Drug Activity Prediction

Name: Sahil Shrikrishna zele

Net id: szele

Miner username: Sahilzele

G no: G01465963

Approaches. And methodologies:

1.Data processing: loaded the training dataset into a sparse matrix, placing 1 where the features are present and 0 otherwise. Separated the labels from the features.

2. Resampling: tried SMOTE for oversampling and Random under sampler for under sampling, tried doing both as the dataset is imbalanced. Using the resampling did not help a lot to increase the f1\_score.

3. Feature selection: since there are alot of features, tried using PCA, TruncatedSVD and SelectKBest, out of these three, SelectKBest gave the best result. Used k = 150 for the same.

4. Classification and F1\_Score: tried using three different models namely DecisionTreeClassifier, GaussianNB and BernoulliNB. Out of these three, bernoulliNB had the best F1\_score and was the final model selected for this problem. Below are the F1\_scores calculated from the final result of 73% on miner.

|  |  |
| --- | --- |
| **Model names** | **F1\_score** |
| DecisionTreeClassifier | 0.926151824 |
| GaussianNB | 0.943264864 |
| BernoulliNB | 0.937320003 |

5. F1-Score: Used cross validation and F1\_sore to measure the accuracy since f1\_Score works best in terms of imbalanced dataset.

6. Predicting the output using BernoulliNB and loading the file.

Different Approaches:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classification models** | **Feature selection** | **Resampling method** | **Score** |
| decision tree | PCA | SMOTE | 0.21 |
| decision tree | TruncatedSVD | SMOTE & Random under sampler | 0.23 |
| GaussianNB | PCA | SMOTE & Random under sampler | 0.21 |
| GaussianNB | TruncatedSVD | SMOTE & Random under sampler | 0.17 |
| BernoulliNB | PCA | SMOTE | 0.39 |
| BernoulliNB | TruncatedSVD | SMOTE | 0.53 |
| BernoulliNB | SelectkBest | SMOTE | 0.67 |
| BernoulliNB | SelectkBest(k = 100) | None | 0.72 |
| BernoulliNB | SelectkBest(k = 150) | None | 0.73 |

References:

1. <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>
2. <https://github.com/scikit-learn/scikit-learn/pull/21109>
3. <https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html>
4. <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html>
5. <https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/>