Drug Activity Prediction

Since "Accurate predictions often only depend on a few molecular features that might be detected even in ultra-small training sets"[1], during this project more focus has been posed on the data pre-processing rather than a sophisticated model design.

1. DATA PRE-PROCESSING

The given data-set consist of 12000 samples, clearly unbalanced, both with regard the label (64%-94% of unknown values) and the task sample distribution. The first step to mitigate such situation was to **get rid of all 0 values** (unknown), reducing, in some instances, the number of available data from 12000 to $\sim$ 600 (e.g. task 4 and 5). After that, the smiles representation have been converted into their **unique canonical form** (we recall that to different representation may correspond the same molecule), and finally translated into numerical form, i.e. **Morgan fingerprints, with radius 4 and 1024 bits**. This was performed on both training and test set. A normalization (with StandardScaler) on the Morgan fingerprints has held no better results, therefore has been disregarded. At this point various algorithm of the SMOTE family have been tested for **resampling**: SMOTE, SVMSMOTE, BorderlineSMOTE and ADASYN. To find the more suitable, the data-set has been divided into training (80%) and validation (20%) sets.

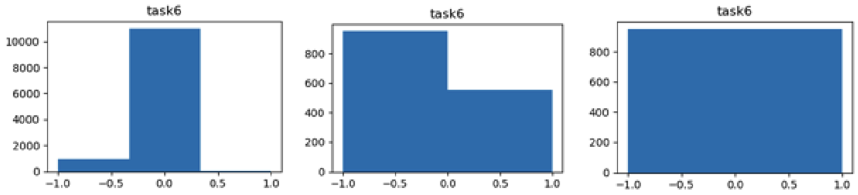

Fig. S1. task 6 data: original (left), SVMSMOTE (center), BorderlineSMOTE (left)

2. TRAINING

A **seed of 42** has been set for reproducibility. A vanilla **Random forest (RF)** with no tuning and no data augmentation has been **trained on each of the 11 tasks** as benchmark, yielding an AUC value of 0.54 on the provided test data. A **grid search** on the RF found the following **optimal parameters[2]: n_estimators=100, criterion='entropy', max_depth= 12, min_samples_leaf= 2, min_samples_split= 5**. This model was used to compare resampling techniques. Also A **Gaussian Naive Bayes (GNB)** was trained with var_smoothing=0.1, but since the result were worse, this has not been further inquired. All the models predicted both a class and a probability distribution for labels -1 and 1.

3. EVALUATION

The RF architecture has been used yielding similar results with huge improvement from the vanilla version, but no decisive difference among augmentation techniques, in fact, the **best performing model on the test set (RF+SVMSMOTE, AUC: 0.704)** is not the best on the validation set (RF+BorderlineSMOTE, AUC: 0.855). The table below summarize the results.

	RF				**GNB**
	SMOTE	**SVMSMOTE**	**BorderlineSMOTE**	**ADASYN**	**SVMSMOTE**
validation set	0.822	0.845	0.855	0.838	0.705
test set	0.667	0.704	0.672	0.653	0.566

Table S1. AUC on validation and test set with different SMOTE resampling algorithms

REFERENCES

[1] Friederike Maite Siemers, Christian Feldmann, Jürgen Bajorath, Minimal data requirements for accurate compound activity prediction using machine learning methods of different complexity, Cell Reports Physical Science, Volume 3, Issue 11, 2022,

[2] Azlim Khan, A.K.; Ahamed Hassain Malim, N.H. Comparative Studies on Resampling Techniques in Machine Learning and Deep Learning Models for Drug-Target Interaction Prediction. Molecules, 2023