

# Table of Contents

<b>PART 1. STATE OF THE ART</b> . . . . .	1
<b>Chapter 1. Introduction</b> . . . . .	3
1.1. Organization of the book . . . . .	6
1.2. Utterance corpus . . . . .	8
1.3. Datasets from the UCI repository . . . . .	10
1.3.1. Wine dataset (wine) . . . . .	10
1.3.2. Wisconsin breast cancer dataset (breast) . . . . .	11
1.3.3. Handwritten digits dataset (Pendig) . . . . .	11
1.3.4. Pima Indians diabetes (diabetes) . . . . .	12
1.3.5. Iris dataset (Iris) . . . . .	13
1.4. Microarray dataset . . . . .	13
1.5. Simulated datasets . . . . .	14
1.5.1. Mixtures of Gaussians . . . . .	14
1.5.2. Spatial datasets with non-homogeneous inter-cluster distance . . . . .	14
<b>Chapter 2. State of the Art in Clustering and Semi-Supervised Techniques</b> . . . . .	15
2.1. Introduction . . . . .	15
2.2. Unsupervised machine learning (clustering) . . . . .	15
2.3. A brief history of cluster analysis . . . . .	16
2.4. Cluster algorithms . . . . .	19
2.4.1. Hierarchical algorithms . . . . .	19

2.4.1.1. Agglomerative clustering . . . . .	19
2.4.1.2. Divisive algorithms . . . . .	23
2.4.2. Model-based clustering . . . . .	24
2.4.2.1. The expectation maximization (EM) algorithm . . . . .	25
2.4.3. Partitional competitive models . . . . .	30
2.4.3.1. $K$ -means . . . . .	30
2.4.3.2. Neural gas . . . . .	35
2.4.3.3. Partitioning around Medoids (PAM) . .	37
2.4.3.4. Self-organizing maps . . . . .	39
2.4.4. Density-based clustering . . . . .	45
2.4.4.1. Direct density reachability . . . . .	45
2.4.4.2. Density reachability . . . . .	46
2.4.4.3. Density connection . . . . .	46
2.4.4.4. Border points . . . . .	47
2.4.4.5. Noise points . . . . .	47
2.4.4.6. DBSCAN algorithm . . . . .	47
2.4.5. Graph-based clustering . . . . .	49
2.4.5.1. Pole-based overlapping clustering . . .	49
2.4.6. Affection stage . . . . .	52
2.4.6.1. Advantages and drawbacks . . . . .	52
2.5. Applications of cluster analysis . . . . .	52
2.5.1. Image segmentation . . . . .	53
2.5.2. Molecular biology . . . . .	55
2.5.2.1. Biological considerations . . . . .	56
2.5.3. Information retrieval and document clustering . . . . .	60
2.5.3.1. Document pre-processing . . . . .	61
2.5.3.2. Boolean model representation . . . . .	63
2.5.3.3. Vector space model . . . . .	64
2.5.3.4. Term weighting . . . . .	65
2.5.3.5. Probabilistic models . . . . .	71
2.5.4. Clustering documents in information retrieval . . . . .	76
2.5.4.1. Clustering of presented results . . . . .	76

2.5.4.2. Post-retrieval document browsing (Scatter-Gather) . . . . .	76
2.6. Evaluation methods . . . . .	77
2.7. Internal cluster evaluation . . . . .	77
2.7.1. Entropy . . . . .	78
2.7.2. Purity . . . . .	78
2.7.3. Normalized mutual information . . . . .	79
2.8. External cluster validation . . . . .	80
2.8.1. Hartigan . . . . .	80
2.8.2. Davies Bouldin index . . . . .	81
2.8.3. Krzanowski and Lai index . . . . .	81
2.8.4. Silhouette . . . . .	82
2.8.5. Gap statistic . . . . .	82
2.9. Semi-supervised learning . . . . .	84
2.9.1. Self training . . . . .	84
2.9.2. Co-training . . . . .	85
2.9.3. Generative models . . . . .	86
2.10. Summary . . . . .	88

**PART 2. APPROACHES TO SEMI-SUPERVISED  
CLASSIFICATION . . . . . 91**

**Chapter 3. Semi-Supervised Classification Using  
Prior Word Clustering . . . . . 93**

3.1. Introduction . . . . .	93
3.2. Dataset . . . . .	94
3.3. Utterance classification scheme . . . . .	94
3.3.1. Pre-processing . . . . .	94
3.3.1.1. Utterance vector representation . . . . .	96
3.3.2. Utterance classification . . . . .	96
3.4. Semi-supervised approach based on term clustering . . . . .	98
3.4.1. Term clustering . . . . .	99
3.4.2. Semantic term dissimilarity . . . . .	100
3.4.2.1. Term vector of lexical co-occurrences . . . . .	101
3.4.2.2. Metric of dissimilarity . . . . .	102

3.4.3. Term vector truncation . . . . .	104
3.4.4. Term clustering . . . . .	105
3.4.5. Feature extraction and utterance feature vector . . . . .	109
3.4.6. Evaluation . . . . .	110
3.5. Disambiguation . . . . .	113
3.5.1. Evaluation . . . . .	116
3.6. Summary . . . . .	124

## **Chapter 4. Semi-Supervised Classification Using Pattern Clustering . . . . . 127**

4.1. Introduction . . . . .	127
4.2. New semi-supervised algorithm using the cluster and label strategy . . . . .	128
4.2.1. Block diagram . . . . .	128
4.2.1.1. Dataset . . . . .	129
4.2.1.2. Clustering . . . . .	130
4.2.1.3. Optimum cluster labeling . . . . .	130
4.2.1.4. Classification . . . . .	131
4.3. Optimum cluster labeling . . . . .	132
4.3.1. Problem definition . . . . .	132
4.3.2. The Hungarian algorithm . . . . .	134
4.3.2.1. Weighted complete bipartite graph . . . . .	134
4.3.2.2. Matching, perfect matching and maximum weight matching . . . . .	135
4.3.2.3. Objective of Hungarian method . . . . .	136
4.3.2.4. Complexity considerations . . . . .	141
4.3.3. Genetic algorithms . . . . .	142
4.3.3.1. Reproduction operators . . . . .	143
4.3.3.2. Forming the next generation . . . . .	146
4.3.3.3. GAs applied to optimum cluster labeling . . . . .	147
4.3.3.4. Comparison of methods . . . . .	150
4.4. Supervised classification block . . . . .	154
4.4.1. Support vector machines . . . . .	154

- 4.4.1.1. The kernel trick for nonlinearly separable classes . . . . . 156
- 4.4.1.2. Multi-class classification . . . . . 157
- 4.4.2. Example . . . . . 157
- 4.5. Datasets . . . . . 159
- 4.5.1. Mixtures of Gaussians . . . . . 159
- 4.5.2. Datasets from the UCI repository . . . . . 159
  - 4.5.2.1. Iris dataset (Iris) . . . . . 159
  - 4.5.2.2. Wine dataset (wine) . . . . . 160
  - 4.5.2.3. Wisconsin breast cancer dataset (breast) . . . . . 160
  - 4.5.2.4. Handwritten digits dataset (Pendig) . . . . . 160
  - 4.5.2.5. Pima Indians diabetes (diabetes) . . . . . 160
- 4.5.3. Utterance dataset . . . . . 160
- 4.6. An analysis of the bounds for the cluster and label approaches . . . . . 162
- 4.7. Extension through cluster pruning . . . . . 164
  - 4.7.1. Determination of silhouette thresholds . . . . . 166
  - 4.7.2. Evaluation of the cluster pruning approach . . . . . 171
- 4.8. Simulations and results . . . . . 173
- 4.9. Summary . . . . . 179

**PART 3 . CONTRIBUTIONS TO UNSUPERVISED CLASSIFICATION – ALGORITHMS TO DETECT THE OPTIMAL NUMBER OF CLUSTERS . . . . . 183**

**Chapter 5. Detection of the Number of Clusters through Non-Parametric Clustering Algorithms . . . . . 185**

- 5.1. Introduction . . . . . 185
- 5.2. New hierarchical pole-based clustering algorithm . . . . . 186
  - 5.2.1. Pole-based clustering basis module . . . . . 187
  - 5.2.2. Hierarchical pole-based clustering . . . . . 189
- 5.3. Evaluation . . . . . 190
  - 5.3.1. Cluster evaluation metrics . . . . . 191
- 5.4. Datasets . . . . . 192

5.4.1. Results . . . . .	192
5.4.2. Complexity considerations for large databases . . . . .	195
5.5. Summary . . . . .	197
<b>Chapter 6. Detecting the Number of Clusters through Cluster Validation . . . . .</b>	<b>199</b>
6.1. Introduction . . . . .	199
6.2. Cluster validation methods . . . . .	201
6.2.1. Dunn index . . . . .	201
6.2.2. Hartigan . . . . .	201
6.2.3. Davies Bouldin index . . . . .	202
6.2.4. Krzanowski and Lai index . . . . .	202
6.2.5. Silhouette . . . . .	203
6.2.6. Hubert's $\gamma$ . . . . .	204
6.2.7. Gap statistic . . . . .	205
6.3. Combination approach based on quantiles . . . . .	206
6.4. Datasets . . . . .	212
6.4.1. Mixtures of Gaussians . . . . .	212
6.4.2. Cancer DNA-microarray dataset . . . . .	213
6.4.3. Iris dataset . . . . .	214
6.5. Results . . . . .	214
6.5.1. Validation results of the five Gaussian dataset . . . . .	215
6.5.2. Validation results of the mixture of seven Gaussians . . . . .	220
6.5.3. Validation results of the NCI60 dataset . . . . .	220
6.5.4. Validation results of the Iris dataset . . . . .	221
6.5.5. Discussion . . . . .	222
6.6. Application of speech utterances . . . . .	223
6.7. Summary . . . . .	224
<b>Bibliography . . . . .</b>	<b>227</b>
<b>Index . . . . .</b>	<b>243</b>