

Lecture Notes in Artificial Intelligence

1623

Subseries of Lecture Notes in Computer Science

Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science

Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Springer

Berlin

Heidelberg

New York

Barcelona

Hong Kong

London

Milan

Paris

Singapore

Tokyo

Thomas Reinartz

Focusing Solutions for Data Mining

Analytical Studies and Experimental Results
in Real-World Domains



Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Author

Thomas Reinartz
DaimlerChrysler AG, Research and Technology
Wilhelm-Runge-Straße 11, D-89081 Ulm, Germany
E-mail: thomas.reinartz@daimlerchrysler.com

Cataloging-in-Publication data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Thomas Reinartz:

Focusing solutions for data mining : analytical studies and experimental results in real-world domains / Thomas Reinartz. - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo : Springer, 1999 (Lecture notes in computer science ; 1623 : Lecture notes in artificial intelligence) ISBN 3-540-66429-7

CR Subject Classification (1998): I.2, F.2, H.3, J.1, J.2

ISBN 3-540-66429-7 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

© Springer-Verlag Berlin Heidelberg 1999
Printed in Germany

Typesetting: Camera-ready by author
SPIN 10705288 06/3142 - 5 4 3 2 1 0 Printed on acid-free paper

To My Family

Preface

This dissertation develops, analyzes, and evaluates focusing solutions for data mining. Data mining is a particular phase in knowledge discovery that applies learning techniques to identify hidden information from data, whereas knowledge discovery is a complex, iterative, and interactive process which covers all activities before and after data mining. Focusing is a specific task in the data preparation phase of knowledge discovery. The motivation of focusing is the existence of huge databases and the limitation of data mining algorithms to smaller data sets. The purpose of focusing is data reduction before data mining, either in the number of tuples, the number of attributes, or the number of values. Then, data mining applies techniques to the reduced data and is still able to achieve appropriate results.

In this dissertation, we first analyze the knowledge discovery process in order to understand relations between knowledge discovery tasks and focusing. We characterize the focusing context which consists of a data mining goal, data characteristics, and a data mining algorithm. We emphasize classification goals, top down induction of decision trees, and nearest neighbor classifiers. Thereafter, we define focusing tasks which include evaluation criteria for focusing success. At the end of the first block, we restrict our attention to focusing tasks for the reduction of the number of tuples.

We start the development of focusing solutions with an analysis of state-of-the-art approaches. We define a unifying framework that builds on three basic techniques: Sampling, clustering, and prototyping. We describe instantiations of this framework and examine their advantages and disadvantages. We follow up the unifying framework and establish an enhanced unified approach to focusing solutions which covers two preparation steps, sorting and stratification, and the application of sampling techniques. We reuse random sampling and systematic sampling from statistics and propose two more intelligent sampling techniques, leader sampling and similarity-driven sampling. We implement the unified approach as a generic sampling algorithm and integrate this algorithm into a commercial data mining system.

Thereafter, we analyze and evaluate specific focusing solutions in different domains. We exemplify an average case analysis to estimate expected average

classification accuracies of nearest neighbor classifiers in combination with simple random sampling. We further conduct an experimental study and consolidate its results as focusing advice which provides heuristics for appropriate selections of best suited focusing solutions. At the end, we summarize the main contributions of this dissertation, describe more related work, raise issues for future work, and state some final remarks.

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

List of Figures

1.1	Structure and Contents of Dissertation	7
2.1	Humans in the Loop of KDD	12
2.2	KDD Phases	14
2.3	Data Preparation Tasks	17
2.4	Data Selection and Focusing	20
2.5	Data Mining Goals	21
2.6	Decision Tree Example	34
3.1	Focusing Concepts	46
3.2	Focusing Specifications	47
3.3	Focusing Context	52
3.4	Focusing Context Example	54
3.5	Evaluation Strategies for Focusing Success	56
4.1	The Unifying Framework of Existing Focusing Solutions	86
4.2	Classification Criteria for Clustering Techniques in Statistics	98
4.3	Strata Tree	124
4.4	Equal-Width and Equal-Frequency Discretization	129
4.5	Similarity Thresholds and Focusing Output Size in LEASAM	137
4.6	Leader Sampling and Hierarchical Clustering	141
4.7	Generic Sampling	151
4.8	CLEMENTINE Data Mining System	154
4.9	CITRUS Architecture	155
4.10	Generic Sampling in CLEMENTINE	157

5.1 Well-Separated Clustering and Not Well-Separated Clustering . . 162

5.2 Experimental Validation of Average Case Analysis for 16 Clusters 171

5.3 Experimental Validation of Average Case Analysis for 25 Clusters 172

6.1 Experimental Procedure 174

6.2 Minimum Filter Evaluation 183

6.3 Minimum Wrapper Evaluation for C4.5 189

6.4 Minimum Wrapper Evaluation for IB 196

6.5 Filter Evaluation and Wrapper Evaluation for C4.5 202

6.6 Filter Evaluation and Wrapper Evaluation for IB 209

6.7 Wrapper Evaluation for C4.5 and Wrapper Evaluation for IB . . 216

E.1 Minimum Filter Evaluation II 284

E.2 Average Filter Evaluation 285

E.3 Average Filter Evaluation II 286

E.4 Maximum Filter Evaluation 287

E.5 Maximum Filter Evaluation II 288

E.6 Minimum Wrapper Evaluation for C4.5 II 289

E.7 Average Wrapper Evaluation for C4.5 290

E.8 Average Wrapper Evaluation for C4.5 II 291

E.9 Maximum Wrapper Evaluation for C4.5 292

E.10 Maximum Wrapper Evaluation for C4.5 II 293

E.11 Minimum Wrapper Evaluation for IB II 294

E.12 Average Wrapper Evaluation for IB 295

E.13 Average Wrapper Evaluation for IB II 296

E.14 Maximum Wrapper Evaluation for IB 297

E.15 Maximum Wrapper Evaluation for IB II 298

E.16 Filter Evaluation and Wrapper Evaluation for C4.5 II 299

E.17 Filter Evaluation and Wrapper Evaluation for IB II 300

E.18 Wrapper Evaluation for C4.5 and Wrapper Evaluation for IB II . 301

List of Tables

2.1	(Database) Table	26
2.2	(Database) Table Example	27
2.3	Similarity Matrix Example	43
3.1	Two Types of Errors in Hypothesis Testing	60
3.2	Isolated and Comparative Filter Evaluation Criteria	70
3.3	Isolated and Comparative Wrapper Evaluation Criteria	78
4.1	Existing Reusable Components in State of the Art Efforts	110
4.2	Similarity Matrix Example	138
4.3	Focusing Contexts and Attribute Relevance	150
6.1	Parameter Settings in GENSAM	177
6.2	η Values for Evaluation Criteria	178
6.3	A Selection of UCI Databases	180
6.4	Rankings of Results with Filter Evaluation	185
6.5	Rankings of Results with Wrapper Evaluation for C4.5	190
6.6	Rankings of Results with Wrapper Evaluation for IB	197
6.7	Focusing Advice for Filter Evaluation	225
6.8	Focusing Advice for Wrapper Evaluation for C4.5	226
6.9	Focusing Advice for Wrapper Evaluation for IB	227
A.1	Notations for Indices, Variables, and Functions	257
A.2	Notations for Algorithms and Procedures	264
E.1	More Experimental Results	283