

KNOWLEDGE-BASED CLUSTERING

From Data to Information Granules

Witold Pedrycz

*Department of Electrical and Computer Engineering
University of Alberta
Edmonton, Canada*

and

*Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland*



A JOHN WILEY & SONS, INC., PUBLICATION

KNOWLEDGE-BASED CLUSTERING

KNOWLEDGE-BASED CLUSTERING

From Data to Information Granules

Witold Pedrycz

*Department of Electrical and Computer Engineering
University of Alberta
Edmonton, Canada*

and

*Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland*



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2005 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Pedrycz, Witold, 1953-

Knowledge-based clustering : from data to information granules / Witold Pedrycz.
p. cm.

"A Wiley-Interscience publication."

Includes bibliographical references and index.

ISBN 0-471-46966-1 (cloth)

1. Soft computing. 2. Granular computing. 3. Fuzzy systems. I. Title.

QA76.9.S63P45 2005

006.3-dc22

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Ewa, Adam, and Barbara

Contents

Foreword	xiii
Preface	xv
1 Clustering and Fuzzy Clustering	1
1.1 Introduction	1
1.2 Basic Notions and Notation	1
1.2.1 Types of Data	2
1.2.2 Distance and Similarity	2
1.3 Main Categories of Clustering Algorithms	6
1.3.1 Hierarchical Clustering	6
1.3.2 Objective Function-Based Clustering	8
1.4 Clustering and Classification	10
1.5 Fuzzy Clustering	11
1.6 Cluster Validity	18
1.7 Extensions of Objective Function-Based Fuzzy Clustering	19
1.7.1 Augmented Geometry of Fuzzy Clusters: Fuzzy C Varieties	19
1.7.2 Possibilistic Clustering	20
1.7.3 Noise Clustering	22
1.8 Self-Organizing Maps and Fuzzy Objective Function-Based Clustering	23
1.9 Conclusions	25
References	26
2 Computing with Granular Information: Fuzzy Sets and Fuzzy Relations	28
2.1 A Paradigm of Granular Computing: Information Granules and Their Processing	28
2.2 Fuzzy Sets as Human-Centric Information Granules	31
2.3 Operations on Fuzzy Sets	32
2.4 Fuzzy Relations	33
2.5 Comparison of Two Fuzzy Sets	35
2.6 Generalizations of Fuzzy Sets	37
2.7 Shadowed Sets	38

2.8	Rough Sets	44
2.9	Granular Computing and Distributed Processing	46
2.10	Conclusions	47
	References	47
3	Logic-Oriented Neurocomputing	50
3.1	Introduction	50
3.2	Main Categories of Fuzzy Neurons	51
3.2.1	Aggregative Neurons	52
3.2.2	Referential (Reference) Neurons	55
3.3	Architectures of Logic Networks	59
3.4	Interpretation Aspects of the Networks	61
3.5	Granular Interfaces of Logic Processing	62
3.6	Conclusions	64
	References	64
4	Conditional Fuzzy Clustering	66
4.1	Introduction	66
4.2	Problem Statement: Context Fuzzy Sets and Objective Function	68
4.3	The Optimization Problem	70
4.4	Computational Considerations of Conditional Clustering	80
4.5	Generalizations of the Algorithm Through the Aggregation Operator	81
4.6	Fuzzy Clustering with Spatial Constraints	82
4.7	Conclusions	86
	References	86
5	Clustering with Partial Supervision	87
5.1	Introduction	87
5.2	Problem Formulation	88
5.3	Design of the Clusters	90
5.4	Experimental Examples	91
5.5	Cluster-Based Tracking Problem	93
5.6	Conclusions	96
	References	96
6	Principles of Knowledge-Based Guidance in Fuzzy Clustering	97
6.1	Introduction	97
6.2	Examples of Knowledge-Oriented Hints and Their General Taxonomy	99
6.3	The Optimization Environment of Knowledge-Enhanced Clustering	102

6.4	Quantification of Knowledge-Based Guidance Hints and Their Optimization	105
6.5	Organization of the Interaction Process	107
6.6	Proximity-Based Clustering (P-FCM)	112
6.7	Web Exploration and P-FCM	117
6.8	Linguistic Augmentation of Knowledge-Based Hints	126
6.9	Conclusions	127
	References	127
7	Collaborative Clustering	129
7.1	Introduction and Rationale	129
7.2	Horizontal and Vertical Clustering	131
7.3	Horizontal Collaborative Clustering	132
7.3.1	Optimization Details	135
7.3.2	The Flow of Computing of Collaborative Clustering	137
7.3.3	Quantification of the Collaborative Phenomenon of Clustering	138
7.4	Experimental Studies	140
7.5	Further Enhancements of Horizontal Clustering	150
7.6	The Algorithm of Vertical Clustering	151
7.7	A Grid Model of Horizontal and Vertical Clustering	153
7.8	Consensus Clustering	155
7.9	Conclusions	157
	References	157
8	Directional Clustering	158
8.1	Introduction	158
8.2	Problem Formulation	159
8.2.1	The Objective Function	160
8.2.2	The Logic Transformation Between Information Granules	161
8.3	The Algorithm	163
8.4	The Development Framework of Directional Clustering	166
8.5	Numerical Studies	167
8.6	Conclusions	174
	References	176
9	Fuzzy Relational Clustering	178
9.1	Introduction and Problem Statement	178
9.2	FCM for Relational Data	179
9.3	Decomposition of Fuzzy Relational Patterns	181

9.3.1	Gradient-Based Solution to the Decomposition Problem	182
9.3.2	Neural Network Model of the Decomposition Problem	184
9.4	Comparative Analysis	188
9.5	Conclusions	189
	References	189
10	Fuzzy Clustering of Heterogeneous Patterns	191
10.1	Introduction	191
10.2	Heterogeneous Data	192
10.3	Parametric Models of Granular Data	194
10.4	Parametric Mode of Heterogeneous Fuzzy Clustering	195
10.5	Nonparametric Heterogeneous Clustering	198
10.5.1	A Frame of Reference	198
10.5.2	Representation of Granular Data Through the Possibility-Necessity Transformation	200
10.5.3	Dereferencing	205
10.6	Conclusions	207
	References	208
11	Hyperbox Models of Granular Data: The Tchebyshev FCM	209
11.1	Introduction	209
11.2	Problem Formulation	210
11.3	The Clustering Algorithm—Detailed Considerations	211
11.4	Development of Granular Prototypes	218
11.5	Geometry of Information Granules	220
11.6	Granular Data Description: A General Model	223
11.7	Conclusions	223
	References	224
12	Genetic Tolerance Fuzzy Neural Networks	226
12.1	Introduction	226
12.2	Operations of Thresholding and Tolerance: Fuzzy Logic-Based Generalizations	227
12.3	Topology of the Logic Network	231
12.4	Genetic Optimization	235
12.5	Illustrative Numeric Studies	236
12.6	Conclusions	244
	References	245

13 Granular Prototyping	246
13.1 Introduction	246
13.2 Problem Formulation	247
13.2.1 Expressing Similarity Between Two Fuzzy Sets	247
13.2.2 Performance Index (Objective Function)	248
13.3 Prototype Optimization	251
13.4 Development of Granular Prototypes	263
13.4.1 Optimization of the Similarity Levels	263
13.4.2 An Inverse Similarity Problem	264
13.5 Conclusions	268
References	268
14 Granular Mappings	270
14.1 Introduction and Problem Statement	270
14.2 Possibility and Necessity Measures as the Computational Vehicles of Granular Representation	271
14.3 Building the Granular Mapping	272
14.4 Designing Multivariable Granular Mappings Through Fuzzy Clustering	275
14.5 Quantification of Granular Mappings	278
14.6 Experimental Studies	278
14.7 Conclusions	280
References	282
15 Linguistic Modeling	283
15.1 Introduction	283
15.2 Cluster-Based Representation of Input-Output Mapping	285
15.3 Conditional Clustering in the Development of a Blueprint of Granular Models	287
15.4 The Granular Neuron as a Generic Processing Element in Granular Networks	290
15.5 The Architecture of Linguistic Models Based on Conditional Fuzzy Clustering	293
15.6 Refinements of Linguistic Models	294
15.7 Conclusions	295
References	296
Bibliography	297
Index	315

Foreword

It is always a challenging task to write a foreword to a work authored by Professor Pedrycz. The reason is that, as a rule, what he writes about goes far beyond what can be found in the existing literature. This is particularly true in the instance of *Knowledge-Based Clustering: From Data to Information Granules* or *Knowledge-Based Clustering*, for short. *Knowledge-Based Clustering* is a magnum opus which touches upon some of the most basic facets of human cognition. It does so with authority, originality, erudition, insight, and high expository skill. Profusion of examples, figures, and references make Professor Pedrycz's work a pleasure to read.

In *Knowledge-Based Clustering*, Professor Pedrycz addresses a vast array of linked subjects. Starting with an exposition of clustering and fuzzy clusters, he moves to computing with granular information, with a granule being a clump of attribute-values drawn together by indistinguishability, equivalence, similarity, proximity, or functionality. Professor Pedrycz's co-authorship of a recent text on granular computing provides him with an effective framework for linking clustering with granular computing. In granular computing, the objects of computation are granules rather than singletons. In its general form, granular computing subsumes computing with intervals, computing with rough sets, and computing with probability distributions. The linkage between granular computing and cluster analysis plays a pivotal role throughout Professor Pedrycz's work, and is an important novel feature of his approach to cluster analysis.

The chapters that are focused on granular computing serve as a foundation for the core of the book—knowledge-based clustering. In this mode of clustering, clustering is guided by the knowledge that underlies data. There is much that is new in this part of the book, especially in chapters dealing with conditioned fuzzy clustering, collaborative clustering, directional clustering, fuzzy relational clustering, and clustering of nonhomogeneous patterns. The last part of Professor Pedrycz's work is an informative exposition of applications of knowledge-based clustering to generic models. In this part, we find a range of unconventional concepts and techniques, among them hyperbox modeling, linguistic modeling, and granular mapping.

To see the importance of Pedrycz's work in a proper perspective, an observation is in order. As we move further into the age of machine intelligence and automated reasoning, a daunting problem becomes harder and harder to master. How can we cope with the explosive growth in data, information, and knowledge? How can we locate and infer from decision-relevant information that is embedded in a large database that is unstructured, imprecise, and not totally reliable?