

ADVANCED TOPICS IN SCIENCE AND TECHNOLOGY IN CHINA

ADVANCED TOPICS

IN SCIENCE AND TECHNOLOGY IN CHINA

Zhejiang University is one of the leading universities in China. In Advanced Topics in Science and Technology in China, Zhejiang University Press and Springer jointly publish monographs by Chinese scholars and professors, as well as invited authors and editors from abroad who are outstanding experts and scholars in their fields. This series will be of interest to researchers, lecturers, and graduate students alike.

Advanced Topics in Science and Technology in China aims to present the latest and most cutting-edge theories, techniques, and methodologies in various research areas in China. It covers all disciplines in the fields of natural science and technology, including but not limited to, computer science, materials science, life sciences, engineering, environmental sciences, mathematics, and physics.

Zengchang Qin
Yongchuan Tang

Uncertainty Modeling for Data Mining

A Label Semantics Approach

With 61 figures



ZHEJIANG UNIVERSITY PRESS
浙江大学出版社



Springer

Authors

Prof. Zengchang Qin
Intelligent Computing and Machine
Learning Lab, School of ASEE,
Beihang University, Beijing, China
E-mail: zengchang.qin@gmail.com

Prof. Yongchuan Tang
College of Computer Science
Zhejiang University,
Hangzhou, Zhejiang, China
E-mail: tyongchuan@gmail.com

ISSN 1995-6819

e-ISSN 1995-6827

Advanced Topics in Science and Technology in China

Zhejiang University Press, Hangzhou

ISBN 978-3-642-41250-9

ISBN 978-3-642-41251-6 (eBook)

Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2013949181

© Zhejiang University Press, Hangzhou and Springer-Verlag Berlin Heidelberg 2014

This work is subject to copyright. All rights are reserved by the Publishers, 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. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publishers; locations, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law. 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.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publishers can accept any legal responsibility for any errors or omissions that may be made. The publishers make no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

*This book is dedicated to my parents
Li-zhong Qin (1939–1995) and Feng-xia
Zhang (1936–2003)*

Zengchang Qin

Preface

Uncertainty is one of the characteristics of the nature. Many theories have been proposed in dealing with uncertainties. Fuzzy logic has been one of such theories. Both of us were inspired by Zadeh's fuzzy theory and Jonathan Lawry's label semantics theory when we both worked in University of Bristol.

Machine learning and data mining are inseparably connected with uncertainty. To begin with, the observable data for learning is usually imprecise, incomplete or noisy. Even the observations are perfect, the generalization beyond that data is still afflicted with uncertainty; e.g., how can we be sure which one from a set of candidate theories that all of them explain the data. Though Occam's razor tells us to favor the simplest models, this principle does not guarantee this simple model is the truth of the data. In recent research, we have found that some complex models seem to be more appropriate comparing to simple ones because of our complex nature and the complicated mechanism of data generation in social problems.

In this book, we introduce a fuzzy logic based theory for modeling uncertainty in data mining. The content of this book can be roughly split into three parts: Chapters 1-3 give a general introduction of data mining and the basics of label semantics theory. Chapters 4-8 introduce a number of data mining algorithms based on label semantics and detailed theoretical aspects, and experimental results are given. Chapters 9-12 introduce prototype theory interpretation of label semantics and data mining algorithms developed based on this interpretation. This book is for the readers like postgraduates and researchers in AI, data mining, soft computing and other related areas.

Zengchang Qin
Pittsburgh, PA, USA
Yongchuan Tang
Hangzhou, China
July, 2013

Acknowledgements

First of all we would like to express sincere thanks to our mentors, colleagues and friends. This book could not have been written without them. Special thank goes to Prof. Jonathan Lawry, our mentor who introduced label semantics theory to us. The first author thanks Prof. Lotfi Zadeh for his insightful comments and support during his two year stay in BISC at UC Berkeley. Many people have helped in our research and providing comments and suggestions, including Trevor Martin (Bristol University), Qiang Shen (Aberystwyth University), Masoud Nikraves (UC Berkeley), Marcus Thint (BT), Zhiheng Huang (Yahoo!), Ines Gonzalez Rodriguez (University of Cantabria), Xizhao Wang (Hebei University), Baoding Liu (Tsinghua University) and Nam Van Huynh (JAIST). Weifeng Zhang, my student at Beihang University, helped to develop the algorithm of data and imprecise clustering. The first author would also like to thank Prof. Katia Sycara for hosting him at Robotics Institute, Carnegie Mellon University. This visit gave him more time to focus on this book and think more deeply about the relations between linguistic labels and natural language.

This work has depended on the generosity of free software LATEX and numerous contributors of Wikipedia. Zhejiang University Press and Springer have provided excellent support throughout all the stages of preparation of this book. We thank Jiaying Xu, our editor, for her patience and support to provide help when we are behind the schedule.

This book is funded by Beihang Series in Space Technology and Applications. The research presented in this book is funded by the National Basic Research Program of China (973 Program) under Grant No. 2012CB316400, and National Natural Science Foundation of China (NSFC) (Nos. 61075046 and 60604034), the joint funding of NSFC and MSRA (No. 60776798), the Natural Science Foundation of Zhejiang Province (No. Y1090003), and the New Century Excellent Talents (NCET) program from the Ministry of Education, China. Finally, we would like to thank our families for being hugely supportive in our work.

Contents

1	Introduction	1
1.1	Types of Uncertainty	1
1.2	Uncertainty Modeling and Data Mining	4
1.3	Related Works	6
	References	9
2	Induction and Learning	13
2.1	Introduction	13
2.2	Machine Learning	14
2.2.1	Searching in Hypothesis Space	16
2.2.2	Supervised Learning	18
2.2.3	Unsupervised Learning	20
2.2.4	Instance-Based Learning	22
2.3	Data Mining and Algorithms	23
2.3.1	Why Do We Need Data Mining?	24
2.3.2	How Do We do Data Mining?	24
2.3.3	Artificial Neural Networks	25
2.3.4	Support Vector Machines	27
2.4	Measurement of Classifiers	29
2.4.1	ROC Analysis for Classification	30
2.4.2	Area Under the ROC Curve	31
2.5	Summary	34
	References	34
3	Label Semantics Theory	39
3.1	Uncertainty Modeling with Labels	39
3.1.1	Fuzzy Logic	39
3.1.2	Computing with Words	41
3.1.3	Mass Assignment Theory	42
3.2	Label Semantics	44
3.2.1	Epistemic View of Label Semantics	45

3.2.2	Random Set Framework	46
3.2.3	Appropriateness Degrees	50
3.2.4	Assumptions for Data Analysis	51
3.2.5	Linguistic Translation	54
3.3	Fuzzy Discretization	57
3.3.1	Percentile-Based Discretization	58
3.3.2	Entropy-Based Discretization	58
3.4	Reasoning with Fuzzy Labels	61
3.4.1	Conditional Distribution Given Mass Assignments	61
3.4.2	Logical Expressions of Fuzzy Labels	62
3.4.3	Linguistic Interpretation of Appropriate Labels	65
3.4.4	Evidence Theory and Mass Assignment	66
3.5	Label Relations	69
3.6	Summary	73
	References	74
4	Linguistic Decision Trees for Classification	77
4.1	Introduction	77
4.2	Tree Induction	77
4.2.1	Entropy	79
4.2.2	Soft Decision Trees	82
4.3	Linguistic Decision for Classification	82
4.3.1	Branch Probability	85
4.3.2	Classification by LDT	88
4.3.3	Linguistic ID3 Algorithm	90
4.4	Experimental Studies	92
4.4.1	Influence of the Threshold	93
4.4.2	Overlapping Between Fuzzy Labels	95
4.5	Comparison Studies	98
4.6	Merging of Branches	102
4.6.1	Forward Merging Algorithm	103
4.6.2	Dual-Branch LDTs	105
4.6.3	Experimental Studies for Forward Merging	105
4.6.4	ROC Analysis for Forward Merging	109
4.7	Linguistic Reasoning	111
4.7.1	Linguistic Interpretation of an LDT	111
4.7.2	Linguistic Constraints	113
4.7.3	Classification of Fuzzy Data	115
4.8	Summary	117
	References	118

5	Linguistic Decision Trees for Prediction	121
5.1	Prediction Trees	121
5.2	Linguistic Prediction Trees	122
5.2.1	Branch Evaluation	123
5.2.2	Defuzzification	126
5.2.3	Linguistic ID3 Algorithm for Prediction	128
5.2.4	Forward Branch Merging for Prediction	128
5.3	Experimental Studies	130
5.3.1	3D Surface Regression	131
5.3.2	Abalone and Boston Housing Problem	134
5.3.3	Prediction of Sunspots	135
5.3.4	Flood Forecasting	137
5.4	Query Evaluation	143
5.4.1	Single Queries	143
5.4.2	Compound Queries	144
5.5	ROC Analysis for Prediction	145
5.5.1	Predictors and Probabilistic Classifiers	145
5.5.2	AUC Value for Prediction	149
5.6	Summary	152
	References	152
6	Bayesian Methods Based on Label Semantics	155
6.1	Introduction	155
6.2	Naive Bayes	156
6.2.1	Bayes Theorem	157
6.2.2	Fuzzy Naive Bayes	158
6.3	Fuzzy Semi-Naive Bayes	159
6.4	Online Fuzzy Bayesian Prediction	161
6.4.1	Bayesian Methods	161
6.4.2	Online Learning	164
6.5	Bayesian Estimation Trees	165
6.5.1	Bayesian Estimation Given an LDT	165
6.5.2	Bayesian Estimation from a Set of Trees	167
6.6	Experimental Studies	168
6.7	Summary	169
	References	171
7	Unsupervised Learning with Label Semantics	177
7.1	Introduction	177
7.2	Non-Parametric Density Estimation	178
7.3	Clustering	180
7.3.1	Logical Distance	181
7.3.2	Clustering of Mixed Objects	185
7.4	Experimental Studies	187
7.4.1	Logical Distance Example	187

7.4.2	Images and Labels Clustering	190
7.5	Summary	191
	References	192
8	Linguistic FOIL and Multiple Attribute Hierarchy for Decision Making	193
8.1	Introduction	193
8.2	Rule Induction	193
8.3	Multi-Dimensional Label Semantics	196
8.4	Linguistic FOIL	199
8.4.1	Information Heuristics for LFOIL	199
8.4.2	Linguistic Rule Generation	200
8.4.3	Class Probabilities Given a Rule Base	202
8.5	Experimental Studies	203
8.6	Multiple Attribute Decision Making	206
8.6.1	Linguistic Attribute Hierarchies	206
8.6.2	Information Propagation Using LDT	209
8.7	Summary	213
	References	213
9	A Prototype Theory Interpretation of Label Semantics	215
9.1	Introduction	215
9.2	Prototype Semantics for Vague Concepts	217
9.2.1	Uncertainty Measures about the Similarity Neighborhoods Determined by Vague Concepts	217
9.2.2	Relating Prototype Theory and Label Semantics	220
9.2.3	Gaussian-Type Density Function	223
9.3	Vague Information Coarsening in Theory of Prototypes	227
9.4	Linguistic Inference Systems	229
9.5	Summary	231
	References	232
10	Prototype Theory for Learning	235
10.1	Introduction	235
10.1.1	General Rule Induction Process	235
10.1.2	A Clustering Based Rule Coarsening	236
10.2	Linguistic Modeling of Time Series Predictions	238
10.2.1	Mackey-Glass Time Series Prediction	239
10.2.2	Prediction of Sunspots	244
10.3	Summary	250
	References	252

11	Prototype-Based Rule Systems	253
11.1	Introduction	253
11.2	Prototype-Based IF-THEN Rules	254
11.3	Rule Induction Based on Data Clustering and Least-Square Regression	257
11.4	Rule Learning Using a Conjugate Gradient Algorithm	260
11.5	Applications in Prediction Problems	262
11.5.1	Surface Predication	262
11.5.2	Mackey-Glass Time Series Prediction	265
11.5.3	Prediction of Sunspots	269
11.6	Summary	274
	References	274
12	Information Cells and Information Cell Mixture Models	277
12.1	Introduction	277
12.2	Information Cell for Cognitive Representation of Vague Concept Semantics	277
12.3	Information Cell Mixture Model (ICMM) for Semantic Representation of Complex Concept	280
12.4	Learning Information Cell Mixture Model from Data Set	281
12.4.1	Objective Function Based on Positive Density Function	282
12.4.2	Updating Probability Distribution of Information Cells	282
12.4.3	Updating Density Functions of Information Cells	283
12.4.4	Information Cell Updating Algorithm	284
12.4.5	Learning Component Number of ICMM	285
12.5	Experimental Study	286
12.6	Summary	290
	References	290

Acronyms

AI Artificial Intelligence
ANN Artificial Neural Networks
AUC Area Under the ROC Curve
AVE Average Error
BLDT Bayesian LDT
BP Back Propagation
CAD Computer Aided Diagnosis
CW Computing with Words
D-S Dempster-Shafer
DT Decision Tree
EM Expectation-Maximization
FDT Fuzzy Decision Tree
FLDT Forest of LDTs
FOIL First-Order Inductive Learning
FPR False Positive Rate
FRBS Fuzzy Rule-Based Systems
FRIL Fuzzy Relational Inference Language
FSNB Fuzzy Semi-Naive Bayes
GTU General Theory of Uncertainty
IBL Instance-Based Learning
ICMM Information Cell Mixture Model
ID3 Iterative Dichotomiser 3
IG Information Gain
ILP Inductive Logical Programming
KDD Knowledge Discovery in Database
 k -NN k -Nearest Neighbors
LD Linguistic Data
LDT Linguistic Decision Tree
LFOIL Linguistic FOIL
LID3 Linguistic ID3

XVIII Acronyms

LLE	Locally Linear Embedding
LLR	Locally Linear Reconstruction
LPT	Linguistic Prediction Tree
LS	Least Square
LT	Linguistic Translation
MB	Merged Branch
MLP	Multi-Layer Perceptrons
MSE	Mean Square Error
MW	Modeling with Words
NB	Naive Bayes
NN	Neural Networks
PDF	Probability Density Function
PET	Probability Estimation Tree
PNL	Precisiated Natural Language
QP	Quadratic Programming
ROC	Receiver Operating Characteristics
SNB	Semi-Naive Bayes
SRM	Structural Risk Minimization
SVM	Support Vector Machines
SVR	Support Vector Regression
TPR	True Positive Rate

Notations

$ A $	Absolute value of A when A is a number or cardinality of A when A is a set
DB	Database with the size of $ DB $: $DB = \{\mathbf{x}_1, \dots, \mathbf{x}_{ DB }\}$
$\mathbf{x}_i$	n -dimensional variable that: $\mathbf{x}_i \in DB$ for $i = 1, \dots, DB $
$\mathbb{L}_x$	Set of labels defined on random variable x
LE	Logical expressions set given $\mathbb{L}$
$\mathbb{F}_x$	Focal set of random variable x
T	Linguistic decision tree that contains $ T $ branches: $T = \{B_1, \dots, B_{ T }\}$
$\mathbb{B}$	A set of branches: $\mathbb{B} = \{B_1, \dots, B_M\}$ $T \equiv \mathbb{B}$ iff: $M = T $
B	A branch of LDT, it has $ B $ focal elements: $B = \{F_1, \dots, F_{ B }\}$
$\mathbb{C}$	A set of classes: $\mathbb{C} = \{C_1, \dots, C_{ \mathbb{C} }\}$
m_x	Mass assignment of x
$m_{\mathbf{x}}$	Mass assignment on a multi-dimensional variable $\mathbf{x}$
$\mu_L(x)$	Appropriateness degree of using label L to describe x
$\mu_{\theta}(x)$	Appropriateness measure of using logical expression θ to describe x where $\theta \in LE$
$p(x y)$	Conditional probability of x given y
$Bel(\cdot)$	Belief function
$Pl(\cdot)$	Plausibility function
$\lambda(\theta)$	λ -function to transfer the logical expression θ into a set of labels
$\mu_{\theta}x$	Appropriateness measure of using logical expression θ to label x
$IG(\cdot)$	Information Gain function
FD	Fuzzy database $FD = \{\langle \theta_1(i), \dots, \theta_n(i) \rangle : i = 1, \dots, N\}$
$\hat{x}$	Estimated value of x based on a training database
$\tilde{p}$	Updated value of p at iterative updating process
$P(x m)$	Conditional distribution of x given mass assignment m
$pm(\cdot)$	Prior mass assignment
$\mathcal{L} \mathcal{P}$	Information cell mixture model $\mathcal{L} \mathcal{P} = \langle \mathbb{L}, Pr \rangle$