA RF



(d) Feature Select SVM (e) 3D PCA SVM (f) 4D PCA SVM

Figure 6: RF: Random Forest, SVM: Support Vector Machine

The RF classifier trained on the 4D PCA data and the SVM based on the reduced data set yielded the highest accuracy values (both around 96.5%).

6 Discussion

This project was good example for the typical tasks of a data-scientist: We got data from experts and a task (e.g. explain the data to us, get fancy plots and numbers, prepare ML project). The basic statistic analyses needed to be run to get to know all statistical characteristics of the data. The data had to be prepared for the ML step. Different approaches were explored to select features for different ML methods in order to enable a good accuracy and performance.

References

[1] Street W. Nick et al. "Nuclear Feature Extraction For Beast Tumor Diagnosis" *IS&T/SPIE 1993 International Symposium on ELectric Imaging: Science and Technology*, vol 1905 861-970