Computer Engineering Department

Course Name: CMPE255 – Data Mining

Student Name: Sahitya Mullapudi

Sjsu Id: 011545404

Program 2: Drug Activity Prediction

Semester: Fall,2017

Rank: 29

Accuracy: 0.7742

Classification Algorithms: Random Forest, SVC, Naïve Bayes, Neural networks

Dimensionality Reduction: PCA, SVD

Methodology:

1. The train set is divided into classes and data.

2. The test set and train set data are then combined.

3. The documents are split into documents of words.

4. csr matrix is built for the processed data.

5. The test and train set are divided.

6. A dimensionality reduction method is selected and test and train data are transformed to fit.

7. A classification algorithm is selected and applied on train set data and classes of the train set.

8. The same algorithm is used to predict the classes of test set.

Approach:

I have followed two approaches:

1.

The train set is divided into classes and data. Test set data and trainset data are combined. A dense matrix is generated with 100001 columns and 1150 rows. This is then passed to PCA , using sklearn libraries. This is tried with different number of components. The train set and test set are divided.

The classification algorithms, Neural networks and Naïve Bayes, Random Forest are used. This is tried with different parameters in random forest and neural networks.

data	Dimensionality reduction	classification	F1 score
dense	PCA	Random forest	0.64
dense	PCA	Neural networks	0.6842
dense	PCA	Naïve Bayes	0.20

2.

The train set is divided into classes and data. Test set data and trainset data are combined. A CSR matrix is created. For dimensionality reduction, Truncated SVD is used. And for classification- SVM , neural networks, random forest are tried. CSR_idf is also tested. But the results did not vary much.

data	Dimensionality reduction	classification	F1 score
Csr matrix	Truncated SVD	Random forest	**0.7742**
Csr matrix	Truncated SVD	Neural networks	0.7097
Csr matrix	Truncated SVD	SVC	0.20

Conclusion:

The classification algorithms, SVC and naïve bayes ‘s performance is not even considerable. The random forest and neural networks produced results in the same range. Changing of the parameters in algorithms gave lesser F1 score always. With the above score, a sparse matrix with dimensionality reduction done with truncated SVD and classification with random forest is higher.