Contents

1 Introduction **1**

1.1 Knowledge Discovery in Databases and Data Mining 1

1.2 Focusing for Data Mining 4

1.3 Overview 6

2 Knowledge Discovery in Databases **11**

2.1 Knowledge Discovery Process 11

2.1.1 Humans in the Loop 11

2.1.2 KDD Project Phases 13

2.2 Data Preparation 16

2.2.1 From Business Data to Data Mining Input 17

2.2.2 Data Selection and Focusing 19

2.3 Data Mining Goals 20

2.3.1 From Understanding to Predictive Modeling 21

2.3.2 Classification 23

2.4 Data Characteristics: Notations and Definitions 25

2.4.1 Database Tables 25

2.4.2 Statistical Values 29

2.5 Data Mining Algorithms 31

2.5.1 Classification Algorithms 32

2.5.2 Top Down Induction of Decision Trees 33

2.5.3 Nearest Neighbor Classifiers 37

2.6 Selecting the Focusing Context 44

3 Focusing Tasks 45

3.1 Focusing Concepts: An Overview 45

3.2 Focusing Specification 47

3.2.1 Focusing Input 48

3.2.2 Focusing Output 49

3.2.3 Focusing Criterion 50

3.3 Focusing Context 52

3.3.1 Data Characteristics 54

3.3.2 Data Mining Algorithms 55

3.4 Focusing Success 55

3.4.1 Filter Evaluation 57

3.4.2 Wrapper Evaluation 64

3.4.3 Evaluation Criteria 70

3.5 Selecting the Focusing Task 83

4 Focusing Solutions 85

4.1 State of the Art: A Unifying View 85

4.1.1 The Unifying Framework of Existing Focusing Solutions . 85

4.1.2 Sampling 87

4.1.3 Clustering 95

4.1.4 Prototyping 104

4.2 More Intelligent Sampling Techniques 109

4.2.1 Existing Reusable Components 111

4.2.2 Advanced Leader Sampling 113

4.2.3 Similarity-Driven Sampling 134

4.3 A Unified Approach to Focusing Solutions 149

4.3.1 Generic Sampling 150

4.3.2 Generic Sampling in a Commercial Data Mining System . 153

5 Analytical Studies 159

5.1 An Average Case Analysis 159

5.2 Experimental Validation of Theoretical Claims 170

6 Experimental Results 173

6.1 Experimental Design 173

6.1.1 Experimental Procedure 173

6.1.2 Data Characteristics 179

6.2 Results and Evaluation 182

6.2.1 Filter Evaluation 182

6.2.2 Wrapper Evaluation for C4.5 188

6.2.3 Wrapper Evaluation for IB 195

6.2.4 Comparing Filter and Wrapper Evaluation for C4.5 . . . 201

6.2.5 Comparing Filter and Wrapper Evaluation for IB 208

6.2.6 Comparing Wrapper Evaluation for C4.5 and IB 215

6.3 Focusing Advice 222

6.3.1 Sorting, Stratification, and Prototype Weighting 222

6.3.2 Focusing Solutions in Focusing Contexts 223

7 Conclusions 231

7.1 Summary and Contributions 231

7.2 More Related Work 235

7.3 Future Work 236

7.4 Closing Remarks 238

Bibliography 239

Acknowledgments 253

A Notations 257

A.1 Indices, Variables, and Functions 257

A.2 Algorithms and Procedures 264

B More Evaluation Criteria 267

B.1 Filter Evaluation Criteria 267

B.2 Wrapper Evaluation Criteria 272

C Remaining Proofs **277**

D Generic Sampling in `GenSam` **281**

E More Experimental Results **283**

Index **303**

Curriculum Vitae **309**