IN THE NAME OF GOD

TEXT CLASSIFICATION BASED ON PROBABILISTIC LEARNING MODELS

BY
HEYDAR DAVOODI

THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN
PARTIAL FULFILMENT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (M.Sc)

IN
COMPUTER ENGINEERING-
ARTIFICIAL INTELLIGENCE AND ROBOTICS
SHIRAZ UNIVERSITY
SHIRAZ, IRAN

EVALUATED AND APPROVED BY THE THESIS COMMITTEE AS: **EXCELLENT**

S. Katebi ....**S. D. KATEBI, Ph.D., PROF. OF COMPUTER ENGINEERING (CHAIRMAN)**

M. Zolghadri ....**M. ZOLGHADRI JAHROMI, Ph.D., ASSISTANT PROF. OF COMPUTER ENGINEERING (CHAIRMAN)**

A. Towhidi ....**A. TOWHIDI, Ph.D., ASSISTANT PROF. OF COMPUTER ENGINEERING**

April 2001

۳۸۳۷۷

ACKNOWLEDGEMENT

I would like to take this opportunity to thank Prof. S. D. Katebi for his supports and comments to my research work, his helps were like the signs on the road. Also, I would like to thank Dr. M. Zolghadri for his help and numerous advises. Special thanks go to Dr. Towhidi for his advises. I wish to thank Dr. Haddad for his cooperation. Finally I give my thanks to the people who helped me when I was a student at Shiraz University.

۳۸۳۷۷

Dedicated to

my mother

for the first lullaby

ABSRACT

Text Classification Based on Probabilistic Learning Models

By

Heydar Davoodi

As the amount of information available to human increases, the role of automatic information organization becomes more important. The area of classifying the text documents into several groups has been the subject of intensive researches during the last decade.

Classification is an important method for data analysis and there are several techniques in artificial intelligence and pattern recognition which have been proposed for this task, but these methods can not be applied efficiently in text classification task because in this problem we will encounter a high number of features. This thesis aims to investigate the concepts that must be considered in text classification task, like: feature extraction document representation, feature subset selection and proper machine learning algorithms for this task. In feature subset selection we have tried to show the relation between probabilistic classification and criterion which is used in our work. Also, we completely considered a probabilistic framework for text classification. In this framework the probability of belonging one document to all classes will be estimated. In this method we use Bayesian networks as an effective and efficient way of saving a joint probability distribution of variables. Then we have proposed architecture for building a knowledge model using Bayesian Networks. By measuring the performance and comparing the results with one classic algorithm, which commonly used in Information Retrieval, we have shown that these proposed approaches are efficient. We used Reuters data for learning and test and experiments have been carried out in 5 categories of these data. Finally we have shown that inference is also efficient in the proposed models.

TABLE OF CONTENTS

Content **Page**

LIST OF TABLES..XI

LIST OF FIGURES XIII

CHAPTER 1: INTRODUCTION.................................1

1.1 Motivation 1

1.2 Text-Learning and Intelligent Agents.................................3

1.2.1 The Content-Based Approach 3

1.2.2 The Collaborative Approach 4

1.3 Machine Learning and Text Classification Task 6

1.4 Text Categorization Applications.......................................9

1.5 Reuters-21578 Training Data Sets.................................... 11

1.5.1 History 11

1.5.2 Format of Reuters Data 12

1.5.3 Categories 13

1.5.4 Using Reuters-21578 for Text Categorization ....

Research 15

1.6 Domain Space Specifications.. 16

1.7 Thesis Overview 17

CHAPTER 2: TEXT REPRESENTATION AND FEATURE SUBSET SELECTION..19

Content **Page**

2.1 Introduction 19

2.2 Text Representation 20

2.2.1 Vector Space Model 20

2.2.1.1 Term Frequency Measure........................ 21

2.2.1.2 Document Frequency Measure............... 21

2.2.1.3 *IDFTF* Measure................................ 22

2.2.2 Extension To Vector Space Model 23

2.2.3 Preprocessing 24

2.2.3.1 Stop-Word Removing........................ 24

2.2.3.2 Stemming.. 25

2.2.3.3 Pruning Of Low Frequent Words............ 26

2.3 Feature Subset Selection 28

2.3.1 Filtering Approach 29

2.3.2 Wrapper Approach 30

2.3.3 Approach Used in Text Learning 31

2.4 Feature Scoring in Text Domains 32

2.4.1 Some Information Theory Basis 32

2.4.1.1 Huffman's Algorithm........................ 32

2.4.1.2 Entropy.. 34

2.4.1.3 Cross Entropy................................ 35

2.4.2 Mutual Information for Scoring 36

2.4.2.1 Theoretical Framework........................ 36

CHAPTER 3: MACHINE LEARNING FOR TEXT

Content Page

CLASSIFICATION..42

3.1 Introduction...42

3.2 Knowledge Engineering Approach 42

3.2 Machine Learning Approach 43

3.2.1 Decision Trees 43

3.2.2 Support Vector Machine 47

3.2.3 Inductive Rule Learning..51

3.2.4 Other Learning Methods 53

3.2.5.1 Neural Networks................................53

3.2.5.2 Rocchio Method.................................53

3.2.5.3 Linear Least Squares Fit.......................54

3.2.5.4 Sleeping Experts.................................55

3.2.5.5 K-Nearest Neighbor............................55

3.2.5.6 Word Algorithm.............................56

3.3 Final Note 56

CHAPTER 4: LEARNING USING PROBABILISTIC MODELS..58

4.1 Introduction 58

4.2 Preliminaries...59

4.2.1 Introduction to Probability 59

4.2.1.1 Probability as a Measure......................59

4.2.1.2 Bayes Theorem..................................61

4.2.1.3 Probability Distributions on Sets of

Content **Page**

Variables..62

4.2.2 Graph Theory..65

4.2.3 Forms of Independence..67

4.2.4 Factorized Distributions..70

4.3 Bayesian Networks..72

4.3.1 Introduction to Bayesian Networks..72

4.3.2 Learning Bayesian Networks..75

4.3.3 Bayesian Network as a Classifier..82

4.3.4 Models for Classification Task..84

4.3.4.1 Naive Bayes..84

4.3.4.2 Tree Augmented Naive-Bayes (Tan)..85

4.3.4.3 Extension to TAN Model..85

CHAPTER 5: BUILDING A PROBABILISTIC TEXT CLASSIFIER..87

5.1 Introduction..87

5.2 Architecture of a Text Classifier..88

5.3 Performance Measures..94

5.3.1 Precision and Recall..94

5.3.2 Breakeven Point..96

5.4 Experimental Results for Classifiers..98

5.4.1 Naive Bayes Classifier..98

5.4.1.1 Multi-Variate Bernoulli Model..100

5.4.1.2 Multinomial Model..102

Content **Page**

5.4.1.3 Feature Selection for Models................105

5.4.1.4 Inference for Naive Bayes....................107

5.4.2 Tree Augmented Naive Bayes Classifier............... 110

5.4.3 Bayesian Learning for Filtering............................. 118

5.4.4 Extension to TAN Model 126

5.4.5 Inference in a Bayesian Network........................... 129

5.4.6 Rocchio a Classic Algorithm................................. 138

5.5 Analysis of The Results 140

5.6 Conclusion.. 146

5.7 Future Works 148

REFERENCES...150

ABSRTACT AND TITLE PAGE IN PERSIAN

LIST OF TABLES

Table **Page**

Table 1.1: Examples of music rating given by imaginary web user..5
Table 1.2: Category sets of Reuters training data..................13
Table 1.3: Number of Training and Test item in top ten categories of Reuters data...............................15
Table 1.4: Domain space in our experiments.......................16
Table 1.5: Number of overlaps of each category with others in training data...17
Table 5.1: A contingency table.......................................94
Table 5.2: Prior parameters for the categories in our domain space...102
Table 5.3: Parameters of first 10 features for category "corn" in Bernoulli model.......................................102
Table 5.4: Parameters of first 10 features for category "corn" in Multinomial model (using 100 features).............104
Table 5.5: Parameters of first 10 features for category "corn" in Multinomial model (using 10 features)...............104
Table 5.6: First 10 features of category "corn" (Bernoulli).....108
Table 5.7: First 10 features of category "wheat" (Bernoulli)....108

Table **Page**

Table 5.8: First 10 features of category "ship" (Bernoulli......108

Table 5.9: First 10 features of category "interest" (Bernoulli)..108

Table 5.10: First 10 features of category "trade" (Bernoulli)...108

Table 5.11: First 10 features of category "corn" (Multinomial)...109

Table 5.12: First 10 features of category "wheat" (Multinomial)...109

Table 5.13: First 10 features of category "ship" (Multinomial)...109

Table 5.14: First 10 features of category "interest" (Multinomial)...109

Table 5.15: first 10 features of category "trade" (Multinomial)...109

Table 5.16: Parameters of first 10 features for category "corn" using TAN model.......................................126

Table 5.17: Parameters organization for inference in TAN model..134

Table 5.18: Matrix, which holds network structure for inference...134

LIST OF FIGURES

Figure **Page**

Figure 1.1: Example of Reuters data in SGML format............14

Figure 2.1: Bag-of-Word representation of a document using term Frequency..22

Figure 2.2: A sample of stop words Used in Our Experiment...25

Figure 2.3: Illustration of Zipf law for category "corn" in our experiment..27

Figure 2.4: Illustration of the filtering approach to feature subset selection..29

Figure 2.5: Illustration of the wrapper approach to feature subset selection..30

Figure 2.6: Illustration of the approach to feature subset selection commonly used in text data............................31

Figure 3.1: A decision tree example..............................45

Figure 3.2: A tree structuring of partitioned objects...............46

Figure 3.3: An example of separable data..........................49

Figure 3.4: A learned rule set for category "Painting"............51

Figure 3.5: Illustration of vector space for Rocchio method. Distance of new instance to centriod of each category measure by some similarity measures.....54

Figure **Page**

Figure 4.1: An hypothetical set of variables associated with sample space of humans.................................65

Figure 4.2: A graph associated with the variable set of Figure 4.1, An extra node has been introduced in order to simplify the graph......................................66

Figure 4.3: *X* is conditionally independent of *Y* given *Z*..........69

Figure 4.4: *X* and *Y* conditionally dependent given *Z*.............69

Figure 4.5: X and *Y* are conditionally independent given *Z*......70

Figure 4.6: Example of a simple Bayesian network (conditional probabilities with respect to structure store in tables)..73

Figure 4.7: Illustration of Naive Bayes structure.................76

Figure 4.8: A simple TAN structure................................85

Figure 4.9: A 2-dependent Bayesian network......................85

Figure 5.1: An architecture for designing a text classifier........91

Figure 5.2: Preprocessing results for 4 categories in our domain...92

Figure 5.3: Preprocessing result for category "trade".............93

Figure 5.4: Illustration of interpolation for finding Breakeven point...97

Figure 5.5: Precision and recall for category "corn" using Bernoulli Naive Bayes model..........................111

Figure 5.6: Precision and recall for category "wheat" using

Figure Page

Bernoulli Naive Bayes model........................111

Figure 5.7: Precision and recall for category "ship" using Bernoulli Naive Bayes model........................112

Figure 5.8: Precision and recall for category "interest" using Bernoulli Naive Bayes model........................112

Figure 5.9: Precision and recall for category "trade" using Bernoulli Naive Bayes model........................112

Figure 5.10: Precision and recall for category "corn" using Multinomial Naive Bayes model....................113

Figure 5.11: Precision and recall for category "wheat" using Multinomial Naive Bayes model....................114

Figure 5.12: Precision and recall for category "ship" using Multinomial Naive Bayes model....................114

Figure 5.13: Precision and recall for category "interest" using Multinomial Naive Bayes model....................115

Figure 5.14: Precision and recall for category "trade" using Multinomial Naive Bayes model....................115

Figure 5.15: Illustration of TAN network with 10 features for category "corn"......................................119

Figure 5.16: Illustration of TAN network with 10 features for category "interest"..................................120

Figure 5.17: Illustration of several beta distributions...........123

Figure 5.18 : Illustration of 2-dependence Bayesian network with