CSI5155: Assignment 2 report

Adrien Heymans

# Introduction

This report will discuss of the evaluation of 6 computational models over 5 different datasets, using techniques such as over-sampling, under-sampling, cross-validation, and hyper-tuning. The goal is to determinate whether there exist significant differences between those algorithms over different datasets, using Friedman’s test and the Nemenyi post-hoc test.

# Datasets

The experiment will use 5 different datasets, in details:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Names | Number of instances | Number of features | Instances of the positive class | Instances of the negative class |
| Dataset | 1884 | 12 | 1265 | 619 |
| DB1 | 2530 | 12 | 1265 | 1265 |
| DB2 | 1238 | 12 | 619 | 619 |
| Heart disease | 302 | 13 |  |  |
| Labor Negotiation | 54 | 11 |  |  |

The dataset “DB1” was built from “Dataset” by oversampling it using the SMOTE method.

The dataset “DB2” was also built form the dataset “Dataset” by under-sampling using the “Near-miss” method.

The “Heart disease” dataset was already balanced and did not need any extr workj to be cleaned.

The “Labor Negotiation” dataset needed pre-processing as there were some missing values, categorical or ordinal data and the data needed to be normalized.

# Computational models

For this experiment we will be using 5 different computational models, namely:

* Decision Tree
* Support Vector Machine
* K-Nearest Neighbor
* Gradient Boosting
* Random Forest
* Multi-layer-perceptron

Each of these models will be trained on the 5 different datasets using 10-fold cross-validation, as well as hyper-tuning to ensure the best results possible.

# Results

The table below describes the results obtained for each algorithm on the different dataset. The score represent the balanced accuracy on the positive and negative classes..

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Tree | Support Vector Machine | K-Nearest Neighbor | Gradient Boosting | Random Forest | Multi-layer Perceptron |
| Labor negotiation | 0.87 | 0.71333333 | 0.75 | 0.83 | 0.91333333 | 0.92666667 |
| Heart Disease | 0.95075269 | 0.95075269 | 0.95731183 | 0.95075269 | 0.97064516 | 0.98354839 |
| Dataset | 0.77403467 | 0.69260666 | 0.43147022 | 0.75013509 | 0.75017168 | 0.78505291 |
| DB1 | 0.80869565 | 0.71660079 | 0.50197628 | 0.76679842 | 0.82213439 | 0.80118577 |
| DB2 | 0.95951351 | 0.95060976 | 0.95141621 | 0.95060976 | 0.95142276 | 0.98137949 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Average Rank* | 2.6 | 5 | 4.8 | 4.2 | 2.2 | 1.4 |

To understand where the difference lie, we used Friedman’s test. With a value a=0.05 the Friedman statistic gives 350.7, which is bigger than the value in the chi-square table, meaning there is a significant difference between those algorithms.

The critical difference has a value of 3.38. Using Nemenyi test, we can visualize where the difference lies.

Diagram, schematic

Description automatically generated

Figure : Critical Difference Diagram

From the critical difference diagram, we can see that MLP outperforms SVM and KNN but is similar to RF, Tree and GB.