PR2 – Drug Activity Prediction

Accuracy: 82.35%

Problem Statement

Determine whether a compound is active (1) or not (0).

Solution

Feature selection method is implemented to reduce dimension and consider only important feature for prediction. Using selected features different classifiers are validated for maximum F1 score as the data is highly imbalanced.

Methodology

1. Splitting train data into train and validation set. It is split in the same ratio as training data
2. Perform dimensionality reduction
3. Perform cross validation and check F1 score
4. Validate F1 score of set of classifiers

Split data

Since, the data is highly imbalanced it is not advisable to train the system as it is, as it may lead to overfitting of data. Below are the methods tried.

1. **SMOTE** over sampling – SMOTE stands for Synthetic Minority Over-sampling Technique. It is the process of creating new minority class from the dataset. First step, it ignores majority class examples. Then for every minority instance, choose 'K' nearest neighbors. Create new instances halfway between the first instance and its neighbors. Though, this process increases minority samples, it did not give better results due to nature of the data.
2. Second method tried was to split training and validation from the original train set with **same distribution as train set**. Minority and majority class is separated say as 722 and 78 respectively. Out of 560 data points in train set (70% of train) – 504 data points are from majority and 56 are taken from minority (70% each). Now out of 240 validation points – 216 are from majority and 24 are from minority set (30% each)

Dimensionality reduction

Dimensionality reduction is done over training and validation data. Various dimensionality method is tried and evaluated. **PCA, SVD, LDA, Random Projections along with combinations of dimensionality reduction methods** were tried. LDA was not giving proper reduction as the data set was close and not exactly linearly separable. F1 score and accuracy was poor for random projection and combination of dimensionality methods (SVD + PCA / PCA+ LDA). PCA was better than the above two methods but it was not able to give proper reduction as they don't maintain object distance. SVD was the best out of all. It was giving feature reduction from 100001 to 800

and it is possible as they preserve distance between objects giving proper components covering maximum variance.

Cross-validation

Before using classifier to predict status of molecule, it is important cross-validate the model and adjust necessary parameters. Two methods of cross validations have been attempted here.

1. **K-Fold cross validation**: Training set is split into 'k' smaller set. Model is trained using 'K-1' of the folds of training data. The resulting model is validated on the remaining part of the data. There are ways to shuffle training the set so that performance can be increased.
2. Training using split train set and validate using created validation set. From this **F1, accuracy, precision and recall values of every classifiers** is obtained.

Accuracy and F1 value obtained from this step is the key for classifier selection.

Classification step

Different classifiers such as Linear SVM, Logistic Regression, KNN, Gradient Boosting, Decision tree, random forest, neural network, naïve bayes, adaboost and extra tree are tried. Out of all these Adaboost, Neural network and Naïve bayes gave significant output F1 score. Adaboost was the best out of all as they fit classifier on the original dataset and then fit additional copies of the classifier on the same dataset. Doing so, weights will be adjusted for incorrectly classified samples as they focus more on incorrectly classified on in subsequent steps. Below is the accuracy table of the obtained in different trials.

S.NO	Dimensionality Reduction	Algorithm	Validation	Validation				Test F1 Score
			Accuracy	F1	Precision	Recall	Accuracy	
1	PCA	Gradient Boosting Classifier	0.91	0.9377	0.9442	0.9456	0.9456	
		Adaboost	0.91	0.9449	0.9468	0.9497	0.9497	66.67
		Neural Network	0.9	0.907	0.919	0.899	0.899	72.22
		Logistic Regression with SMOTE	0.9	0.88	0.89	0.85	0.86	71
		Logistic Regression	0.9	0.8711	0.9299	0.8451	0.8451	60
		Decision Tree	0.9	0.8798	0.881	0.8786	0.8786	52.63
		Extra tree	0.91	0.858	0.8167	0.903	0.9037	NC
		Naïve Bayes	0.9	0.8356	0.8485	0.8242	0.9037	NC
		Linear SVM	0.9	0.858	0.8167	0.903	0.9037	NC
		Nearest Neighbors	0.9	0.8582	0.8167	0.9307	0.9037	72
		Random Forest	0.9	0.8859	0.9234	0.9163	0.9163	NC
2	SVD	Gradient Boosting Classifier	0.91	0.9514	0.9512	0.9539	0.9539	NC
		Adaboost	0.91	0.9318	0.9397	0.9414	0.9414	82.35
		Neural Network	0.9	0.9352	0.9339	0.9372	0.9372	72.22
		Logistic Regression	0.9	0.8859	0.9234	0.9163	0.9163	54.17
		Decision Tree	0.9	0.895	0.9065	0.887	0.887	NC
		Extra tree	0.91	0.858	0.8167	0.9037	0.9037	NC
		Naïve Bayes	0.9	0.8328	0.8475	0.82	0.82	78
		Linear SVM	0.9	0.858	0.8167	0.9037	0.9037	78.79
		Nearest Neighbors	0.9	0.858	0.8167	0.9037	0.9037	NC
		Random Forest	0.9	0.8942	0.9269	0.9205	0.9205	NC
3	LDA	Does not work as consider all classification as one						
		Logistic Regression	0.358	0.342	0.258	0.508	0.508	54
4	Sparse PCA	Decision Tree	0.89	0.98	0.98	0.98	0.98	23

Best Accuracy: 82.35 % : SVD + Adaboost with estimator as 100