Lluís Giménez Gabarró & Pablo Pérez Cano 17 Oct 2023

Statistical Learning. Report Lab 2 Intro SLT

Brief summary

The ML models described in this practicum are used in the biomedical context to classify patients with prostate cancer, benign prostatic hyperplasia, and control subjects. The data consists of 654 mass spectra from 327 subjects. The analysis involved replicate averaging and log transformation preprocessing steps and an unsupervised feature filtering was applied before implementing k-nearest neighbors (k-NN) and Nearest Centroid Classifier. The k-NN method underwent parameter optimization with a 2-fold cross-validation. We found that k-NN with optimized parameters and feature selection significantly improved classification accuracy and the 10-fold yielded significant better results than 2-fold CV.

Introduction and Motivation

Prostate cancer and benign prostatic hyperplasia (BPH) are two prevalent conditions that affect millions of people. Early and accurate diagnosis of these conditions is highly important to ensure an effective intervention. Therefore, emerging technologies, such as mass spectrometry, have opened new opportunities for disease detection and classification.
In the pursuit of more precise diagnostic tools, we have performed a machine learning analysis of MS data, aiming to discriminate between patients with prostate cancer, patients with BPH, and control subjects.

Our analysis begins with a preprocessing protocol, involving replicate averaging and log transformation. To evaluate the predictive potential of our models, we partition the dataset into a training subset and an external validation subset, ensuring that both subsets maintain a balanced representation of subjects from each condition group. Regarding our classification methodologies, we employ the k-nearest neighbors (k-NN) algorithm in the initial analysis, opting for a preliminary k value without optimization. We compute the classification rate and quantify the associated uncertainty, which forms the baseline for our subsequent investigations. A Nearest Centroid Classifier is also employed to provide a distinct perspective on the classification task, affording a comparative evaluation.

Furthermore, we performed a parameter optimization for the k-NN algorithm, searching for the most suitable k value using cross-validation techniques. This process leads to the identification of an optimal k value and underscores the statistical significance of its selection in comparison to other candidate values. These findings illuminate the importance of parameter tuning in enhancing model accuracy.

In an effort to reduce the dimensionality of our feature space and optimize the classification process, we implement a variance feature filter to eliminate the uninformative features of the data, thereby simplifying the model while maintaining its predictive power.

Data

Our dataset is Prostate2000Raw and can be obtained from the ChemometricsWithR R package.

This data was made public in papers [1] and [2] that can be found at the references.

All the data from the dataset was generated and measured from prostate tissues using Surface-Enhanced Laser Desorption/Ionization (SELDI) Mass Spectrometry. It contains 654 mass spectra from 327 patients (each subject has 2 replicates). All patients are divided into 3 groups; patients with prostate cancer (pca), patients with benign prostatic hyperplasia (bph), control subjects (control). They took blood and serum samples from patients.

Regarding the dimensionality of the data, it has 3 different classes so it is a multiclass dataset with 327 samples and 10523 features (points of spectra).

Methods

Replicate averaging and log transformation

a. Preprocessing steps
b. DOI: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/log , base package
c. The preprocessing steps refers to everything you must do before ML, which is application dependent. Thus, in our case we have done replicate averaging (because we had two replicates per subject) and log transformation of the data to obtain a gaussian distribution.

Variance feature filter

a. Unsupervised feature filtering
b. Own programing. Attached R file submitted.
c. It is used to remove uninformative features from the analysis. The procedure consists in computing the variance of all features, sorting them and removing the features that explain a low variance of the data. In our case, we have optimized this threshold.

Nearest-Centroid Classifier

a. Classification algorithm
b. DOI: 10.1016/j.neuroimage.2005.06.017 , class package
c. The Nearest-Centroid Classifier is a classification algorithm used because of its simplicity and efficiency, especially when the classes are well separated. It is based on the following discriminant function, which will output a large value when the distance between the centroid and the feature vector is close (feature will correspond to the class that has the nearest centroid):

$$g_i(x) = -(x-\mu_i)^T(x-\mu_i)$$

k-Nearest Neighbors

a. Classification algorithm
b. DOI: 10.1109/2.78568 , caret package

c. It is a more complex classification algorithm used to handle complex decision boundaries, such as when the class-conditional densities are multi-modal and non-linearly separable. It operates on the principle that data points with similar features tend to belong to the same class. It consists in finding the "k" closest labeled examples in the training dataset and assigning the new feature vectors to the class that appears most frequently within the k-subset. Note that it is computationally and memory expensive, so we should not use it when we work with a huge dataset.

Hold-out validation

a. Data partition
b. DOI: 10.1111/j.2517-6161.1974.tb01479.x , caret package
c. The Hold-Out Method divides the dataset into two portions: a training set and a testing set. The principle is to train the model on the training set and evaluate its performance on the testing set. It allows for the assessment of how well the model generalizes to unseen data based on a single random split.

k-Fold cross-validation

a. Validation technique
b. DOI: 10.1111/j.2517-6161.1974.tb01479.x , caret package
c. K-Fold Cross-Validation divides the dataset into k approximately equal-sized folds or subsets. The model is trained and tested k times, with each fold used as a testing set once. The performance is averaged over these iterations, reducing the impact of random variability. The principle is to assess the model's performance while ensuring that all data is used for both training and testing.

Results and Interpretation

Note that depending on the seed used to perform the analysis, we will obtain different results. In fact, it could happen that we obtain a bigger CR in the external validation than in the internal validation. This could happen because there is an element of randomness associated with the data partition process. It could happen that the data in the external validation is more similar to the training than the data in internal validation. Therefore, we get a better balanced classification rate (BCR).

The initial step in our analysis involved the preprocessing of mass spectra data. By implementing the replicate averaging, we reduced the impact of measurement variability within subjects. By implementing the log transformation we normalized the intensity distribution and highlighted major spectral peaks. This preparation of the data enhanced its suitability for subsequent machine learning analysis.

The workflow followed to implement a variance feature filter (unsupervised) consists in calculating the variance for each feature, sorting the obtained variances and visually assigning a threshold to remove the features that show no significant variability in the full dataset. We have optimized this threshold in the internal validation by calculating the variance for each feature, sorting the features, selecting the top 90% of features with higher variance (we could also optimize the percentage of features that we keep, but we won't do it) and perform the analysis again.

We repeat this process until we have only 5 features and we obtain the figure of merit (CR in this case) for each threshold. Then we choose the optimal threshold by simply selecting the threshold that gives a higher CR.

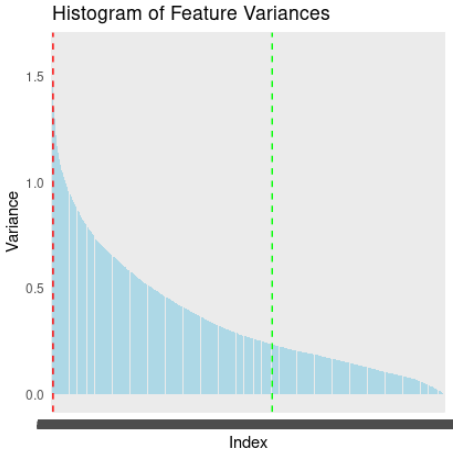

Figure 1. *Variance feature filter. Ordered histogram of the variance of each feature. The green vertical line corresponds to the threshold that we visually applied and the red vertical line corresponds to the optimized threshold.*

Regarding the comparison of the different k-folds, with k = 2 cross-validation we obtained an optimal value for k = 1 with a CR = 78% and a BCR = 45% . Moreover, we obtained a mean performance of the model of 64%. In contrast, with k = 10 cross-validation, we obtained an optimal value for k = 3 with a CR = 96% and a BCR = 64%. Also, for k = 10 cross-validation, we obtained a mean performance of the model of 71% a little bit better than k = 2 cross validation. Also, for the case of k = 10 cross-validation as we are increasing the size of the training set we will have a better representation of the data. Therefore, larger values of k provide more stable and reliable estimates but require more computational time and smaller values of k can be faster but may result in more variability in the estimates.
After optimizing the value of K using k-folds for k = 10 cross-validation, we found out that the optimum value for k was 3. By computing the test statistic for the difference in proportions so we check that the p-value was < 0.05, concluding that there is a statistically significant difference between both values of k.

In the hypothetical case we only had patients with prostate cancer and controls, the model would only classify the subjects into these two categories and it would not be able to predict if a sample has BPH or not. Obviously, if we tried to predict a BPH patient, it would incorrectly classify the BPH samples to one of these two conditions.

General conclusion

In light of the above, our study demonstrates the efficacy of ML in discriminating between the three conditions using mass spectra data. Optimizing model parameters, including k-value in k-NN and feature selection threshold, significantly enhances classification accuracy.

Bibliographic references

[1]. [B.L. Adam, Y.Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, Y. Yasui, Z. Feng, G.L. Wright, “Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men”, Cancer Res. 63, 3609-3614, (2002)]

[2]. [Y. Qu, B.L. Adam, Y. Yasui, M.D. Ward, L.H. Cazares, P.F. Schellhammer, Z. Feng, O.J. Semmes, G.L. Wright , “Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients”, Clinical Chemistry, 48, 1835-1843, (2002)].