Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Machine Learning Techniques in Language Classification

Marina Meilă MMP@STAT.WASHINGTON.EDU
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Division of Computer Science and Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Editor: Kevin Murphy and Bernhard Schölkopf

Abstract

This research project report describes a series of text classification experiments using Naive Bayes and Support Vector Machine classifiers to identify the language of an article. The English, German, French and Spanish articles were retrieved from the ¡wikipedia corpus¿ for use in training and testing. Naïve Bayes and SVM classifiers were trained against the corpus using a variety of feature selection techniques, such as Binomial Separation and Term Frequency, to evalute performance tradeoffs and develop more intuition into text classification problems.

Keywords: Bayesian Networks, Support Vector Machines, Feature Selection, Natural Language Processing

1. Introduction

Natural Language Processing (NLP) describes an expansive problem domain in the automated analysis of text. ¡Talk about Spam¿ ¡Talk about languages¿ ¡Talk about entity recognition¿ ¡talk about feature selection¿

2. Statement of Problem

3. Approach

- corpus taken from wikipedia: http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

- English, French, German, Spanish Start with a simple Naive Bayes classifier for each language, trained against the highest frequency words. Evaluate it using F1 measure, Accuracy, and maybe the AUC. Iteratively improve on the classifier by selecting word features using TF-IDF, BNS and KL-Divergence. Evaluate again using the same metrics. Switch from Naive Bayes to SVMs and repeat 1 2. Try different SVM kernels (RBF).

I could pivot my project to the problem of training classifiers for recognizing a series of closely related languages: English, French, German and Latin. I'm focusing on these four because the English vocabulary draws substantially from them, creating an interesting over-

lap. There is a corpus of Wikipedia articles containing translations between language pairs that I could use as the experimental data: http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/. I can parse out the article content into small chunks - perhaps paragraphs of various lengths. It is a large corpus. There were more than 280,000 articles in the English-French article archive I pulled down, so I don't think I'll be at a lack for data. Wikipedia Comparable Corpora Linguatools linguatools.org The Wikipedia Comparable Corpora are bilingual document-aligned text corpora. They have been extracted from the Wikipedia Monolingual Corporas XML files using the crosslanguage links.

The ML work would be as follows: Start with a simple Naive Bayes classifier for each language, trained against the highest frequency words. Evaluate it using F1 measure, Accuracy, and maybe the AUC. Iteratively improve on the classifier by selecting word features using TF-IDF, BNS and KL-Divergence. Evaluate again using the same metrics. Switch from Naive Bayes to SVMs and repeat 1 2. Try different SVM kernels (RBF). I'm hopeful I can show an interesting progression in classifier performance, elaborating on why a specific feature selection technique did/did not help and examine the tradeoffs between Naive Bayes and SVM. SVMs are not something I've had a chance to use yet, so I'd like to explore them a little more. Joachims[1] was cited quite a bit by Chee and did an interesting paper on SVMs in text classification that I'm drawing my understanding from.

4. Background and Related Work in Field

I've been working on a backup idea with these concerns. The dissertation I selected was Dr. Brant Chee's Exploring Machine Learning Techniques Using Patient Interactions in Online Health Forums to Classify Drug Safety. He encountered several problems with forum message quality that introduced me to problems in NLP and text classification. Chee opted for simple, lexicon based techniques to solve some of these problems - such as classifying messages as English or not using counts of English/non-English terms, then linear combination to score the message. However, the citations he included and associated reading I did gave me more insight into how SVMs and Naive Bayes classifiers have been used in text classification problems, like identifying SPAM. It was also my first exposure to some feature identification techniques like KL-Divergence, Binomial Separation (BNS), and using TF-IDF scores as word features.

5. Results

6. Discussion and Analysis of Results

7. Future Work

8. Grading Criteria

- Statement of problem: 15% - Approach : 20% - Background and Related Work: 15% - Results: 15% - Discussion and Analysis of Results: 20% - Relevant Future work: 5% - Construction and readability of the work: 10%

Appendix A.

In this appendix we prove the following theorem from Section 6.2:
Theorem *Let u, v, w be discrete variables such that v, w do not co-occur with u (i.e., $u \neq 0 \;\Rightarrow\; v = w = 0$ in a given dataset $\mathcal{D}$). Let N_{v0}, N_{w0} be the number of data points for which $v = 0, w = 0$ respectively, and let I_{uv}, I_{uw} be the respective empirical mutual information values based on the sample $\mathcal{D}$. Then*

$$N_{v0} \;>\; N_{w0} \;\;\Rightarrow\;\; I_{uv} \;\leq\; I_{uw}$$

with equality only if u is identically 0. ■
Proof. We use the notation:

$$P_v(i) \;=\; \frac{N_v^i}{N}, \quad i \neq 0; \quad P_{v0} \;\equiv\; P_v(0) \;=\; 1 - \sum_{i \neq 0} P_v(i).$$

These values represent the (empirical) probabilities of v taking value $i \neq 0$ and 0 respectively. Entropies will be denoted by H. We aim to show that $\frac{\partial I_{uv}}{\partial P_{v0}} < 0$....

Remainder omitted in this sample. See http://www.jmlr.org/papers/ for full paper.

References