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

Uncertainty Modeling for Data Mining

A Label Semantics Approach

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

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.

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

1 Introduction

So far as the laws of mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality.

—Albert Einstein (1879–1955), "Geometry and Experience"

1.1 Types of Uncertainty

Our nature is uncertain. Given this fact, there are two main streams of philosophy to understand uncertainty. First, the nature is incomplete and is full of uncertainties. Uncertainty is an objective and undeniable fact of nature. The second stream implies that the nature is governed by orders and laws. However, we cannot perceive all these laws from our limited cognitive abilities. That is where the uncertainties come from. The existence of uncertainty is because of the lack of information. Following these two streams of philosophy, uncertainty can be roughly classified into the following two categories:

(1) Epistemic or systematic uncertainties are due to things we could in principle know but don't in practice. This may be either because we have not measured a quantity sufficiently accurately, or because our model neglects certain effects. The uncertainty comes from an imprecise nature which is involved with mixture of truths. As gray is a mixture of black and white.
(2) Aleatoric or statistical uncertainties are unknowns that differ each time we would make the same experiment. We assume there exists an ideological and undeniable fact which is the reason for a phenomenon. However, it cannot be perceived due to the limitation of human cognitive abilities. Each experiment is actually the observable evidence of this "fact" from which we can know better about this fact by conducting repeated experiments.

Vagueness or ambiguity is sometimes described as "second order uncertainty", where there is uncertainty even about the definitions of uncertain states or outcomes. To quote Lindley [1]:

> There are some things that you know to be true, and others that you know to be false; yet, despite this extensive knowledge that you have, there remain many things whose truth or falsity is not known to you. We say that you are uncertain about them. You are uncertain, to varying degrees, about everything in the future; much of the past is hidden from you; and there is a lot of the present about which you do not have full information. Uncertainty is everywhere and you cannot escape from it.

Philosophically, uncertainty is ubiquitous. However, in the practice of science and engineering, what we are concerned with is how to predict future events by using uncertain information with a proper measure. Probability is a way of expressing knowledge or belief that an event will occur or has occurred using uncertainty information. Generally, there are two broad categories of probability interpretations: frequentists and Bayesians. Frequentists consider probability to be the relative frequency of occurrence from repeating games. Bayesians use probability as a measure of an individual's degree of belief. Such belief can be updated by new observable evidence from a prior [2]. In the last few decades, Bayesian probability has been widely used in probabilistic reasoning and statistical inference [3,4]. Many successful algorithms have been proposed and applications have been used in real-world practice. Bayesian probability theory assumes that uncertainty exists because of the limitation of our cognitive abilities and lack of information①. Some other uncertainty theories have been proposed to assume that the nature itself is uncertain and independent from the limited abilities of acquiring this information. Among them, Fuzzy Logic is the most successful and widely-used theory of modeling such a type of uncertainty.

Proposed by Zadeh in 1965 [5], fuzzy logic is a superset of conventional Boolean logic that has been extended to handle the concept of partial truth (an interpretation of the uncertainty of being true) — truth values between "completely true" and "completely false". Three hundred years B.C., the Greek philosopher, Aristotle, came up with binary logic of true and false, which is now the principle foundation of mathematics. Two centuries before Aristotle, Buddha, had the belief which contradicted the black-and-white world, which went beyond the bivalent cocoon and sees the world as it is, filled with contradictions. Such beliefs are popular especially in oriental cultures, such as the Chinese Yin-Yang concept which is used to describe how polar or seemingly contrary forces are interconnected and interdependent in the natural world, and how they give rise to each other in turn [6].

Both fuzzy logic and probability theory can be used to represent subjective belief. Fuzzy set theory uses the concept of fuzzy set membership (i.e., how much a variable is in a set), and probability theory (Bayesian) uses the concept of subjective probability (i.e., how probable do I think that a variable is in a set). While this distinction is mostly philosophical, there is no such situation where this variable is partially in the set; the variable is either in the set or not, absolutely. However, we do not have such absolute belief because of the lack of information. The

① According to Jaynes, probability is an extension of logic given incomplete information [2].

fuzzy-logic-derived possibility measure is inherently different from the probability measure; hence, they are not directly equivalent [7]. The work presented in this book actually uses both fuzzy logic and probability for modeling uncertainty and making predictions based on observable evidence. The nature of uncertainty is modeled by fuzzy labels and the reasoning for using evidence is probabilistic.

A prediction or forecast is a statement about the way things will happen in the future. A basic difference between a good predictor and a random guesser is that a good predictor always uses the previous experience or embedded knowledge when making predictions. We human beings are using such a way for making wise decisions or predictions. The research of studying how to effectively use machines to make predictions using given historic data is referred to as machine learning [8]. In this information age, we are buried by a tremendous amount of data. How we use machine learning algorithms to exploit the data for discovering useful patterns is called *data mining*.

Machine learning and data mining research has developed rapidly in recent decades. As one of the most successful branches of artificial intelligence (AI), it has had a tremendous impact on the current world ②. Many new technologies have emerged or been reborn with its development such as bioinformatics [9], natural language processing [10], computer vision [11], information theory [12], and information retrieval [13]. Traditionally machine learning and data mining research has focused on learning algorithms with high classification or prediction accuracy. From another perspective, however, this is not always sufficient for some real world applications that require good algorithm transparency. By the latter we refers to the interpretability of models; that is, the models need to be easily understood and provide information regarding underlying trends and relationships that can be used by practitioners in the relevant fields. Transparent models should allow for a qualitative understanding of the underlying system in addition to giving quantitative predictions of behavior. The intuition behind this idea is the way of human reasoning with imprecise concepts. It has been a well-accepted fact that computers have beaten the human being in numerical calculations in both accuracy and speed. However, the capability of imprecise reasoning is still Achilles' heel for machines.

Uncertainty and imprecision are often inherent in modeling these real-world applications and it is desirable that these should be incorporated into learning algorithms. In this book, we shall investigate the effectiveness of a high-level modeling framework from the dual perspectives of accuracy and interpretability. The reasoning is that by enabling models to be defined in terms of linguistic expressions we can enhance robustness, accuracy and transparency. We need a higher level modeling language which is to be truly effective and it must

② In 2011, IBM's *Watson*, an artificial intelligence computer system capable of answering questions posed in natural language, beat other human competitors on a famous American quiz show *Jeopardy* and became the biggest winner. Its core algorithm, *DeepQA*, basically uses advanced machine learning and information retrieval technologies. This is a big event for attracting people's attention to the long lasting human-machine competition since the last breakthrough by *Deep Blue*, the world champion chess player, also from IBM.

provide a natural knowledge representation framework for inductive learning. As such it is important that it allows for the modeling of uncertainty, imprecision and vagueness in a semantically clear manner. Here we present such a higher level knowledge representation framework centered on the Modeling with Words (MW)[14] paradigm.

We need to notice that the underlying semantics of our approach is quite different from computing with words (CW) ③ proposed by Zadeh[15]. In this book, the framework is used mainly for modeling and building intelligent data mining systems. In such systems, we use words or fuzzy labels for modeling uncertainties and use probabilistic approaches for reasoning. Therefore, the framework we will introduce is an achievement of the research of modeling with words (MW) rather than CW. The new framework we shall use in this book, label semantics[16], is a random set based semantics for modeling imprecise concepts where the degree of appropriateness of a linguistic expression, as a description of a value, is measured in terms of how the set of appropriate labels for that value varies across a population. Different from traditional fuzzy logic, fuzzy memberships are viewed as being fixed point coverage functions of random sets, themselves representing uncertainty or variations in underlying crisp definition of an imprecise concept. Also, label semantics allows linguistic queries and information fusion in a logical representation of linguistic expressions. Therefore, label semantics provides us with an ideal framework for modeling uncertainty with good transparency.

1.2 Uncertainty Modeling and Data Mining

Since the invention of fuzzy logic, it has been widely applied in engineering especially in control problems by handling the uncertainty information as a set of expert rules. However, in this information age, we are facing some new challenges. Nowadays, a tremendous amount of data and information has flooded us. Contributing factors including the widespread use of the World Wide Web (WWW) and other digital innovations in electronics and computing, such as digital cameras, intelligent mobile phones, PDAs and new portal computing devices such as iPad, Blackberry, Kindle, and etc. Most importantly, all the classical communication tools such as papers, books, photos, videos are digitalized and have never been so easily accessed as today. We are in the age of overwhelming information. The ability to find the useful information has never been so important in history. Valuable information may be hiding behind the data, but it is difficult for human beings to extract this without powerful tools. We have already been living in a "data rich but information poor" environment since the invention of these innovative IT infrastructures and devices. To relieve such a plight, data mining research emerged and has developed rapidly in the past few decades.

③ CW is focused on developing a calculus of using linguistic terms directly for reasoning based on a fuzzy logic framework. More details on modeling with words are available in Reference [14], in which Zadeh pointed out the differences between CW and MW in the foreword of this book.

Data mining has become one of the most active and exciting areas for its omnipresent applicability in the current world. Approaches to data mining research mainly include three perspectives according to Zhou[17]: databases, machine learning, and statistics. Especially from the perspective of machine learning, many data mining algorithms have been developed to accomplish a limited set of tasks and produce a particular enumeration patterns over data sets. But more theoretical and practical problems still block our way to gain knowledge from data. Among these obstacles, uncertainty is one of the most intractable. The traditional data mining algorithms, such as decision trees[18,19] and K-means clustering[20], are crisp and each database value may be classified into at most one cluster. This is unlikely to satisfy everyday life experiences where a value may be partially classified into one or more categories.

Probabilistic approaches for data mining have been the main stream of this research for handling the statistical uncertainties. We generally assume some prior probabilities in the hypothesis space, by inference on observations, to yield the best hypothesis that can explain the observations best. Form another perspective, systemic uncertainties are not well handled in such a probabilistic reasoning framework. Imprecise data, missing data, and human subjectivity, all could cause such uncertainty. Fuzzy logic is a good means for handling these uncertainties, and also provides an inference methodology to enable the principles of approximate human reasoning capabilities to be systematically used as a basis for knowledge-based systems. In contrast to a classic set, the boundary of a fuzzy set is blurred. This smooth transition is characterized by membership functions which give a fuzzy set flexibility in modeling linguistic expressions. The appearance of fuzzy logic becomes an important milestone in not only mathematics and logic but also scientific philosophy — it is complementary to our classical 0-or-1, black-or-white view of the nature[21]. Interpretations of membership degrees include similarity, preference, and uncertainty[22]: they can state how similar an object or case is to a prototypical one, they can indicate preferences between suboptimal solutions to a problem, or they can model uncertainty about the true situation that is described in imprecise terms. Generally, due to their closeness to human reasoning, solutions obtained using fuzzy approaches are easy to understand and apply.

Uncertainty may exist in data mining models in various different ways:

(1) The model structure, i.e., how accurately a mathematical model describes the true system for a real-life situation, may be known only approximately. Models are almost always only approximations to reality.
(2) The numerical approximation, i.e., how appropriately a numerical method is used in approximating the operation of the system. Most models are too complicated to solve exactly. For example, the finite element method may be used to approximate the solution of a partial differential equation, but this introduces an error (the difference between the exact and the numerical solutions).
(3) Input and/or model parameters may be known only approximately due to the noise of data.