

Using Unlabeled Data For Text Classification

Vom Fachbereich Mathematik
der Technischen Universität Darmstadt
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)
genehmigte

Dissertation

von
Dipl. Inf. Jens Mehnert
aus Saarbrücken

Referent: Prof. Dr. Michael Kohler
Koreferentin: Prof. Dr. Iryna Gurevych
Tag der Einreichung: 7. Juli 2009
Tag der mündlichen Prüfung: 29. Oktober 2009

Darmstadt 2009

D 17

Contents

1. Introduction	1
1.1. The Classes Fact And Non-Fact	3
1.2. Related Work	6
1.2.1. Text Classification	6
1.2.2. Using Unlabeled Data	9
1.3. Prediction Models	9
1.3.1. Linear Model	10
1.3.2. Nonlinear Model	10
1.3.3. Model Free	10
1.4. Plug-In Estimator	11
2. Classification Algorithms	13
2.1. K-Nearest Neighbour	13
2.2. AdaBoost	14
2.3. Neural Network	15
3. A Semi-Supervised Classification Algorithm	19
3.1. Data Preparation And Sentence Statistics	19
3.1.1. Tree Tagger	20
3.1.2. Word Types	20
3.1.3. Sentence Statistics	21
3.2. Short Summary Of The Algorithm	22
3.3. Manual Classification Of Sentences	22
3.3.1. Classifier Learning Data	23
3.4. Manual Classification And Verification Of Words	23
3.4.1. Preselection Of Words	23
3.4.2. Word Class Verification	24
3.5. Automatic Sentences Mining	27
3.6. Automatic Word Mining	30
3.6.1. Word Mining Example	31
Nouns	31
Verbs	33
Adjectives	35
Adverbs	36
3.6.2. Resulting Word List	36
3.7. A Word Type Vector	39
3.8. About The Final Algorithm	40

4. Test Section	43
4.1. Test System	44
4.2. Source Code	44
4.3. Bag Of Words	45
4.4. Our Approach	45
5. Consistency	49
5.1. Function Space	49
5.2. Regression Estimator	49
5.3. Consistency Of The Estimate	50
6. Rate Of Convergence	63
6.1. Function Space	63
6.2. Rate Of Convergence Of The Estimate	63
A. Used Theorems	73
B. Used Lemmata	77
C. Preliminary Mathematics	79
C.1. VC Dimension	79
C.2. Covering Numbers	79
C.3. Packing Numbers	81
C.4. Universal Consistency	82
C.5. O_P Notation	82
C.6. Fourier Transform And Related Definitions	82