

Charu C. Aggarwal

Data Mining

The Textbook

 Springer

Churu C. Aggarwal

Data Mining

The Textbook

 Springer

Charu C. Aggarwal
IBM T.J. Watson Research Center
Yorktown Heights
New York
USA

A solution manual for this book is available on Springer.com.

ISBN 978-3-319-14141-1 ISBN 978-3-319-14142-8 (eBook)
DOI 10.1007/978-3-319-14142-8

Library of Congress Control Number: 2015930833

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Printed by Markono Print Media Pte Ltd

To my wife Lata,
and my daughter Sayani

Contents

1	An Introduction to Data Mining	1
1.1	Introduction	1
1.2	The Data Mining Process	3
1.2.1	The Data Preprocessing Phase	5
1.2.2	The Analytical Phase	6
1.3	The Basic Data Types	6
1.3.1	Nondependency-Oriented Data	7
1.3.1.1	Quantitative Multidimensional Data	7
1.3.1.2	Categorical and Mixed Attribute Data	8
1.3.1.3	Binary and Set Data	8
1.3.1.4	Text Data	8
1.3.2	Dependency-Oriented Data	9
1.3.2.1	Time-Series Data	9
1.3.2.2	Discrete Sequences and Strings	10
1.3.2.3	Spatial Data	11
1.3.2.4	Network and Graph Data	12
1.4	The Major Building Blocks: A Bird's Eye View	14
1.4.1	Association Pattern Mining	15
1.4.2	Data Clustering	16
1.4.3	Outlier Detection	17
1.4.4	Data Classification	18
1.4.5	Impact of Complex Data Types on Problem Definitions	19
1.4.5.1	Pattern Mining with Complex Data Types	20
1.4.5.2	Clustering with Complex Data Types	20
1.4.5.3	Outlier Detection with Complex Data Types	21
1.4.5.4	Classification with Complex Data Types	21
1.5	Scalability Issues and the Streaming Scenario	21
1.6	A Stroll Through Some Application Scenarios	22
1.6.1	Store Product Placement	22
1.6.2	Customer Recommendations	23
1.6.3	Medical Diagnosis	23
1.6.4	Web Log Anomalies	24
1.7	Summary	24

1.8	Bibliographic Notes	25
1.9	Exercises	25
2	Data Preparation	27
2.1	Introduction	27
2.2	Feature Extraction and Portability	28
2.2.1	Feature Extraction	28
2.2.2	Data Type Portability	30
2.2.2.1	Numeric to Categorical Data: Discretization	30
2.2.2.2	Categorical to Numeric Data: Binarization	31
2.2.2.3	Text to Numeric Data	31
2.2.2.4	Time Series to Discrete Sequence Data	32
2.2.2.5	Time Series to Numeric Data	32
2.2.2.6	Discrete Sequence to Numeric Data	33
2.2.2.7	Spatial to Numeric Data	33
2.2.2.8	Graphs to Numeric Data	33
2.2.2.9	Any Type to Graphs for Similarity-Based Applications	33
2.3	Data Cleaning	34
2.3.1	Handling Missing Entries	35
2.3.2	Handling Incorrect and Inconsistent Entries	36
2.3.3	Scaling and Normalization	37
2.4	Data Reduction and Transformation	37
2.4.1	Sampling	38
2.4.1.1	Sampling for Static Data	38
2.4.1.2	Reservoir Sampling for Data Streams	39
2.4.2	Feature Subset Selection	40
2.4.3	Dimensionality Reduction with Axis Rotation	41
2.4.3.1	Principal Component Analysis	42
2.4.3.2	Singular Value Decomposition	44
2.4.3.3	Latent Semantic Analysis	47
2.4.3.4	Applications of PCA and SVD	48
2.4.4	Dimensionality Reduction with Type Transformation	49
2.4.4.1	Haar Wavelet Transform	50
2.4.4.2	Multidimensional Scaling	55
2.4.4.3	Spectral Transformation and Embedding of Graphs	57
2.5	Summary	59
2.6	Bibliographic Notes	60
2.7	Exercises	61
3	Similarity and Distances	63
3.1	Introduction	63
3.2	Multidimensional Data	64
3.2.1	Quantitative Data	64
3.2.1.1	Impact of Domain-Specific Relevance	65
3.2.1.2	Impact of High Dimensionality	65
3.2.1.3	Impact of Locally Irrelevant Features	66
3.2.1.4	Impact of Different L_p -Norms	67
3.2.1.5	Match-Based Similarity Computation	68
3.2.1.6	Impact of Data Distribution	69

	3.2.1.7	Nonlinear Distributions: ISOMAP	70
	3.2.1.8	Impact of Local Data Distribution	72
	3.2.1.9	Computational Considerations	73
	3.2.2	Categorical Data	74
	3.2.3	Mixed Quantitative and Categorical Data	75
3.3		Text Similarity Measures	75
	3.3.1	Binary and Set Data	77
3.4		Temporal Similarity Measures	77
	3.4.1	Time-Series Similarity Measures	77
	3.4.1.1	Impact of Behavioral Attribute Normalization	78
	3.4.1.2	L_p -Norm	79
	3.4.1.3	Dynamic Time Warping Distance	79
	3.4.1.4	Window-Based Methods	82
	3.4.2	Discrete Sequence Similarity Measures	82
	3.4.2.1	Edit Distance	82
	3.4.2.2	Longest Common Subsequence	84
3.5		Graph Similarity Measures	85
	3.5.1	Similarity between Two Nodes in a Single Graph	85
	3.5.1.1	Structural Distance-Based Measure	85
	3.5.1.2	Random Walk-Based Similarity	86
	3.5.2	Similarity Between Two Graphs	86
3.6		Supervised Similarity Functions	87
3.7		Summary	88
3.8		Bibliographic Notes	89
3.9		Exercises	90
4		Association Pattern Mining	93
	4.1	Introduction	93
	4.2	The Frequent Pattern Mining Model	94
	4.3	Association Rule Generation Framework	97
	4.4	Frequent Itemset Mining Algorithms	99
	4.4.1	Brute Force Algorithms	99
	4.4.2	The Apriori Algorithm	100
	4.4.2.1	Efficient Support Counting	102
	4.4.3	Enumeration-Tree Algorithms	103
	4.4.3.1	Enumeration-Tree-Based Interpretation of Apriori	105
	4.4.3.2	TreeProjection and DepthProject	106
	4.4.3.3	Vertical Counting Methods	110
	4.4.4	Recursive Suffix-Based Pattern Growth Methods	112
	4.4.4.1	Implementation with Arrays but No Pointers	114
	4.4.4.2	Implementation with Pointers but No FP-Tree	114
	4.4.4.3	Implementation with Pointers and FP-Tree	116
	4.4.4.4	Trade-offs with Different Data Structures	118
	4.4.4.5	Relationship Between FP-Growth and Enumeration-Tree Methods	119
	4.5	Alternative Models: Interesting Patterns	122
	4.5.1	Statistical Coefficient of Correlation	123
	4.5.2	χ^2 Measure	123
	4.5.3	Interest Ratio	124

4.5.4	Symmetric Confidence Measures	124
4.5.5	Cosine Coefficient on Columns	125
4.5.6	Jaccard Coefficient and the Min-hash Trick	125
4.5.7	Collective Strength	126
4.5.8	Relationship to Negative Pattern Mining	127
4.6	Useful Meta-algorithms	127
4.6.1	Sampling Methods	128
4.6.2	Data Partitioned Ensembles	128
4.6.3	Generalization to Other Data Types	129
	4.6.3.1 Quantitative Data	129
	4.6.3.2 Categorical Data	129
4.7	Summary	129
4.8	Bibliographic Notes	130
4.9	Exercises	132
5	Association Pattern Mining: Advanced Concepts	135
5.1	Introduction	135
5.2	Pattern Summarization	136
	5.2.1 Maximal Patterns	136
	5.2.2 Closed Patterns	137
	5.2.3 Approximate Frequent Patterns	139
	5.2.3.1 Approximation in Terms of Transactions	139
	5.2.3.2 Approximation in Terms of Itemsets	140
5.3	Pattern Querying	141
	5.3.1 Preprocess-once Query-many Paradigm	141
	5.3.1.1 Leveraging the Itemset Lattice	142
	5.3.1.2 Leveraging Data Structures for Querying	143
	5.3.2 Pushing Constraints into Pattern Mining	146
5.4	Putting Associations to Work: Applications	147
	5.4.1 Relationship to Other Data Mining Problems	147
	5.4.1.1 Application to Classification	147
	5.4.1.2 Application to Clustering	148
	5.4.1.3 Applications to Outlier Detection	148
	5.4.2 Market Basket Analysis	148
	5.4.3 Demographic and Profile Analysis	148
	5.4.4 Recommendations and Collaborative Filtering	149
	5.4.5 Web Log Analysis	149
	5.4.6 Bioinformatics	149
	5.4.7 Other Applications for Complex Data Types	150
5.5	Summary	150
5.6	Bibliographic Notes	151
5.7	Exercises	152
6	Cluster Analysis	153
6.1	Introduction	153
6.2	Feature Selection for Clustering	154
	6.2.1 Filter Models	155
	6.2.1.1 Term Strength	155
	6.2.1.2 Predictive Attribute Dependence	155

	6.2.1.3	Entropy	156
	6.2.1.4	Hopkins Statistic	157
	6.2.2	Wrapper Models	158
6.3		Representative-Based Algorithms	159
	6.3.1	The k -Means Algorithm	162
	6.3.2	The Kernel k -Means Algorithm	163
	6.3.3	The k -Medians Algorithm	164
	6.3.4	The k -Medoids Algorithm	164
6.4		Hierarchical Clustering Algorithms	166
	6.4.1	Bottom-Up Agglomerative Methods	167
	6.4.1.1	Group-Based Statistics	169
	6.4.2	Top-Down Divisive Methods	172
	6.4.2.1	Bisecting k -Means	173
6.5		Probabilistic Model-Based Algorithms	173
	6.5.1	Relationship of EM to k -means and Other Representative Methods	176
6.6		Grid-Based and Density-Based Algorithms	178
	6.6.1	Grid-Based Methods	179
	6.6.2	DBSCAN	181
	6.6.3	DENCLUE	184
6.7		Graph-Based Algorithms	187
	6.7.1	Properties of Graph-Based Algorithms	189
6.8		Non-negative Matrix Factorization	191
	6.8.1	Comparison with Singular Value Decomposition	194
6.9		Cluster Validation	195
	6.9.1	Internal Validation Criteria	196
	6.9.1.1	Parameter Tuning with Internal Measures	198
	6.9.2	External Validation Criteria	198
	6.9.3	General Comments	201
6.10		Summary	201
6.11		Bibliographic Notes	201
6.12		Exercises	202
		Cluster Analysis: Advanced Concepts	205
7.1		Introduction	205
7.2		Clustering Categorical Data	206
	7.2.1	Representative-Based Algorithms	207
	7.2.1.1	k -Modes Clustering	208
	7.2.1.2	k -Medoids Clustering	209
	7.2.2	Hierarchical Algorithms	209
	7.2.2.1	ROCK	209
	7.2.3	Probabilistic Algorithms	211
	7.2.4	Graph-Based Algorithms	212
7.3		Scalable Data Clustering	212
	7.3.1	CLARANS	213
	7.3.2	BIRCH	214
	7.3.3	CURE	216
7.4		High-Dimensional Clustering	217
	7.4.1	CLIQUE	219
	7.4.2	PROCLUS	220