ID 2214
Lab 4, Group 4

Lab report

Lab assignment 4

Lucas Larsson *lulars@kth.se*

Mihaela Bakšić *baksic@kth.se*

Contents

1 Introduction 3

2 Feature extraction 3

3 Models and Oversampling 5

4 Results 6

5 Discussion 6

1 Introduction

In this lab, we will be developing a predictive model that will be used to determine if a certain chemical compounds is active or not.

Relatively simple task, the tricky part is determine which model to use out of the many available ones. And understaning what makes a chemical compounds active. For this task we used models imported from the scikit-learn library.

Our goal is to maximise the area under the ROC curve also known as AUC score. We achieved this through applying the scientific method, we formed a hypothesis on which features and models to use and tested them and then optimized for the maximum AUC score.

2 Feature extraction

In simple terms this section aims to answer the question, ***what features determine if a chemical compounds is active or not?***

Our approche for this was to start researching what makes a molecule active, and if it is possible to account for it somehow and assign it a metric.

The preliminary list contained the following attributes:

1. Charge (CHARGE)
2. Number of Saturated Carbocycles (NUM SAT CARB)
3. Mass (MASS)
4. Number of Heavy Atoms (NUM HEAVY)
5. Number of NH or OH (NUM NHOH)
6. Number of Rotatable Bonds (ROTATABLE BONDS)
7. Number of Rings (NUM RINGS)
8. Number of Aromatic Rings (NUM AROM RINGS)
9. Number of Saturated Rings (NUM SAT RINGS)
10. Number Hydrogen Acceptors/Donors (ACCEPTORS, DONORS)
11. Morgan Fingerprints (FP_i for each bit in the vector)
12. Number of Aliphatic Hydroxyl Groups

13. Number of Atoms (NUM C, NUM N, NUM O)
14. Number of Saturated Heterocycles
15. Number of Functional Groups
16. Wildman-Crippen LogP Value (LOG P)
17. Wildman-Crippen MR Value (MR)

Next step as mentioned above is determine which features are helpful and remove the rest, even Possibly discuss the need for more features if needed.

The main distinction was made between the hand crafted and extracted features and Morgan Fingerprints. These two sets have been used separately and together to train models.

Considering that the extracted features were based on our limited knowledge of factors contributing to the activity of a molecule, we conducted a feature importance study. The study was conducted using a random forest classifier. We analyzed the mean decrease in Gini impurity when splitting a tree using our extracted features. In Figure 1 we can observe substantial difference in feature importances.

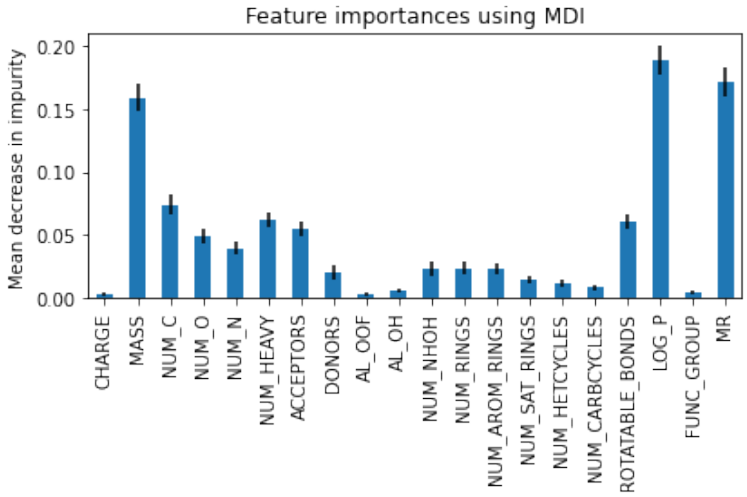

Figure 1: Feature importance using MDI

Further optimization was carried out by removing features with MDI below a certain threshold, training random forest classifier on said feature set and observing change in AUC.

In Table 1 we can observe noticeable increase in AUC while adding features until the first row, meaning that adding features NUM SAT RINGS, NUM HETCYCLES was not informative for the classifier and we deem those unneeded.

This resulted in the final custom feature set consisting of MASS, NUM C, NUM O, NUM N,

FEATURES	AUC
MASS, NUM C, NUM O, NUM N, NUM HEAVY, ACCEPTORS, DONORS, NUM NHOH, NUM RINGS, NUM AROM RINGS, NUM SAT RINGS, NUM HETCYCLES, ROTATABLE BONDS, LOG P, MR	0.9492
MASS, NUM C, NUM O, NUM N, NUM HEAVY, ACCEPTORS, DONORS, NUM NHOH, NUM RINGS, NUM AROM RINGS, ROTATABLE BONDS, LOG P, MR	0.9464
MASS, NUM C, NUM HEAVY, ACCEPTORS, ROTATABLE BONDS, LOG P, MR	0.9122
MASS, LOG P, MR	0.8107

Table 1: Table to test captions and labels.

NUM HEAVY, ACCEPTORS,DONORS, NUM NHOH, NUM RINGS, NUM AROM RINGS, ROTATABLE BONDS, LOG P and MR features.

Regarding Morgan fingerprints features, a grid search over multiple values and models gave us an optimal fingerprint size of 512 bits. This length is used in further experiments.

All features were computed using the open source toolkits RDKit and molmass.

3 Models and Oversampling

As shown in Table 2, the training data is extremely imbalanced, which can result in a biased model to the majority class. To counteract it we used **oversampling**.

0	154528
1	1730

Table 2: Count of chemical compounds and there Activity class.

The two main ways we used oversampling are with **ROS** and **SMOTE**. Both are great techniques for addressing the issue of imbalanced data sets.

ROS works by randomly duplicating observations from the minority class in order to increase its representation in the dataset. This can improve the performance of the model by providing a more balanced dataset for training. However, it does have the potential to introduce noise and overfitting.

SMOTE, on the other hand, creates synthetic observations of the minority class by interpolating between existing observations. This can be a more effective method of balancing the dataset, as it can create a more diverse range of observations for the minority class. However, it can also potentially reduce the model's performance on previously unseen data.

Both methods were used to increase the size of the underrepresented class (ACTIVE=1) to 10% of the total dataset size.

As mentioned above both techniques have their own drawbacks. We use both SMOTE, ROS and the original dataset to train the models to achieve the best results.

Three datasets were created as a basis for our feature combinations, one with fingerprint features, one with custom extracted features and one with both. Afterwards, 6 more were created using SMOTE and ROS.

All features have been scaled using the standard scaler.

Models used for classification are logistic regression, Naive Bayes classifier, random forest classifier and XGBoost.

For determining the best model, we focused on AUC score calculated in a six-fold cross validation setup.

For tree-based models, grid search over number of learners has been performed.

4 Results

Most models displayed improved accuracy, F1 score and AUC score for SMOTE and ROC oversampled datasets in comparison to the original dataset.

Furthermore, both fingerprint and custom feature datasets have achieved satisfying results for at least some models. However, the combined feature set has not reached satisfying AUC score (less than 0.75) with any of the considered models. Despite the fingerprints often being augmented with additional custom features to improve classifier performance, this did not show good results in case of this dataset, and combined feature set displayed poorest AUC in comparison with two other feature sets.

Generally, using SMOTE and ROS improved performance of classifiers.

The best performing model is random forest classifier (50 learners) on the Morgan fingerprints dataset oversampled using SMOTE, with accuracy of 0.9874, F1 score of 0.9268 and AUC of 0.9374.

5 Discussion

We expected more complex models, such as random forest classifier and XGBoost, to be more suitable for the fingerprint feature set and the simpler ones, such as logistic regression and Naive

Bayes to be more suitable for the custom feature set.

That was overall true, except for the case of XGBoost that performed worse in comparison with the other tree. This could be either due to lack of data or poor reaction of the model to the disbalance between the classes.

Furthermore, the custom features set could be further improved with a more extensive analysis of all features that can be extracted from the rdKit library. However, we believe this was out of scope for this assignment.