ADVANCED TOPICS IN SCIENCE AND TECHNOLOGY IN CHINA

Zengchang Qin
Yongchuan Tang

Uncertainty Modeling for Data Mining

A Label Semantics Approach

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

Springer

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

图书在版编目 (CIP) 数据

基于不确定性建模的数据挖掘= Uncertainty modeling for data mining 英文 / 秦曾昌，汤永川著．—杭州：浙江大学出版社，2013.11
ISBN 978-7-308-12106-4

I. ①基… II. ①秦… ②汤… III. ①数据采集—研究—英文 IV. ①TP274

中国版本图书馆 CIP 数据核字(2013)第 195373 号

基于不确定性建模的数据挖掘
秦曾昌　汤永川　著

责任编辑　许佳颖
封面设计　俞亚彤
出版发行　浙江大学出版社
网址：http://www.zjupress.com
Springer-Verlag GmbH
网址：http://www.springer.com
排　　版　杭州理想广告有限公司
印　　刷　浙江印刷集团有限公司
开　　本　710mm×1000mm　1/16
印　　张　19.5
字　　数　602 千
版 印 次　2013 年 11 月第 1 版　2013 年 11 月第 1 次印刷
书　　号　ISBN 978-7-308-12106-4 (浙江大学出版社)
ISBN 978-3-642-41250-9 (Springer-Verlag GmbH)
定　　价　120.00 元

浙江大学出版社发行部联系方式:0571-88925591；http://zjdxcbs.tmall.com

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.

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 basesd 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 Nikravesh (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.

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

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 $\|DB\|$: $DB = \{\mathbf{x}_1, \ldots, \mathbf{x}_{\|DB\|}\}$
$\mathbf{x}_i$	n-dimensional variable that: $\mathbf{x}_i \in DB$ for $i = 1, \ldots, \|DB\|$
$\mathbb{L}_x$	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: $T = \{B_1, \ldots, B_{\|T\|}\}$
$\mathbb{B}$	A set of branches: $\mathbb{B} = \{B_1, \ldots, B_M\}$ $T \equiv \mathbb{B}$ iff: $M = \|T\|$
B	A branch of LDT, it has $\|B\|$ focal elements: $B = \{F_1, \ldots, F_{\|B\|}\}$
$\mathbb{C}$	A set of classes: $\mathbb{C} = \{C_1, \ldots, C_{\|\mathbb{C}\|}\}$
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 $FD = \{\langle \theta_1(i), \ldots, \theta_n(i) \rangle : i = 1, \ldots, N\}$
$\hat{x}$	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
$\mathscr{L}\mathscr{P}$	Information cell mixture model $\mathscr{L}\mathscr{P} = \langle \mathbb{L}, Pr \rangle$

Contents