NC STATE UNIVERSITY

EM Algorithm for Text Classification

Yufei Cai[1] Boming Miao[2] [1]SUSTech [2]Northeastern University

Introduction

Based on expectation maximization (EM) classification algorithm, we are to investigate the effectiveness of several modern classifiers under the semi-supervised learning framework.

Methods

Naive Bayes Classifier

y unknown, $x_1, \dots x_n$ mutually independent

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n \mid y)}{P(x_1, \dots, x_n)}$$

$$\forall i \in \{1, 2, \dots, n\}, P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y)$$

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\,\Pi_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)} \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\hat{y} = \operatorname*{argmax}_y \prod_{i=1}^{n} P(x_i \mid y)$$

BERT Classifier

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google.

EM Semi-supervised Learning

Algorithm 1 EM Semi-supervised Learning for Classification

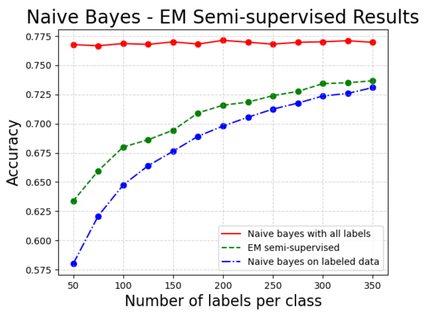
Input: $classifier$, $(\boldsymbol{X_l}, \boldsymbol{y_l})$, $\boldsymbol{X_u}$, $iteration\_times$
Output: $classifier$

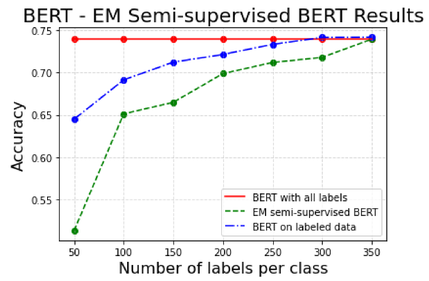
```
function EM_SEMI_SUPERVISED_LEARNING_FOR_CLASSIFICATION(classifier, (X_l, y_l), X_u, iteration_times)
    classifier.train(X_l, y_l)
    X ← X_l ∪ X_u
    for i = 1 to iteration_times do
        y_u ← classifier.predict(X_u)
        y ← y_l ∪ y_u
        classifier.train(X, y)
    end for
end function
```

Dataset

The twenty newsgroups dataset contains 1000 text articles posted to each of 20 online newgroups for a total of 20,000 articles whose labels are hierarchically organized as a tree.

Results





Results

Classifier	Hierarchy	Accuracy
Naive Bayes	Non-hierarchical	0.7719
Naive Bayes	Hierarchical	0.7392
Bert	Non-hierarchical	0.7400

Accuracy of Hierarchical and Non-hierarchical Classification

Conclusion

The semi-supervised EM algorithm can increase the accuracy when combined with naive bayes classifiers but will decrease the accuracy when combined with BERT owing to the fact that BERT requires labels more accurate in order to function well.

The accuracy of the hierarchical naive bayes classifier is lower than that of the non-hierarchical one owing to the fact that each classifier on the tree is train by less data so that it cannot distinguish different classes with less information given.

References

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[2] Kamal Nigam, Andrew McCallum, and Tom M Mitchell. Semi-supervised text classification using em., 2006.

caiyf2018@mail.sustech.edu.cn

bomingmiao@gmail.com