Building a machine learning model to predict tumor types from RNA-seq gene expression data

Yorgo EL MOUBAYED, Aix-Marseille Université, Marseille, France

1. Introduction

This project aims to observe which features are most helpful in predicting different classes of tumors: BRCA, KIRC, COAD, LUAD and PRAD and to see general trends that may aid us in model selection and hyperparameter selection. The goal is to classify the tumors. To achieve this, a comprehensive comparison between machine learning classification methods to fit a function that can predict the discrete class of new input will be performed.

2. Materials and methods

2.1. Data preparation and preprocessing

2.1.1. Data import and preparation

This collection of data is part of the RNA-Seq (HiSeq) PANCAN dataset **[1]**. It is a random extraction of gene expressions of patients having different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD. Data was recovered from "The Cancer Genome Atlas (TCGA)" **[2]**. Research Network has profiled and analyzed large numbers of human tumors to discover molecular aberrations at the DNA, RNA, protein and epigenetic levels. The resulting rich data provide a major opportunity to develop an integrated picture of commonalities, differences and emergent themes across tumor lineages. The Pan-Cancer initiative compares the first 12 tumor types profiled by TCGA. Analysis of the molecular aberrations and their functional roles across tumor types reveals how to extend therapies effective in one cancer type to others with a similar genomic profile. The collection of data contains two separate datasets:

- ***data.csv*:** samples (instances) are stored row-wise and variables (attributes) of each sample are RNA-Seq gene expression levels measured by illumina HiSeq platform. A dummy name (gene_XX) is given to each attribute. The dataset contains 801 rows (samples) and 20532 columns (genes).

- ***labels.csv***: different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD relative to each sample.

2.1.2. Data exploration and preprocessing

Before building the model, we splitted the data into two parts: a *training set* and a *test set*. The training set is used to train and evaluate the model during the development stage. Then, the trained model is used to make predictions on the unseen test set. This approach gives a sense of the model's performance and robustness.

RNA-seq data is messy. It often contains noise from technical artefacts, batch effects, and other confounders. Before analyzing the data, it is necessary to assess and correct for as much of this unwanted variation as possible. When a large number of candidate variables are present, a dimension reduction procedure is usually conducted to reduce the variable space before the subsequent analysis is carried out. The goal of dimension reduction is to find a list of candidate genes with a more operable length ideally including all the relevant genes. Leaving many uninformative genes in the analysis can lead to biased estimates and reduced power. Therefore, dimension reduction is often considered a necessary predecessor of the analysis because it can not only reduce the cost of handling numerous variables, but also has the potential to improve the performance of the downstream analysis algorithms.

Principal component analysis (**PCA)**, is a statistical technique to convert high dimensional data to low dimensional data by selecting the most important features that capture maximum information about the dataset. The features are selected on the basis of variance that they cause in the output. The feature that causes highest variance is the first principal component. The feature that is responsible for second highest variance is considered the second principal component, and so on. It is important to mention that principal components do not have any correlation with each other.

A feature set must be normalized before applying PCA because the variance scale is huge. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results. Thus, the training time of the algorithm is reduced with a smaller number of features. Finally, the last point to remember before we start coding is that PCA is a statistical technique and can only be applied to numeric data. Therefore, categorical features are required to be converted into numerical features before PCA can be applied.

Now that the datasets are loaded and preprocessed the different machine learning classifiers[3] can be built.

2.2. Building and training the models

There are many models for machine learning, and each model has its own strengths and weaknesses. In the given datasets, the outcome variable has five sets of values (classes of tumors): BRCA, KIRC, COAD, LUAD and PRAD. Hence, classification algorithms of supervised learning are employed. Different types of classification algorithms in machine learning are available: logistic regression, nearest neighbor, support vector machines (SVM), kernel SVM, naïve bayes, decision tree algorithm and random forest classification. Using these different classification algorithms, the classifiers were built.

2.4. Evaluating the model’s accuracy

The classification accuracy method is employed to find the accuracy of the models. The classification accuracy is the ratio of number of correct predictions to the total number of input samples.

$$\textbf{Accuracy} = \frac{\textbf{Number of correct predictions}}{\textbf{Number of predictions made}}$$

3. Results and discussions

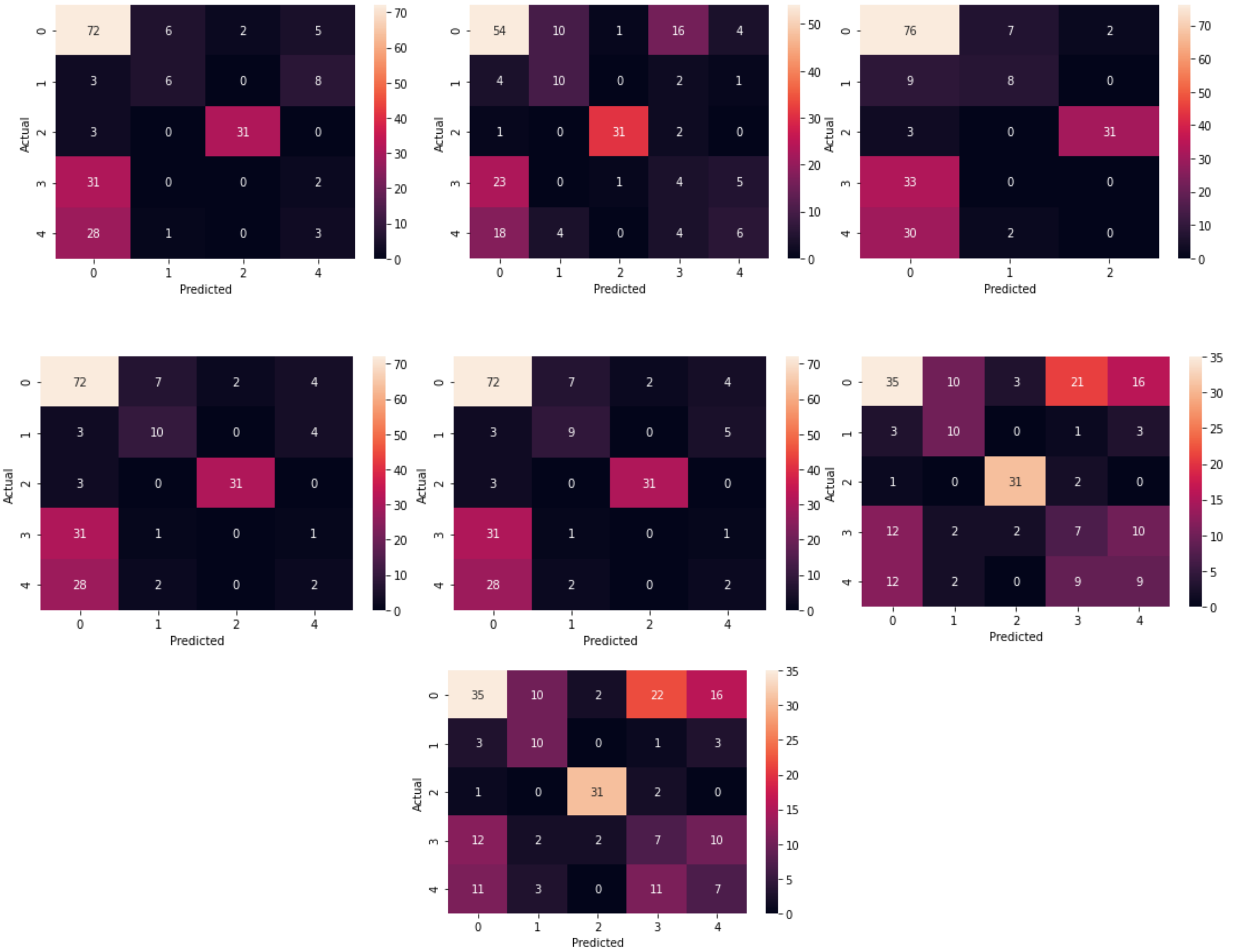

***Figure 1* - Confusion matrices of the different types of classification algorithms.** (A) Logistic regression, (B) Nearest neighbor, (C) SVM, (D) Kernel SVM, (E) Naive Bayes, (F) Decision tree algorithm, (G) Random forest classification. Correct predictions are shown on the diagonal. Total number of predictions is fixed to 201.

Classifier	Accuracy
Logistic regression	0.55
Nearest neighbor	0.52
SVM	0.57
Kernel SVM	0.57
Naive Bayes	0.57
Decision tree algortihm	0.45
Random forest classification	0.45

***Figure 2* - Given accuracy of the different types of classification algorithms.**

When compared to other models, SVM, kernel SVM and Naïve Bayes classifiers perform better than the others: results reveal that they have the better testing accuracy. Nonetheless, the given accuracies are still low. Hence, we still need to optimize these classifiers in order to get better results. The models have a low accuracy could be caused by overfitting. This happens because the models are trying too hard to capture the noise in the training dataset. By noise we mean the data points that don't really represent the true properties of the data, but random chance. Learning such data points, makes the model more flexible, at the risk of overfitting. A technique that helps in avoiding overfitting and also increasing model interpretability is regularization. This is a form of regression, that shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

4. Conclusions

In this study, we compared classifiers and showed how to train models and predict new samples. We should note that the model performance depends on several criteria, such as normalization and transformation methods, gene-wise overdispersions, number of classes etc. Hence, should the given performances be considered as a generalization to any RNA-Seq dataset? Neural networks are well known for classification problems, for example, they are used in handwritten digits classification, but the question is will it be fruitful if we used them for regression problems?

5. Supplementary information

A GitLab repository is available on https://gitlab.com/Yorgomoubaed/biostatistiques-approfondies It contains the source code and datasets with detailed explanations to setup and run simulations.

6. Refrences

1. 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.

2. Cancer Genome Atlas Research Network *et al.* **The Cancer Genome Atlas Pan-Cancer analysis projec**t. *Nat. Genet.* **45**, 1113–1120 (2013).

3. Pedregosa, F. *et al.* **Scikit-learn: Machine Learning in Python**.