Shahjalal University of Science and Technology
Department of Computer Science and Engineering

CSE 476

Machine Learning Classifiers applied on Document Classification and Stylogenetics

Umme Sumaya Jannat
Reg. No.: 2014331022
4^{th} year, 1^{st} Semester

Syed Md. Hasnayeen
Reg. No.: 2014331025
4^{th} year, 1^{st} Semester

Department of Computer Science and Engineering

Course Teacher
Ayesha Tasnim
Assistant Professor
Department of Computer Science and Engineering

July 15, 2018

1. Document Categorization

Our Objective was to categorize an article based on the topic.

1.1. Dataset

The dataset was collected from https://scdnlab.com/corpus/ . The dataset contains articles of 2 categories in 2 different folders. The number of articles each folder contains is given below.

1. Accident : 6350 articles
2. Crime : 8840 articles

1.2. Feature Extraction

For the project,we TF-IDF values of each unique words in a document as a feature vector for a single document.

1.3. Experiment Procedure

The task is mainly classification. For each categories we took 80% articles in our train dataset and 20% articles in our test dataset. We fed the TF-IDF vectors of train dataset in a classifier and checked how the classifer worked on test dataset. We used Different Classifiers for this task. The results of our experiments is given below.

1.4. Experiment Results

1.4.1. Naïve Bayes:

We achieved accuracy rate of 88.8 using Naïve Bayes Classifier. We used MultinomialNB as classifier of sklearn. Multinomial Naive Bayes has a parameter alpha, which acts as a additive smoothing parameter. We tuned this value and got different accuracy rates.

1.4.2. Decision Tree:

We achieved accuracy rate of 87.0 using Decision Tree Classifier.

1.4.3. Support Vector Machine:

We achieved accuracy rate of 57.8 using Support Vector Machine with linear kernel.

1.4.4. Neural Network:

We achieved accuracy rate of 92.03 using Neural Network.

1.4.5. K-nearest Neighbor:

We achieved accuracy rate of 66.8 using K-nearest neighbor(KNN).

2. Stylogenetics

Our Objective was to categorize writings based on authorship attribute.

2.1. Dataset

The dataset was collected from CSE batch 2013 https://drive.google.com/open?id=11bCA8dzCOz2T
. The dataset contains articles of 6 authors in 6 different folders. The number of articles each folder contains is given below.
1. Emon Jubayer : 378 articles
2. Hasan Mahbub : 174 articles
3. MZI : 121 articles
4. Nir Shondhani : 225 articles
5. Rono Dipon Bashu : 202
6. Tareq Anu : 343 articles

2.2. Feature Extraction

For the project,we TF-IDF values of each unique words in a document as a feature vector for a single document.

2.3. Experiment Procedure

The task is mainly classification. For each categories we took 50% articles in our train dataset and 50% articles in our test dataset. We fed the TF-IDF vectors of train dataset in a classifier and checked how the classifer worked on test dataset. We used Different Classifiers for this task. The results of our experiments is given below.

2.4. Experiment Results

2.4.1. Naïve Bayes:

We achieved accuracy rate of 51.2 using Naïve Bayes Classifier. We used MultinomialNB as classifier of sklearn. Multinomial Naive Bayes has a parameter alpha, which acts as a additive smoothing parameter. We tuned this value and got different accuracy rates.

2.4.2. Decision Tree:

We achieved accuracy rate of 73.4 using Decision Tree Classifier.

2.4.3. Support Vector Machine:

We achieved accuracy rate of 44.4 using Support Vector Machine with linear kernel.

2.4.4. Neural Network:

We achieved accuracy rate of 63.01 using Neural Network.

2.4.5. K-nearest Neighbor:

We achieved accuracy rate of 64.95 using K-nearest neighbor(KNN).