HW2 - Drug Activity Prediction

**Name:** Aditya Varma Vetukuri

**Miner name:** AdityaVetukuri

**Accuracy:** 0.80

**Rank:** 3

**The objectives of this assignment are the following**:

* Use/implement a feature selection/reduction technique.
* Implement Naive Bayes classifier.
* Experiment with various classification models.
* Think about dealing with imbalanced data.
* F1 Scoring Metric

**Packages** used for this assignment:

* **Numpy**
* **Sklearn**
* **Scipy**
* **Tqdm**
* **Imblearn**

**Feature** **Selection** techniques that I have used for my Naïve Bayes Classifier:

* **Select K Best Feature Reduction Technique :** This technique takes all the features and scores each feature with a scoring function. I used chi2 as my scoring function. It returns the top K-highest scoring features.
* **Principal Component Analysis:** It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Because, the data set is binary matrix PCA was converting it to different co-ordinate system which will make it non binary data. So it was not giving me promising outputs.
* **Truncated SVD:** As I was working with a csr\_matrix, I’ve read in Internet that for matrix feauture reduction Truncated SVD is the best matrix reduction technique to use. It has a parameter called n\_components. I’ve reduced the components to 3000 from 10000 features.

**Implementation of NaiveBayes Classifier:**

* Firstly, I’ve calculated the frequencies of each feature occurring in both the classes and overall frequency.
* Then I have calculated the prior probability of each feature belonging to a particular class and summed up all priors which will give a posterior probability belonging to a class.
* Later I’ve chosen a value which maximizes the value of our Posterior probability
* The prediction is then appended to my resultant list

**Other Classification models that I have tried for this assignment**

* **Guassian Naïve Bayes** (I’ve also implemented this classifier)
* **RandomForest Classifier**
* **Artificial Neural Networks Classifier**
* **Decision Tree Classifier**

Dealing with **Imbalanced** data:

* The dataset is very imbalanced because the number of active classes are very low with (78) actives and (722) in actives.
* We can balance the dataset by Resampling the data.

We can Resample the data by two ways

* **Oversampling** the minority Class
* **Undersampling** the majority Class
* I’ve used inbuilt **RandomUnderSamping** and **RandomOverSampling** from sklearn package.
* Implemented **SMOTE** Over sampling followed by **Tomek** **links**.
* The “**balanced**” parameter in **Decision** **Tree** **Classifier** and **Random** **Forest** **Classifier** worked the best for F1-Score
* The balanced mode automatically adjusts the class weights using the class labels this method worked the best for balancing data than any Sampling technique.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Feature Selection | Balancing method | F-1 Score | RunTime |
| Naïve Bayes | **None** | **None** | **0** | **10seconds** |
| Naïve Bayes | SelectKbest (k = 10) | **None** | 0.7179487 | **10seconds** |
| Naïve Bayes | SelectKbest (k = 5) | **None** | 0.7368421 | **9seconds** |
| Naïve Bayes | SelectKbest (k = 10) | **SMOTE** | 0.8062015 | **9seconds** |
| Naïve Bayes | SelectKbest (k = 5) | **SMOTE** | 0.8538461 | **9seconds** |
| Decision Tree | **None** | **Class\_weight =**  **“balanced”** | 0.8138621 | **60seconds** |
| DecisionTree | Truncated SVD (n\_components=4000) | **Class\_weight = “balanced”** | 0.9362351 | **2mins** |
| DecisionTree | Truncated SVD (n\_components=4000) | **None** | 0.5641022 | **2mins** |
| RandomForest | None | **Class\_weight = “balanced”** | 0.6573428 | **1min** |
| RandomForest | Truncated SVD (n\_components=4000 | **Class\_weight = “balanced”** | 0.7423123 | **2min** |
| RandomForest | Truncated SVD (n\_components=4000 | **None** | 0.5714285 | **2min** |
| Artificial Neural Networks | **None** | **SMOTE** | 0.8612312 | **30seconds** |
| Artificial Neural Networks | **None** | **None** | 0.5623123 | **30seconds** |
| Artificial Neural Networks | Truncated SVD (n\_components=4000 | **SMOTE** | 0.9123242 | **30seconds** |

* The above accuracies are for the training data by using test-train cross validation the accuracy for this and on Miner are not relative because of the imbalanced data that we have in our dataset
* Smote is increasing the f1\_score drastically by re-sampling the minority class.

Below are my Top accuracies on Miner.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Feature Selection | Balancing Method | F-1 Score |
| Decision Tree Classifier | Truncated SVD | Class\_weight = “balanced” | 0.80 |
| Naïve Bayes | SelectKbest (k =10) | SMOTE | 0.78 |
| Naïve Bayes | SelectKbest (k = 20) | NONE | 0.77 |
| Decision Tree | NONE | Class\_weight = “balanced” | 0.76 |
| Artificial Neural Networks | NONE | NONE | 0.71 |

**Things that I have learned with this Assignment** :

* Dealing with Binary Data, I converted all the features to a **CSR** **matrix** with rows as drug and columns as their features and the row-column value is a floating value 1 that it exists.
* Detail implementation of Naïve Bayes Classifier and how a feature which does not exist may effect the entire posterior probability. I used the **Laplace** **Estimate** to fix this.
* How **SelectKbest** works better than PCA on binary data as I said earlier PCA was converting the binary data to another dimension for visualization and thus effecting the f1\_score.
* Reduction of my features in **csr\_matrix** to less features from 10,0000 to 4,000.
* Some Resampling techniques like Random Oversampling, Random **UnderSampling**, **SMOTE** and **SMOTE**.
* Working with Tree Classifiers, How good Decision Tree and RandomForest Trees were working on imbalanced datasets with their inbuilt parameter (class\_weight = “Balanced”) which automatically balances the class labels with their weights.