Assignment 2: *Nested CV on Hepatitis C dataset*

Konstantinos Mexis

Department of Process Engineering, NTUA, Iroon Politechniou 6 Zografou, Athens, Greece

Abstract

Hepatitis C infection is an insidious disease and early diagnosis of the disease is not possible in many cases. Early diagnosis and treatment have an important place for the disease. The diagnosis of the disease can be made with the use of machine learning methods. In this study, a complete ML pipeline was created in order to classify future patients as Hepatis C patients or healthy blood donors. For this task, we built a *nested Cross Validation (nCV)* pipeline to compare systematically the expected performance of multiple classification algorithms on unseen data. The pipeline in this study can be used as an alternative method in the diagnosis of Hepatitis C-related diseases.

**Keywords**: Hepatitis C; machine learning; nested CV;

* 1. Introduction

Hepatitis C is a viral infection that causes the inflammation of the liver. It is caused by the hepatitis C virus (HCV), which is one of the most important global health problems in the world. Worldwide, 350,000 people out of 185 million patients infected with HCV die from diseases caused by HCV. HCV infections pose serious problems on a global scale. Since there is no vaccine yet to prevent HCV infection, it is essential to prevent infection. Early detection of patients and people at risk is critical to prevent the spread of HCV infection. The contribution of this study is to create a complete ML pipeline to process the collected data and classify successfully future patients.

* 1. Methods
     1. Dataset description

The data set used in this study is a part of the HCV data set in the UCI Machine Learning Repository. The given dataset consists of 204 rows and 12 features. The features are described in the following table:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Data type** |
| Age | Numerically it is the age value in years | int64 |
| Sex | Female=1, Male=0 | int64 |
| ALB | Numerical value of laboratory test data | float64 |
| ALP | Numerical value of laboratory test data | float64 |
| ALT | Numerical value of laboratory test data | float64 |
| AST | Numerical value of laboratory test data | float64 |
| BIL | Numerical value of laboratory test data | float64 |
| CHE | Numerical value of laboratory test data | float64 |
| CHOL | Numerical value of laboratory test data | float64 |
| CREA | Numerical value of laboratory test data | float64 |
| GGT | Numerical value of laboratory test data | float64 |
| PROT | Numerical value of laboratory test data | float64 |

The target variable is the *label* column. ‘label=1’ corresponds to a Hepatitis C patient (positive class) and ‘label=0’ corresponds to a healthy blood donor (negative class).

* + 1. Data pre-processing

Data pre-processing is very important for the correct operation and high performance of AI algorithms. In the given dataset, there were no missing data. The dataset was splitted into a feature matrix (X) of size 204x12 and a target matrix (y) of size 204x1. We also normalize the data (feature matrix X), in order to ensure that all of the features

* 1. Conclusion

The production of large scale metabolic kinetic models is hindered by the uncertainty of predicting the kinetic parameters and producing physiologically relevant and robust kinetic models. Using the ORACLE framework, we are able to generate large populations of kinetic models but the biggest percentage of them are not stable. In this work the uncertainty in the model analysis is reduced through the use of machine learning principles. Using machine learning classification and explainability techniques we were able to raise the stability index of the generated models. We were able to reach up to 97.7% stability on our generated models implementing simultaneously all the extracted rules which leads to postulate that it is better to constrain such systems as much as possible to get more feasible results.

References

Andreozzi, S., Miskovic, L., & Hatzimanikatis, V., 2016, iSCHRUNK – In Silico Approach to Characterization and Reduction of Uncertainty in the Kinetic Models of Genome-scale Metabolic Networks. *Metabolic Engineering*, *33*, 158–168.

Barakat, N., & Bradley, A. P., 2010, Rule extraction from support vector machines: A review. *Neurocomputing*, *74*(1–3), 178–190

Goix, N., Birodkar, V., Gardin, F., Schertzer, J.-M., JEONG, H., manoj kumar, Gramfort, A., Staley, T., la Tour, T. D., Deng, B., C, Pedregosa, F., Wu, L., Rokem, A., Jackson, K., & mrahim., 2020, *scikit-learn-contrib/skope-rules v1.0.1* (v1.0.1) [Computer software]. Zenodo