

Contents

| | |
|---|-------------|
| Acknowledgments | iii |
| List of Figures | x |
| List of Tables | xii |
| Abbreviations | xiii |
| Abstract | xiv |
| 1 Introduction | 1 |
| 1.1 Motivations for Dimensionality Reduction | 4 |
| 1.1.1 Curse of Dimensionality and Empty Space Phenomenon | 5 |
| 1.1.2 The Peaking Phenomenon | 5 |
| 1.1.3 Irrelevant and Redundant Features | 6 |
| 1.2 Techniques for Dimensionality Reduction | 6 |
| 1.2.1 Feature Selection Algorithms | 6 |
| 1.2.2 Feature Extraction Algorithms | 7 |
| 1.3 Research Contributions | 7 |
| 1.4 Organization of this Dissertation | 10 |
| 1.5 Notations Used | 12 |
| 2 Feature Selection | 13 |
| 2.1 Feature Selection and Learning Problems | 14 |
| 2.1.1 Unsupervised Feature Selection | 14 |
| 2.1.2 Supervised Feature Selection | 14 |
| 2.2 Stages in Feature Selection | 15 |
| 2.2.1 Feature Subset Generation | 15 |
| 2.2.2 Feature Subset Evaluation | 16 |
| 2.2.3 Stopping Criterion | 16 |
| 2.2.4 Result Validation | 16 |
| 2.3 Categories of Feature Selection Algorithms | 17 |
| 2.3.1 Feature Ranking Algorithms | 17 |
| 2.3.1.1 Ranking Based on Correlation Measures | 18 |
| 2.3.1.2 Ranking Based on Information Theory | 20 |

| | | |
|----------|---|-----------|
| 2.3.1.3 | Ranking Based on Distances between Probability Distributions | 21 |
| 2.3.2 | Feature Subset Selection Algorithms | 23 |
| 2.3.2.1 | Filter Methods | 24 |
| 2.3.2.2 | Wrapper Methods | 26 |
| 2.3.2.3 | Embedded Methods | 27 |
| 2.3.2.4 | FR versus FSS for Feature Selection | 27 |
| 2.3.3 | Feature Selection and Causality | 27 |
| 2.4 | Feature Selection Related Issues | 30 |
| 2.4.1 | Evaluation of Feature Selection Algorithms | 30 |
| 2.4.2 | Stability of Feature Selection Algorithms | 31 |
| 2.5 | Summary | 32 |
| 3 | Evaluating Rankings of Binary Features for Feature Selection | 33 |
| 3.1 | The Diff-criterion: A New Feature Weighting Scheme | 34 |
| 3.1.1 | Class-dependent Density of a Binary Feature | 35 |
| 3.1.2 | Class-dependent Sparsity of a Binary Feature | 35 |
| 3.1.3 | The Max-criterion | 36 |
| 3.1.4 | A Motivating Example | 36 |
| 3.1.5 | The Diff-criterion | 37 |
| 3.1.6 | Time Complexity | 38 |
| 3.2 | CDFE: A New Feature Ranking Algorithm | 38 |
| 3.2.1 | Time Complexity | 41 |
| 3.3 | Rationale of the Diff-criterion | 41 |
| 3.3.1 | Mathematical Analysis | 41 |
| 3.3.2 | Graphical Analysis | 46 |
| 3.3.3 | Comparing the Characteristics of Diff-criterion and Mutual Information | 48 |
| 3.3.3.1 | Range of the Weight values | 48 |
| 3.3.3.2 | Time-Complexity | 49 |
| 3.4 | Experiments | 50 |
| 3.4.1 | Data Description | 50 |
| 3.4.2 | Performance Evaluation | 51 |
| 3.4.3 | Evaluation on Data sets with Low Dimensionality | 53 |
| 3.4.3.1 | LUCAS0: The Lung Cancer Data set | 53 |
| 3.4.3.2 | SPECT: The Heart Disease Data set | 54 |
| 3.4.4 | Evaluation on Data sets with High Dimensionality | 57 |
| 3.4.4.1 | GINA: The Handwriting Recognition Data set | 58 |
| 3.4.4.2 | HIVA: The Chemoinformatics Data set | 60 |
| 3.4.4.3 | NOVA: The Text Classification Data set | 62 |
| 3.4.4.4 | DOROTHEA: The Drug Discovery Data set | 63 |
| 3.4.5 | Discussion of Results | 66 |
| 3.4.6 | Comparison of CDFE against Winning Entries of the Agnostic Learning Track | 70 |

| | | |
|-----------|--|-----------|
| 3.5 | Summary and Conclusions | 72 |
| 4 | Evaluating Orderings of Binary Features for Correctness | 73 |
| 4.1 | The Correctness Problem | 74 |
| 4.1.1 | A Motivating Example | 75 |
| 4.1.2 | The Importance of Feature Ranking Correctness | 76 |
| 4.1.2.1 | Feature Analysis | 76 |
| 4.1.2.2 | Feature Selection | 77 |
| 4.2 | Related Work | 77 |
| 4.3 | Feature Ranking Evaluation Strategy (FRES) | 78 |
| 4.3.1 | Time Complexity | 82 |
| 4.4 | Experiments | 82 |
| 4.4.1 | Synthetic Data | 83 |
| 4.4.1.1 | Data Description | 83 |
| 4.4.1.2 | Results | 83 |
| 4.4.2 | Real Life Data | 85 |
| 4.4.2.1 | The Working of FRES Does Not Depend upon a Classifier | 85 |
| 4.4.2.2 | Results | 87 |
| 4.4.2.2.1 | Evaluation on Data sets with Low Dimensionality | 87 |
| 4.4.2.2.2 | Evaluation on Data sets with High Dimensionality | 91 |
| 4.4.3 | Discussion of Results | 94 |
| 4.5 | Summary and Conclusions | 98 |
| 5 | Two Stage Feature Selection for High-Dimensional Binary Data | 99 |
| 5.1 | Two-Stage Feature Selection Algorithms Based on the Diff-criterion | 100 |
| 5.1.1 | First Stage: Selection of the Preliminary Feature Subset | 101 |
| 5.1.2 | Second Stage: Selection of the Final Feature Subset | 102 |
| 5.1.2.1 | Koller and Sahami's Markov Blanket Filtering (MBF) Algorithm | 103 |
| 5.1.2.2 | Bernoulli Mixture Model-Based Markov Blanket Filtering (BMM-MBF) Algorithm | 104 |
| 5.2 | Experiments | 106 |
| 5.2.1 | Performance Evaluation | 106 |
| 5.2.2 | Stage-1: Class-Dependent Density-Based Feature Elimination | 107 |
| 5.2.3 | Evaluation of the Two-Stage Algorithm with the MBF Algorithm in Second Stage | 108 |
| 5.2.3.1 | GINA: The Handwriting Recognition Data set | 109 |
| 5.2.3.2 | HIVA: The Chemoinformatics Data set | 110 |

| | | |
|----------|--|------------|
| 5.2.3.3 | NOVA: The Text Classification Data set | 112 |
| 5.2.3.4 | DOROTHEA: The Drug Discovery Data set | 114 |
| 5.2.4 | Evaluation of the Two-Stage Algorithm with the BMM-MBF Algorithm in Second Stage | 115 |
| 5.2.4.1 | GINA: The Handwriting Recognition Data set | 115 |
| 5.2.4.2 | HIVA: The Chemoinformatics Data set | 117 |
| 5.2.4.3 | NOVA: The Text Classification Data set | 118 |
| 5.2.4.4 | DOROTHEA: The Drug Discovery Data set | 120 |
| 5.3 | Discussion of Results | 121 |
| 5.4 | Comparison of Two-Stage Algorithms against the Winning Entries of Agnostic Learning Track | 123 |
| 5.5 | Summary and Conclusions | 124 |
| 6 | Conclusions | 126 |
| 6.1 | Summary of Research Contributions | 127 |
| 6.2 | Future Work | 129 |
| 6.3 | Concluding Remarks | 130 |
| A | Data Sets | 131 |
| A.1 | LUCAS0: The Lung Cancer Data Set | 131 |
| A.2 | SPECT: The Heart Disease Data Set | 131 |
| A.3 | GINA: The Handwriting Recognition Data Set | 132 |
| A.4 | HIVA: The Chemoinformatics Data Set | 132 |
| A.5 | NOVA: The Text Mining Data Set | 133 |
| A.6 | DOROTHEA: The Drug Discovery Data Set | 133 |
| B | Experimental Setup | 134 |
| B.1 | Challenge Learning Object Package (CLOP) | 134 |
| B.1.1 | Classifiers | 134 |
| B.1.1.1 | The Naive Bayes' Classifier | 134 |
| B.1.1.2 | The Kernel Ridge Regression Classifier | 135 |
| B.1.1.3 | The Support Vector Machine with Radial Basis Function Kernel Classifier | 135 |
| B.1.1.4 | The Logistic Regression Classifier | 135 |
| B.1.2 | Cross-Validation | 135 |
| | References | 137 |