Assignment 1

DV2599: Machine Learning

Anna Bergknut
BTH, DVAMI21
anbe21@student.bth.se

I. Introduction

This project focuses on wine quality classification using machine learning. [2] We explore data preprocessing, classifier evaluation, and the impact of class balancing on model performance.

II. Data Preprocessing

Data preprocessing is a vital phase in machine learning, refining raw data for effective model training. It involves cleaning, transforming, and organizing data to enhance algorithm performance and reliability. In this project, I began by checking for missing values using the isnull() method to ensure data completeness. The dataset was then manually split into 80% for training and 20% for testing. To standardize feature scales, the StandardScaler from scikit-learn was applied. To address class imbalance, the imbalanced-learn library was utilized to balance uneven class distributions within the scaled training set. [4]

III. Algorithms

A. Random Forest

I chose Random Forest for its versatility and robustness in handling both classification and regression tasks. By constructing multiple decision trees during training, each on a random subset of data and features, the algorithm navigates complex data relationships, ensuring high accuracy and offering insights into crucial features. Its inherent diversity and resilience make it effective, even in the presence of outliers. [3]

B. KNeighbors

For the second algorithm, I opted for K-Nearest Neighbors. It operates by assigning a class to a new point based on the classes of its nearest neighbors in a plotted space. The 'K' denotes the number of neighboring points taken into account. If the majority of nearby points belong to a particular class, the new point is predicted to belong to that class. [1]

IV. training and testing process

The training and testing process involves preparing and evaluating a machine learning model. During training, the model learned patterns from the training data, aided by standardized feature scales and addressed class imbalance. Four classifiers underwent evaluation through repeated k-fold cross-validation, with Random Forest emerging as the best performer. The final model tested accuracy and performance on the reserved test set, providing insights into its predictive capabilities on new, unseen data. [4]

V. model evaluation/results

Prior to addressing class imbalance, Random Forest exhibited commendable performance with a mean accuracy of 65.23% and a minimal standard deviation of 1.26%, surpassing the K-Neighbors Classifier, which achieved a mean accuracy of 52.31% with a standard deviation of 1.08%.

Upon balancing the training data using RandomUnderSampler, the Random Forest Classifier's mean accuracy experienced a decrease to 18.59%, accompanied by a standard deviation of 9.84%. The K-Neighbors Classifier, though demonstrating an increase to a mean accuracy of 27.78%, maintained a standard deviation of 10.03%.

For the final model both before and after balancing i used Random Forest was chosen due to its superior performance on the balanced training set. The test set accuracy before balancing was 53,98% and after it was 12,55%.

VI. My thoughts

The unexpected decrease in performance after balancing the training data can be attributed to the significant reduction in data points through undersampling. This reduction may have led to a loss of valuable information for the Random Forest model, impacting its ability to generalize well to the test set. Despite the improvement in class distribution, the resulting model struggled to maintain its previous accuracy levels.

The K-Neighbors Classifier, on the other hand, displayed a moderate increase in mean accuracy after balancing. This improvement suggests that the balancing technique had a positive effect on the model's ability to discern patterns within the data. However, the high standard deviation indicates variability in performance, highlighting potential sensitivity to different folds in the cross-validation process.

Ultimately, the choice of Random Forest for the final model was based on its superior performance. The substantial drop in test set accuracy after balancing may raise questions about the trade-off between addressing class imbalance and preserving the model's predictive capabilities. This underscores the nuanced nature of handling imbalanced datasets and emphasizes the need for a careful consideration of the impact on both training and test performance.

References

[1] scikit-learn, Nearest Neighbors, webpage, 2023, https://scikit-learn.org/stable/modules/neighbors.htmlnearest-neighbors-classification

[2] scikit-learn, Choosing the right estimator, webpage, 2023, https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

[3] Geeks for Geeks, amandp13, Random Forest Classifier using Scikit-learn, webpage, 2022, https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/

[4] Simplilearn, Mayank Banoula, Machine Learning Steps: A Complete Guide, webpage, 2023, https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps