



***Also in this series***

Gregoris Mentzas, Dimitris Apostolou, Andreas Abecker  
and Ron Young

*Knowledge Asset Management*

1-85233-583-1

Michalis Vazirgiannis, Maria Halkidi  
and Dimitrios Gunopulos

---

# Uncertainty Handling and Quality Assessment in Data Mining

With 35 Figures



Springer

Michalis Vazirgiannis, PhD  
Department of Informatics, Athens University of Economics and Business  
Maria Halkidi, MSc  
Department of Informatics, Athens University of Economics and Business  
Dimitrios Gunopulos, PhD  
Department of Computer Science and Engineering, University of California,  
Riverside  
*Series Editors*  
Xindong Wu  
Lakhmi Jain

British Library Cataloguing in Publication Data  
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data  
Vazirgiannis, Michalis, 1964-

Uncertainty handling and quality assessment in data mining / Michalis Vazirgiannis,  
Maria Halkidi, and Dimitrios Gunopulos.

p. cm. – (Advanced information and knowledge processing, ISSN 1610-3947)

Includes bibliographical references and index.

ISBN 978-1-4471-1119-1 ISBN 978-1-4471-0031-7 (eBook)

DOI 10.1007/978-1-4471-0031-7

1. Data mining. 2. Data mining- -Quality control. I. Halkidi, Maria 1974- II.

Gunopulos, Dimitrios, 1967- III. Title. IV. Series

QA76.9.D343V39 2003

006.3- -dc21

2003042421

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

AI&KP ISSN 1610-3947

ISBN 978-1-4471-1119-1

<http://www.springer.co.uk>

© Springer-Verlag London 2003

Originally published by Springer-Verlag London Limited in 2003

Softcover reprint of the hardcover 1st edition 2003

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

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Typesetting: Electronic text files prepared by authors

34/3830-543210 Printed on acid-free paper SPIN 10879524

# CONTENTS

---

<b>INTRODUCTION .....</b>	<b>1</b>
<b>DATA MINING PROCESS.....</b>	<b>11</b>
2.1 Introduction to the Main Concepts of Data Mining .....	11
2.2 Knowledge and Data Mining .....	12
2.2.1 Knowledge Discovery in Database vs Data Mining .....	12
2.3 The Data Mining Process .....	16
2.3.1 Data Mining Requirements .....	17
2.4 Classification of Data Mining Methods .....	18
2.5 Overview of Data Mining Tasks .....	19
2.5.1 Clustering.....	20
2.5.1.1 Overview of Clustering Algorithms.....	25
2.5.1.2 Comparison of Clustering Algorithms .....	31
2.5.2 Classification .....	43
2.5.2.1 Bayesian Classification .....	44
2.5.2.2 Decision Trees.....	46
2.5.2.3 Neural Networks .....	49
2.5.2.4 Nearest Neighbor Classification .....	50
2.5.2.5 Support Vector Machines (SVMs).....	50
2.5.2.6 Fuzzy Classification approaches .....	51
2.5.3 Induction of classification rules .....	53
2.5.4 Association Rules .....	55
2.5.5 Sequential Patterns.....	56
2.5.6 Time Series Similarity .....	58

2.5.7	Visualization and Dimensionality Reduction .....	60
2.5.8	Regression .....	61
2.5.9	Summarization .....	61
2.6	Summary .....	61
References	.....	63
<b>QUALITY ASSESSMENT IN DATA MINING .....</b>		<b>73</b>
3.1	Introduction .....	73
3.2	Data Pre-processing and Quality Assessment .....	75
3.3	Evaluation of Classification Methods .....	76
3.3.1	Classification Model Accuracy .....	76
3.3.1.1	Alternatives to the Accuracy Measure .....	77
3.3.2	Evaluating the Accuracy of Classification Algorithms .....	78
3.3.2.1	McNemar's Test .....	79
3.3.2.2	A Test for the Difference of Two Proportions .....	80
3.3.2.3	The Resampled Paired $t$ Test .....	81
3.3.2.4	The $k$ -fold Cross-validated Paired $t$ Test .....	82
3.3.3	Interestingness Measures of Classification Rules .....	82
3.3.3.1	Rule-Interest Function .....	82
3.3.3.2	Smyth and Goodman's J-Measure .....	83
3.3.3.3	General Impressions .....	83
3.3.3.4	Gago and Bento's Distance Metric .....	84
3.4	Association Rules .....	85
3.4.1	Association Rules Interestingness Measures .....	85
3.4.1.1	Coverage .....	86
3.4.1.2	Support .....	86
3.4.1.3	Confidence .....	86
3.4.1.4	Leverage .....	87
3.4.1.5	Lift .....	88
3.4.1.6	Rule Templates .....	89
3.4.1.7	Gray and Orłowska's Interestingness .....	90
3.4.1.8	Dong and Li's Interestingness .....	90
3.4.1.9	Peculiarity .....	91
3.4.1.10	Closed Association Rules Mining .....	92
3.5	Cluster Validity .....	93
3.5.1	Fundamental Concepts of Cluster Validity .....	95

3.5.2	External and Internal Validity Indices .....	96
3.5.2.1	Hypothesis Testing in Cluster Validity .....	96
3.5.2.2	External Criteria .....	98
3.5.2.3	Internal Criteria .....	101
3.5.3	Relative Criteria .....	102
3.5.3.1	Crisp Clustering .....	104
3.5.3.2	Fuzzy Clustering .....	115
3.5.4	Other Approaches for Cluster Validity .....	118
3.5.5	An Experimental Study on cluster validity .....	120
3.5.5.1	A Comparative Study .....	120
3.6	Summary .....	121
References	.....	123
<b>UNCERTAINTY HANDLING IN DATA MINING .....</b>		<b>129</b>
4.1	Introduction .....	129
4.2	Basic Concepts on Fuzzy Logic .....	131
4.2.1	Fuzzy Set Theory .....	132
4.2.2	Membership Functions .....	133
4.2.2.1	Hypertrapezoidal Fuzzy Membership Functions .....	134
4.2.2.2	Joint Degree of Membership .....	136
4.2.3	Fuzzy Sets and Information Measures .....	137
4.3	Basic Concepts on Probabilistic Theory .....	138
4.3.1	Uncertainty Quantified Probabilistically .....	139
4.3.1.1	Bayesian Theorem .....	139
4.4	Probabilistic and Fuzzy Approaches .....	140
4.5	The EM Algorithm .....	141
4.5.1	General Description of EM Algorithm .....	141
4.6	Fuzzy Cluster Analysis .....	143
4.6.1	Fuzzy C-Means and its Variants .....	143
4.6.2	Fuzzy C-Means for Object-Data .....	145
4.6.3	Fuzzy C-Means (FCM) Alternatives .....	146
4.6.4	Applying Fuzzy C-Means Methodology to Relational Data ....	148
4.6.5	The Fuzzy C-Means Algorithm for Relational data .....	148
4.6.5.1	Comments on FCM for Relational Data .....	150
4.6.6	Noise Fuzzy Clustering Algorithm .....	151

4.6.7	Conditional Fuzzy C-Means Clustering .....	152
4.7	Fuzzy Classification Approaches .....	154
4.7.1	Fuzzy Decision Trees .....	154
4.7.1.1	Building a Fuzzy Decision Tree.....	154
4.7.1.2	Inference for Decision Assignment.....	155
4.7.2	Fuzzy Rules .....	159
4.8	Managing Uncertainty and Quality in the Classification Process .....	160
4.8.1	Framework Description .....	161
4.8.2	Mapping to the Fuzzy Domain .....	163
4.8.2.1	Classification Space (CS).....	163
4.8.2.2	Classification Value Space (CVS) .....	165
4.8.3	Information Measures for Decision Support .....	166
4.8.3.1	Class Energy Metric .....	167
4.8.3.2	Attribute Energy Metric .....	168
4.8.4	Queries & Decision Support.....	168
4.8.5	Classification Scheme Quality Assessment .....	170
4.9	Fuzzy Association Rules .....	171
4.9.1	Defining Fuzzy Sets.....	172
4.9.2	Fuzzy Association Rule Definition.....	173
4.9.2.1	Fuzzy Support .....	174
4.9.2.2	Fuzzy Confidence .....	174
4.9.2.3	Fuzzy Correlation.....	174
4.9.3	Mining Fuzzy Association Rules Algorithms.....	175
4.10	Summary .....	177
	References.....	178
	<b>UMINER: A DATA MINING SYSTEM HANDLING UNCERTAINTY AND QUALITY .....</b>	<b>183</b>
5.1	Introduction.....	183
5.2	UMiner Development Approach.....	184
5.3	System Architecture .....	186
5.4	UMiner's Data Mining Tasks.....	187
5.5	Demonstration.....	191



Contents	IX
5.5.1    Clustering process.....	191
5.6 Summary .....	195
References.....	197
<b>CASE STUDIES.....</b>	<b>199</b>
6.1 Extracting Association Rules for Medical Data Analysis .....	199
6.2 The Mining Process .....	200
6.2.1    Collection of Data.....	200
6.2.2    Data Cleaning and Pre-processing.....	200
6.2.3    Further Analysis of Extracted Association Rules .....	201
6.3 Cluster Analysis of Epidemiological Data .....	215
References.....	221
<b>INDEX .....</b>	<b>223</b>