Abstract Classification

[COMP 598 Group Project 2] [*]

Benedicte Leonard-Cannon
McGill University
benedicte.leonard-cannon@mail.mcgill.ca

Faiz Khan
McGill University
faiz.khan@mail.mcgill.ca

Sherry Ruan
McGill University
shanshan.ruan@mail.mcgill.ca

ABSTRACT

TBD

1. INTRODUCTION

What is text classification (or categorization)

Briefly describe our methodology

Talk about potential uses?

2. RELATED WORK

In the following section, we present existing work conducted by other researchers in the task of text categorization.

Genkin and Lewis[X] proposed a Bayesian lasso logistic regression model for binary text categorization that relied on a Laplace prior to reduce the risk of overfitting. Their approach addressed the impracticality of fitting a standard logistic regression model to a dataset containing a large feature space. More precisely, their training algorithm used prior probability distributions of the model parameters to encourage model sparsity. According to them, this approach produced a compact model that is effective and doesnâĂŹt overfit. In practice, their algorithm performed as well as two state-of-the-art categorization models (support vector machines (SVM) and ridge logistic regression) on five standard test sets (ModApte, RCV1-v2, OHSUMED, WebKB and 20 NG). [Regression: http://www.stat.columbia.edu/ madigan/PAPERS/techno.pdf]

Joachims (X) was the first to study the performance of SVMs for text classification in his 1998 paper. Joachims used two SVMs -one based on a polynomial kernel and the other on a radial basis function (RBF) kernel-. He compared both of these models with the following benchmark algorithms: Naive Bayes, Rocchio, k nearest neighbors (k-NN) and C4.5 decision tree. Here again, the performance of these classifiers were assessed through the ModApte and Ohsumed datasets. Prior to fitting, these datasets were reduced to a bag-of-words representation out of which stopwords were discarded. The resulting feature vectors were normalized to unit length and the best features were selected according to their information gain. From the experiments conducted, Joachim concluded that both SVM algorithms outperformed the four benchmark algorithms significantly. http://www.cs.cornell.edu/people/tj/publications/joachims98a.pdf

Some sources I might use:

3. METHODOLOGY

Data preprocessing *Write which libraries were used for each case if applicable* Lower case Remove punctuation Remove stop words Stemming What else? Feature selection Build a dictionary of all words present in abstracts Get rid of words occurring less than X times to reduce dimensionality of dataset Build 2 dataset based on the dictionary. 1 dataset contains word occurrence/absence for each abstract. Other dataset is bag-of-words. Remove features with low variance (scikit learn) Univariate feature selection (scikit learn) ?? if we can

Joachims says âĂIJaggressive feature selection may result in a loss of informationâĂİ even if we only discard the least relevant features. Hence, it is best if we keep as many features as we can handle.

Algorithm selection Multinomial, multivariate Naive Bayes (Bennie write) Nearest Neighbor (Sherry write) Random forest (Bennie write) Try SVMs According to Joachims: âĂIJtheir ability to learn can be independent of the dimensionality of the feature spaceâĂİ Optimization (if required) ? Parameter selection Laplace smoothing alpha for Bayes. k for NN. Other params Random forests have 10 paramsâĂę

4. TESTING AND VALIDATION

detailed analysis of results, NOT Kaggle

5. DISCUSSION

Improvements -Combine bag-of-words with bigrams or trigrams -Normalize the feature vectors by abstract length (here not a big difference since all abstracts are roughly the same length) -Consider formulae (might be easy to map a given formula to a particular field!)

We hereby state that all the work presented in this report is that of the authors.

6. REFERENCES

[*] The complete dataset of this report is available at http://www.acm.org/eaddress.htm

The implementation of the algorithm described in this report is avaiable at http://www.acm.org/eaddress.htm