# Table of Contents

Acknowledgement

Abstract

Table of Contents

List of Figures

List of Tables

Abbreviation

Notations

1. Introduction

1.1. Introduction 1
1.2. Motivation 2
1.3. Need for the Study 4
1.4. Scope and Objectives 6
1.5. Methodology 6
1.6. Organization of the Thesis 7

2. Review of Related Work 9

2.1. Introduction 9
2.2. Definition of Text Categorization 9
2.3. Cross Road of Information Retrieval and Machine Learning 10
2.4. Challenges in Text Categorization 12
2.5. Types of Text Categorization Task 13
   2.5.1. Single Label Vs Multilabel 13
   2.5.2. Hierarchical Vs Non-Hierarchical 14
2.6. Applications of Text Categorization 15
   2.6.1. Automatic Document Indexing 16
   2.6.2. Document Organization 16
   2.6.3. Text Filtering 17
   2.6.4. Word Sense Disambiguation 17
   2.6.5. Hierarchical Categorization of Web Pages 18
2.7. Machine Learning Techniques 18
   2.7.1. Rocchio Method 19
2.7.2. k–Nearest Neighbours
2.7.3. Naive Bayes Classifier
2.7.4. Decision Tree
2.7.5. Support Vector Machine

2.8. Summary

3. Theory of Evidence
3.1. Introduction
3.2. Representation of Uncertainty
3.3. Dempster Shafer Theory
   3.3.1. Frame of Discernment
   3.3.2. Basic Probability Assignment
   3.3.3. Belief and Plausibility Function
   3.3.4. Combination Rule
3.4. Transferable Belief Model
   3.4.1. Credel Level: Entertaining Knowledge
   3.4.2. Pignistic Level: Decision Making
3.5. Applications of Dempster-Shafer Theory
   3.5.1. Data Fusion
   3.5.2. Image Classification
   3.5.3. Color Image Segmentation
   3.5.4. Belief Decision Tree
3.6. Summary

4. Curse of Dimensionality
4.1. Introduction
4.2. Text Representation
4.3. Pre-Processing of Textual Documents
   4.3.1. Word Distribution
   4.3.2. Removal of Stop Words
   4.3.3. Stemming
4.4. Feature Selection
   4.4.1. Filter Methods
   4.4.2. Wrapper Methods
4.5. Feature Weighting
4.5.1. Term Frequency 56
4.5.2. Document Frequency 57
4.5.3. Product of Term Frequency and Inverse Document Frequency 58
4.5.3. Document Normalization 58
4.6. Proposed Feature Selection Method based on TBM 59
4.7. Summary 63

5. An Evident Theoretic $k$-NN Algorithm for Text Categorization 64
  5.1. Introduction 64
  5.2. Review of $k$-Nearest Neighbour Classification 64
    5.2.1. Majority Voting $k$NN 65
    5.2.2. Distance Weighted $k$NN 66
    5.2.3. Similarity and Distance Metrics 67
  5.3. Evidence Representation in $k$NN 68
    5.3.1. Handling Uncertainty 69
    5.3.2. Resolving Uncertainty thru Evidence Theory 70
  5.4. An Improved Evident Theoretic $k$NN for Text Categorization 71
  5.5. Experimental Evaluation 78
  5.6. Summary 79

6. Experimental Setup 81
  6.1. Introduction 81
    6.2.1. Support Vector Machine 81
    6.2.2. $k$-Nearest Neighbour 82
  6.3. Benchmark Datasets 83
    6.3.1. Text Pre-processing 83
    6.3.2. Reuters-21578 Corpus 83
    6.3.3. WebKB Corpus 85
    6.3.4. 20 News Groups Corpus 86
  6.4. Evaluation Metrics 87
    6.4.1. Precision and Recall 87
    6.4.2. Micro and Macro Averaging 89
    6.4.3. $F_1$ Measure 90
List of Figures

2.1 Multi-Level Hierarchy in Text Classification 15
2.2 A Decision tree corresponding to the DNF rule of Table 2.1 25
2.3 Linear separation of positive examples from negative by the hyper plane with Support vectors 27
3.1 A relationship between Belief and Plausibility function in a unit interval 35
4.1 Representation of documents in a vector space model 48
4.2 An example of feature in frame of discernment 61
5.1 Classification of query patterns in 3-Nearest Neighbours 65
5.2 A situation of uncertainty caused by the existence of more than one frequent class label among the nearest neighbours. 70
5.3 A situation of uncertainty caused by the absence of any near neighbours. 70
5.4 An example of query pattern \( t \) in a frame of discernment \( L_1, L_2 \) and \( L_\infty \) 74
6.1 An Architectural diagram of Text Categorization framework 93
7.1 Comparison of COM with others in terms of Micro \( F_1 \) on Reuters for voting \( k\text{NN} \) 98
7.2 Comparison of COM with others in terms of Micro \( F_1 \) on Reuters for SVM 98
7.3 Comparison of COM with others in terms of Micro \( F_1 \) on WebKB for voting \( k\text{NN} \) 99
7.4 Comparison of COM with others in terms of Micro \( F_1 \) on WebKB for SVM 99
7.5 Performance of \( ik\text{NN} \) over other variants on Reuters dataset using ChiSquare feature selection. 104
7.6 Performance of \( ik\text{NN} \) over other variants on Reuters dataset using information gain feature selection. 105
7.7 Performance of \( ik\text{NN} \) over other variants on WebKB dataset using ChiSquare feature selection 105
7.8 Performance of \( ik\text{NN} \) over other variants on WebKB dataset using information gain feature selection 106
7.9  Performance of $i_k$NN over other variants on 20 News Groups dataset using Chi-square gain feature selection

7.10 Performance of $i_k$NN over other variants on 20 News Group dataset using information gain feature selection
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>A Rule-based classifier for the wheat category of Reuters Corpus</td>
<td>10</td>
</tr>
<tr>
<td>3.1</td>
<td>A Probability distribution for the cause of failure.</td>
<td>36</td>
</tr>
<tr>
<td>3.2</td>
<td>Dempster Combination of Expert 1 and Expert 2</td>
<td>36</td>
</tr>
<tr>
<td>4.1</td>
<td>Traditional feature selection methods used in Text Categorization.</td>
<td>54</td>
</tr>
<tr>
<td>4.2</td>
<td>Term frequency component.</td>
<td>57</td>
</tr>
<tr>
<td>4.3</td>
<td>Document frequency component.</td>
<td>58</td>
</tr>
<tr>
<td>5.1</td>
<td>Description of dataset used in the experiments</td>
<td>78</td>
</tr>
<tr>
<td>5.2</td>
<td>The classification accuracy of improved evident theoretic $k$NN algorithm</td>
<td>78</td>
</tr>
<tr>
<td>5.3</td>
<td>The classification accuracy of majority voting $k$NN algorithm.</td>
<td>79</td>
</tr>
<tr>
<td>5.4</td>
<td>The classification accuracy of distance weighted $k$NN algorithm.</td>
<td>79</td>
</tr>
<tr>
<td>6.1</td>
<td>Description of Reuters 21578 dataset.</td>
<td>84</td>
</tr>
<tr>
<td>6.2</td>
<td>Description of WebKB dataset.</td>
<td>85</td>
</tr>
<tr>
<td>6.3</td>
<td>Description of 20 News Groups dataset.</td>
<td>86</td>
</tr>
<tr>
<td>6.4</td>
<td>Confusion matrix</td>
<td>88</td>
</tr>
<tr>
<td>6.5</td>
<td>Micro and Macro averaged precision/recall measures</td>
<td>89</td>
</tr>
<tr>
<td>7.1</td>
<td>Feature Selection Metrics adapted in experiments</td>
<td>95</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>Text Categorization or Text classification</td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>Information Retrieval</td>
<td></td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
<td></td>
</tr>
<tr>
<td>kNN</td>
<td>k-Nearest Neighbour</td>
<td></td>
</tr>
<tr>
<td>IG</td>
<td>Information Gain</td>
<td></td>
</tr>
<tr>
<td>CHI</td>
<td>Chi Square</td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td>Odd Ratio</td>
<td></td>
</tr>
<tr>
<td>COM</td>
<td>Ensemble of Feature Selection metric</td>
<td></td>
</tr>
<tr>
<td>TBM</td>
<td>Transferable Belief Model</td>
<td></td>
</tr>
<tr>
<td>KE</td>
<td>Knowledge Engineering</td>
<td></td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
<td></td>
</tr>
<tr>
<td>WSD</td>
<td>Word Sense Disambiguation</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
<td></td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
<td></td>
</tr>
<tr>
<td>DS</td>
<td>Dempster-Shafer</td>
<td></td>
</tr>
<tr>
<td>VSM</td>
<td>Vector Space Model</td>
<td></td>
</tr>
<tr>
<td>BNS</td>
<td>Bi-Normal Separation</td>
<td></td>
</tr>
<tr>
<td>MI</td>
<td>Mutual Information</td>
<td></td>
</tr>
<tr>
<td>DF</td>
<td>Document Frequency</td>
<td></td>
</tr>
<tr>
<td>IDF</td>
<td>Inverse Document Frequency</td>
<td></td>
</tr>
<tr>
<td>$v_k$NN</td>
<td>Voting $k$nn</td>
<td></td>
</tr>
<tr>
<td>$w_k$NN</td>
<td>Weighted $k$nn</td>
<td></td>
</tr>
<tr>
<td>$i_k$NN</td>
<td>Improved evident theoretic $k$nn</td>
<td></td>
</tr>
</tbody>
</table>
NOTATIONS

\( D \)  
a domain of documents

\( d_j \)  
a document \( j \)

\( c_i \)  
a category \( i \)

\(|C|\)  
number of categories

\(|D|\)  
number of documents

\( \vec{d}_j \)  
a document vector \( j \)

\(|T|\)  
total number of terms in a dictionary

\(|T_r|\)  
total number of training documents

\( w_{kj} \)  
a weight of a term \( t_k \) in a document \( j \)

\( t_k \)  
a term in a document

\( \Phi \)  
unknown target function

\( \hat{\Phi} \)  
a function which approximate the target function

\( q \)  
a query pattern to be classified

\( h \)  
number of neighbourhoods

\( E_i \)  
a neighbourhood \( i \)

\( \Omega \)  
a frame of discernment

\( \omega_i \)  
a possible hypothesis in a frame

\( m(\omega_i) \)  
\textit{mass or bpa} value associated with a hypothesis

\( N_k(q) \)  
\( k \) nearest neighbours of \( q \) in a neighbourhood

\( K \)  
amount of conflict or normalising factor

\( k \)  
number of neighbours considered

\( \text{BetP}() \)  
pignistic probability function

\( A, B, C, X, Y, Z \)  
finite set of objects

\( \gamma, \beta \)  
control parameters

\(|W|\)  
number of Dimensions in a vector

\( F^{-1} \)  
cumulative inverse function