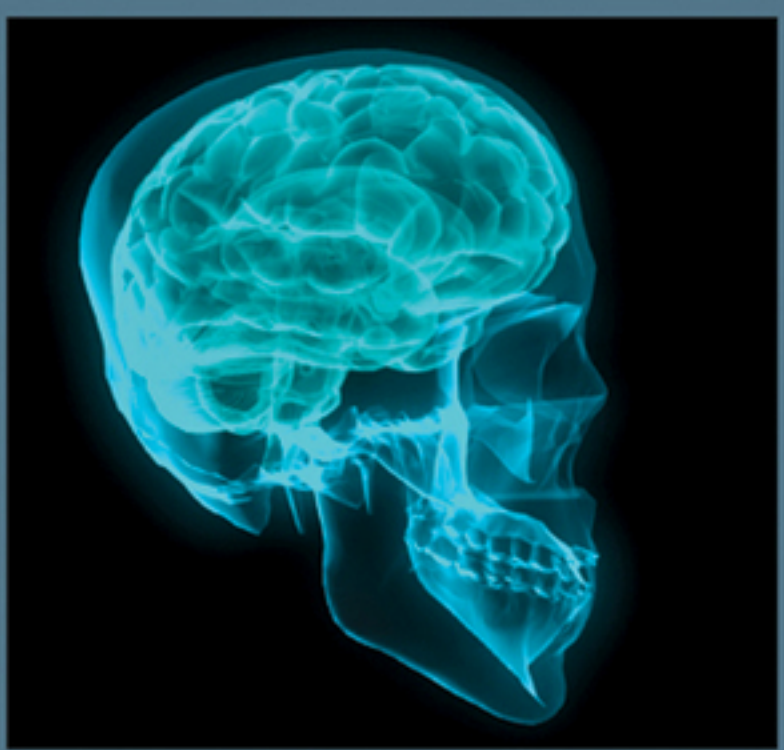
Wiley Series on Bioinformatics: Computational Techniques and Engineering
Yi Pan and Albert Y. Zomaya, Series Editors

Rough-Fuzzy Pattern Recognition

APPLICATIONS IN BIOINFORMATICS AND MEDICAL IMAGING



PRADIPTA MAJI • SANKAR K. PAL

WILEY
IEEE
IEEE computer society

ROUGH-FUZZY PATTERN RECOGNITION

Wiley Series on

Bioinformatics: Computational Techniques and Engineering

Bioinformatics and computational biology involve the comprehensive application of mathematics, statistics, science, and computer science to the understanding of living systems. Research and development in these areas require cooperation among specialists from the fields of biology, computer science, mathematics, statistics, physics, and related sciences. The objective of this book series is to provide timely treatments of the different aspects of bioinformatics spanning theory, new and established techniques, technologies and tools, and application domains. This series emphasizes algorithmic, mathematical, statistical, and computational methods that are central in bioinformatics and computational biology.

Series Editors: **Professor Yi Pan** and **Professor Albert Y. Zomaya**
pan@cs.gsu.edu zomaya@it.usyd.edu.au

Knowledge Discovery in Bioinformatics: Techniques, Methods, and Applications
Xiaohua Hu and Yi Pan

Grid Computing for Bioinformatics and Computational Biology
Edited by El-Ghazali Talbi and Albert Y. Zomaya

Bioinformatics Algorithms: Techniques and Applications
Ion Mandiou and Alexander Zelikovsky

Analysis of Biological Networks
Edited by Björn H. Junker and Falk Schreiber

Computational Intelligence and Pattern Analysis in Biological Informatics
Edited by Ujjwal Maulik, Sanghamitra Bandyopadhyay, and Jason T. L. Wang

Mathematics of Bioinformatics: Theory, Practice, and Applications
Matthew He and Sergey Petoukhov

Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications
Edited by Mourad Elloumi and Albert Y. Zomaya

Mathematical and Computational Methods in Biomechanics of Human Skeletal Systems: An Introduction
Jiří Nedoma, Jiří Stehlík, Ivan Hlaváček, Josef Daněk, Taťjana Dostálová, and Petra Přečková

Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging
Pradipta Maji and Sankar K. Pal

ROUGH-FUZZY PATTERN RECOGNITION

Applications in Bioinformatics and Medical Imaging

PRADIPTA MAJI
Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

SANKAR K. PAL
Center for Soft Computing Research, Indian Statistical Institute, Kolkata, India

IEEE PRESS

A JOHN WILEY & SONS, INC., PUBLICATION

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

Library of Congress Cataloging-in-Publication Data:

Maji, Pradipta, 1976–
Rough-fuzzy pattern recognition : applications in bioinformatics and medical imaging / Pradipta Maji, Sankar K. Pal.
p. cm. – (Wiley series in bioinformatics ; 3)
ISBN 978-1-118-00440-1 (hardback)
1. Fuzzy systems in medicine. 2. Pattern recognition systems. 3. Bioinformatics. 4. Diagnostic imaging–Data processing. I. Pal, Sankar K. II. Title.
R859.7.F89M35 2011
610.285–dc23

2011013787

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To our parents

CONTENTS

FOREWORD

It is my great pleasure to welcome the new book *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging* by the prominent scientists Professor Sankar K. Pal and Professor Pradipta Maji.

Soft computing methods allow us to achieve high-quality solutions for many real-life applications. The characteristic features of these methods are tractability, robustness, low-cost solution, and close resemblance with human-like decision making. They make it possible to use imprecision, uncertainty, approximate reasoning, and partial truth in searching for solutions. The main research directions in soft computing are related to fuzzy sets, neurocomputing, genetic algorithms, probabilistic reasoning, and rough sets. By integration or combination of the different soft computing methods, one may improve the performance of these methods. Among the various integrations realized so far, neuro-fuzzy computing (combing fuzzy sets and neural networks) is the most visible one because of its several real-life applications.

Both fuzzy and rough set theory represent two different approaches to analyzing vagueness. Fuzzy set theory addresses gradualness of knowledge, expressed by the fuzzy membership, whereas rough set theory addresses the granularity of knowledge, expressed by the indiscernibility relation. In 1999, together with Professor Sankar K. Pal, we edited the book *Rough-Fuzzy Hybridization* published by Springer. Since then, great progress has been made in the development of methods based on a combination of these approaches, both on foundations and applications. It is proved that by combining the rough-set and fuzzy-set approaches it is possible to achieve a significant improvement in the performance of methods. They are complementary to each other rather than competitive.

The book is based on a unified framework describing how rough-fuzzy computing techniques can be formulated for building efficient information granules, especially pattern recognition models. These granules are induced from some elementary granules by (hierarchical) fusion. The elementary granules can be induced using rough-set-based methods and/or fuzzy-set-based methods, and also the aggregation of granules can be based on a combination of such methods. In this way, one can consider different approaches such as rough-fuzzy, fuzzy-rough or fuzzy rough-fuzzy. For example, using rough-set-based methods one can efficiently induce some crisp patterns, which can be next fuzzified for making them soft. This can help, for example, in searching for the relevant fusion of granules. Analogously, the discovered fuzzy patterns may contain too many details and then, by using rough-set-based methods, one can obtain simpler patterns with satisfactory quality. Such patterns can be next used efficiently in approximate reasoning, for example, in the discovery of more complex patterns from the existing ones. In all these methods, the rough-set approach and the fuzzy-set approach work synergistically in efficient searching under uncertainty and imprecision for the target granules (e.g., classifiers for complex concepts) with a high quality. Note that in the fusion of granules, inclusion measures also play an important role because they allow us to estimate the quality of the granules constructed.

In a perfect way, the book introduces the reader to the fascinating and successful cooperative game between the rough-set approach and fuzzy-set approach. In this book, the reader will find a nice introduction to the rough-fuzzy approach and fuzzy-set approach. The discussed methods and algorithms cover all major phases of a pattern recognition system (e.g., classification, clustering, and feature selection). The book covers existing results and also presents new results. It was proved experimentally that the performance of the developed algorithms based on a combination of approaches is much better than the performance of algorithms based on approaches taken separately. This is shown in the book for several tasks such as feature selection of real valued data, case selection, image processing and analysis, data mining and knowledge discovery, selection of vocabulary for information retrieval, and decision rule extraction. The high performance of the developed methods is especially emphasized for real-life applications in bioinformatics and medical image processing. The balance of theory, algorithms, and applications will make this book attractive to many readers. The reader will also find in the book a discussion on various challenging issues relevant to application domains and possible ways of handling them with rough-fuzzy approaches.

This book, unique in its character, will be very useful to graduate students, researchers, and practitioners. The authors deserve the highest appreciation for their outstanding work. This is a work whose importance is hard to exaggerate.

ANDRZEJ SKOWRON

Institute of Mathematics
Warsaw University, Poland
December, 2011

PREFACE

Soft computing is a collection of methodologies that work synergistically, not competitively, that, in one form or another, reflects its guiding principle: exploits the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth to achieve tractability, robustness, low cost solution, and close resemblance with human-like decision making. It provides a flexible information processing capability for representation and evaluation of various real-life, ambiguous and uncertain situations and therefore results in the foundation for the conception and design of high machine intelligence quotient systems. At this juncture, the principal constituents of soft computing are fuzzy sets, neurocomputing, genetic algorithms, probabilistic reasoning, and rough sets.

One of the challenges of basic soft computing research is how to integrate its different constituting tools synergistically to achieve both generic and application-specific merits. Application-specific merits point to the advantages of integrated systems not achievable using the constituting tools singly.

Rough set theory, which is considered to be a newer soft computing tool compared with others, deals with uncertainty, vagueness, and incompleteness arising from the indiscernibility of objects in the universe. The main goal of rough set theoretic analysis is to synthesize or construct approximations, in terms of upper and lower bounds of concepts, properties, or relations from the acquired data. The key notions here are those of information granules and reducts. Information granules formalize the concept of finite precision representation of objects in real-life situations, and the reducts represent the core of an information system, both in terms of objects and features, in a granular universe. Its integration with fuzzy set theory, called rough-fuzzy computing, has motivated researchers to design a much stronger paradigm of reasoning and handling uncertainties associated with

the data compared with those of the individual ones. The generalized theories of rough-fuzzy sets (when a fuzzy set is defined over crisp granules), fuzzy-rough sets (when a crisp set is defined over fuzzy granules), and fuzzy rough-fuzzy sets (when a fuzzy set is defined over fuzzy granules) have been applied successfully to problems such as classification, clustering, feature selection of real valued data, case selection, image processing and analysis, data mining and knowledge discovery, selecting vocabulary for information retrieval, and decision rule extraction.

In rough-fuzzy pattern recognition, fuzzy sets are used for handling uncertainties arising from ill-defined or overlapping nature of concepts, classes or regions, in terms of membership values, while rough sets are used for granular computing and handling uncertainties, due to granulation or indiscernibility in the feature space, in terms of lower and upper approximations of concepts or regions. On the one hand, rough information granules are used, for example, in defining class exactness, and encoding domain knowledge. On the other hand, granular computing, which deals with clumps of indiscernible data together rather than individual data points, leads to computation gain, thereby making it suitable for mining large data sets. Furthermore, depending on the problems, granules can be class dependent or independent. Several clustering algorithms have been formulated where the incorporation of rough sets resulted in a balanced mixture between restrictive or hard clustering and descriptive or fuzzy clustering. Judicious integration of the concept of rough sets with the existing fuzzy clustering algorithms has made the latter work faster with improved performance. Various real-life applications of these models, including those in bioinformatics, have also been reported during the last five to seven years. These are available in different journals, conference proceedings, and edited volumes. This scattered information causes inconvenience to readers, students, and researchers.

The current volume is aimed at providing a treatise in a unified framework describing how rough-fuzzy computing techniques can be judiciously formulated and used in building efficient pattern recognition models. On the basis of the existing as well as new results, the book is structured according to the major phases of a pattern recognition system (for example, clustering, classification, and feature selection) with a balanced mixture of theory, algorithm, and applications. Special emphasis is given to applications in bioinformatics and medical image processing.

The book consists of nine chapters. Chapter 1 provides an introduction to pattern recognition and data mining, along with different research issues and challenges related to high dimensional real-life data sets. The significance of soft computing in pattern recognition and data mining is also presented in Chapter 1. Chapter 2 presents the basic notions and characteristics of two soft computing tools, namely, fuzzy sets and rough sets. These are followed by the concept of information granules, the emergence of a rough-fuzzy computing paradigm and their relevance to pattern recognition. It also provides a mathematical framework for generalized rough sets incorporating the concept of fuzziness in defining the granules as well as the set. Various roughness and uncertainty measures with

properties are reported. Different research issues related to rough granules are stated.

Chapter 3 mainly centers on a generalized unsupervised learning (clustering) algorithm, termed as rough-fuzzy-possibilistic c-means. While the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in class definition, the membership function of fuzzy sets enables efficient handling of overlapping partitions. It incorporates both probabilistic and possibilistic memberships simultaneously to avoid the problems of noise sensitivity of fuzzy c-means and the coincident clusters of possibilistic c-means. The concept of crisp lower bound and fuzzy boundary of a class enables efficient selection of cluster prototypes. The algorithm is generalized in the sense that all the existing variants of c-means algorithms can be derived from this as a special case. Superiority in terms of computation time and performance is demonstrated. Several quantitative indices are reported for evaluating the performance on various real-life data sets.

Chapter 4 provides various supervised classification methods based on class-dependent f-granulation and rough set theoretic feature selection. The significance of granulation for better class discriminatory information and neighborhood rough sets for better feature selection is demonstrated. Extensive experimental results with quantitative indices are provided on both fully and partially labeled data sets. Future directions on the use of this concept in other computing paradigms are also provided.

Selection of nonredundant and relevant features of real valued data sets is a highly challenging problem. Chapter 5 addresses this issue. Methods described here are based on fuzzy-rough sets by maximizing the relevance and minimizing the redundancy of the selected features. Various new concepts such as fuzzy equivalence partition matrix, representation of Shannon's entropy for fuzzy approximation spaces, f-information measures to compute both relevance and redundancy of features, and feature evaluation indices are stated along with experimental results.

While several experimental results on both artificial and real-life data sets, including speech and remotely sensed multi-spectral image data, are provided in Chapters 3, 4, and 5 to demonstrate the effectiveness of the respective rough-fuzzy methodologies, the next four chapters are concerned only with certain specific applications, namely, in bioinformatics and medical imaging. Problems considered in bioinformatics include selection of a minimum set of bio-basis strings with maximum information for amino acid sequence analysis (Chapter 6), grouping functionally similar genes from microarray gene expression data through clustering (Chapter 7), and selection of relevant genes from high dimensional microarray data (Chapter 8). Selection of bio-basis strings is done by devising a relational clustering algorithm, called rough-fuzzy c-medoids. It judiciously integrates rough sets, fuzzy sets, and amino acid mutation matrices with hard c-medoids algorithms. The selected bio-basis strings are evaluated with newly defined indices in terms of a homology alignment score. Gene clusters thus identified may contribute to revealing underlying class structures, providing a

useful tool for the exploratory analysis of biological data. The concept of fuzzy equivalence partition matrix, based on the theory of fuzzy-rough sets, is shown to be effective for selecting relevant and nonredundant continuous valued genes from high dimensional microarray gene expression data.

Problems of segmentation of brain MR images for visualization of human tissues during clinical analysis is addressed in Chapter 9 using rough-fuzzy clustering with new indices for feature extraction. The concept of discriminant analysis, based on the maximization of class separability, is used to circumvent the problems of initialization and local minima of rough-fuzzy clustering. Different challenging issues in the respective application domains and the ways of handling them with rough-fuzzy approaches are also discussed.

The relevant existing conventional/traditional approaches or techniques are also included wherever necessary. Directions for future research in the concerned topic are provided in each chapter. Most of the materials presented in the book are from our published works. For the convenience of readers, a comprehensive bibliography on the subject is also appended in each chapter. Some works in the related areas might have been omitted because of oversight or ignorance.

This book, which is unique in its character, will be useful to graduate students and researchers in computer science, electrical engineering, system science, medical science, bioinformatics, and information technology, both as a textbook and a reference book for some parts of the curriculum. Researchers and practitioners in industry and R&D laboratories working in the fields of system design, pattern recognition, machine learning, image analysis, vision, data mining, bioinformatics, and soft computing or computational intelligence will also be benefited.

Finally, the authors take this opportunity to thank Mr. Michael Christian and Dr. Simone Taylor of John Wiley & Sons, Inc., Hoboken, New Jersey, for their initiative and encouragement. The authors also gratefully acknowledge the support provided by Prof. Malay K. Kundu, Dr. Saroj K. Meher, Dr. Debashis Sen, Ms. Sushmita Paul, and Mr. Indranil Dutta of Indian Statistical Institute in the preparation and proofreading of a few chapters of the manuscript. The book was written when one of the authors, Prof. S. K. Pal, held a J. C. Bose National Fellowship of the Government of India.

PRADIPTA MAJI
SANKAR K. PAL

Kolkata, India

ABOUT THE AUTHORS

Pradipta Maji received his BSc degree in Physics, MSc degree in Electronics Science, and PhD degree in the area of Computer Science from Jadavpur University, India, in 1998, 2000, and 2005, respectively.

Currently, he is an assistant professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. He was associated with the Center for Soft Computing Research: A National Facility, Indian Statistical Institute, Kolkata, India, from 2005 to 2009. Before joining the Indian Statistical Institute, he was a lecturer in the Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India, from 2004 to 2005. In 2004, he visited the Laboratory of Information Security & Internet Applications (LISIA), Division of Electronics, Computer and Telecommunication Engineering, Pkyoung National University, Pusan, Korea. During the period of September 2000 to April 2004, he was a research scholar at the Department of Computer Science and Technology, Bengal Engineering College (DU) (currently known as Bengal Engineering and Science University), Shibpur, Howrah, India. From 2002 to 2004, he also served as a Research and Development Consultant of Cellular Automata Research Laboratory (CARL), Kolkata, India. His research interests include pattern recognition, computational biology and bioinformatics, medical image processing, cellular automata, and soft computing. He has published more than 60 papers in international journals and conferences and is a reviewer for many international journals.

Dr. Maji received the 2006 Best Paper Award of the International Conference on Visual Information Engineering from The Institution of Engineering and Technology, UK, the 2008 Microsoft Young Faculty Award from Microsoft Research Laboratory India Pvt., the 2009 Young Scientist Award from the National Academy of Sciences, India, and the 2011 Young Scientist Award from

the Indian National Science Academy, and was selected as the 2009 Associate of the Indian Academy of Sciences.

Sankar K. Pal is a distinguished scientist of the Indian Statistical Institute and a former Director. He is also a J.C. Bose Fellow of the Government of India. He founded the Machine Intelligence Unit and the Center for Soft Computing Research, a national facility in the institute in Calcutta. He received his PhD in Radio Physics and Electronics from the University of Calcutta in 1979, and another PhD in Electrical Engineering along with DIC from Imperial College, University of London, in 1982. He joined his institute in 1975 as a CSIR Senior Research Fellow and later became a Full Professor in 1987, Distinguished Scientist in 1998, and the Director for the term 2005–2010.

He worked at the University of California, Berkeley, and the University of Maryland, College Park, in 1986–1987; the NASA Johnson Space Center, Houston, Texas, in 1990–1992 and 1994; and the US Naval Research Laboratory, Washington, DC, in 2004. Since 1997 he has served as a distinguished visitor of the IEEE Computer Society (USA) for the Asia-Pacific region, and held several visiting positions at universities in Italy, Poland, Hong Kong, and Australia.

Prof. Pal is a Fellow of the IEEE, USA, the Academy of Sciences for the Developing World (TWAS), Italy, International Association for Pattern Recognition, USA, International Association of Fuzzy Systems, USA, and all the four National Academies for Science/Engineering in India. He is a coauthor of 17 books and more than 400 research publications in the areas of pattern recognition and machine learning, image processing, data mining and web intelligence, soft computing, neural nets, genetic algorithms, fuzzy sets, rough sets, and bioinformatics.

He received the 1990 S.S. Bhatnagar Prize (which is the most coveted award for a scientist in India), and many prestigious awards in India and abroad including the 1999 G.D. Birla Award, 1998 Om Bhasin Award, 1993 Jawaharlal Nehru Fellowship, 2000 Khwarizmi International Award from the Islamic Republic of Iran, 2000–2001 FICCI Award, 1993 Vikram Sarabhai Research Award, 1993 NASA Tech Brief Award (USA), 1994 IEEE Transactions Neural Networks Outstanding Paper Award (USA), 1995 NASA Patent Application Award (USA), 1997 IETE-R.L. Wadhwa Gold Medal, the 2001 INSA-S.H. Zaheer Medal, 2005–2006 Indian Science Congress-P.C. Mahalanobis Birth Centenary Award (Gold Medal) for Lifetime Achievement, 2007 J.C. Bose Fellowship of the Government of India, and 2008 Vigyan Ratna Award from Science & Culture Organization, West Bengal.

Prof. Pal is currently or has in the past been an Associate Editor of *IEEE Trans. Pattern Analysis and Machine Intelligence* (2002–2006), *IEEE Trans. Neural Networks* [1994–1998 & 2003–2006], *Neurocomputing* (1995-2005), *Pattern Recognition Letters*, *Int. J. Pattern Recognition & Artificial Intelligence*, *Applied Intelligence*, *Information Sciences*, *Fuzzy Sets and Systems*, *Fundamenta Informaticae*, *LNCS Trans. on Rough Sets*, *Int. J. Computational Intelligence and Applications*, *IET Image Processing*, *J. Intelligent Information Systems*, and *Proc.*

INSA-A; Editor-in-Chief, *Int. J. Signal Processing, Image Processing and Pattern Recognition*; Series Editor, *Frontiers in Artificial Intelligence and Applications*, IOS Press, and *Statistical Science and Interdisciplinary Research*, World Scientific; a Member, Executive Advisory Editorial Board, *IEEE Trans. Fuzzy Systems*, *Int. Journal on Image and Graphics*, and *Int. Journal of Approximate Reasoning*; and a Guest Editor of *IEEE Computer*.

1

INTRODUCTION TO PATTERN RECOGNITION AND DATA MINING

1.1 INTRODUCTION

Pattern recognition is an activity that human beings normally excel in. The task of pattern recognition is encountered in a wide range of human activity. In a broader perspective, the term could cover any context in which some decision or forecast is made on the basis of currently available information. Mathematically, the problem of pattern recognition deals with the construction of a procedure to be applied to a set of inputs; the procedure assigns each new input to one of a set of classes on the basis of observed features. The construction of such a procedure on an input data set is defined as pattern recognition.

A pattern typically comprises some features or essential information specific to a pattern or a class of patterns. Pattern recognition, as per the convention, is the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns. In other words, the discipline of pattern recognition essentially deals with the problem of developing algorithms and methodologies that can enable the computer implementation of many recognition tasks that humans normally perform. The objective is to perform these tasks more accurately, faster, and perhaps more economically than humans and, in many cases, to release them from drudgery resulting from performing routine recognition tasks repetitively and mechanically. The scope of pattern recognition also encompasses tasks at which humans are not good, such as reading bar codes. Hence, the goal

Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging,
First Edition. Pradipta Maji and Sankar K. Pal.

of pattern recognition research is to devise ways and means of automating certain decision-making processes that lead to classification and recognition.

Pattern recognition can be viewed as a twofold task, consisting of learning the invariant and common properties of a set of samples characterizing a class, and of deciding that a new sample is a possible member of the class by noting that it has properties common to those of the set of samples. The task of pattern recognition can be described as a transformation from the measurement space $\mathcal{M}$ to the feature space $\mathcal{F}$ and finally to the decision space $\mathcal{D}$; that is,

$$\mathcal{M} \rightarrow \mathcal{F} \rightarrow \mathcal{D}, \tag{1.1}$$

where the mapping $\delta : \mathcal{F} \rightarrow \mathcal{D}$ is the decision function, and the elements $d \in \mathcal{D}$ are termed as *decisions*.

Pattern recognition has been a thriving field of research for the past few decades [1–8]. The seminal article by Kanal [9] gives a comprehensive review of the advances made in the field until the early 1970s. More recently, a review article by Jain et al. [10] provides an engrossing survey of the advances made in statistical pattern recognition till the end of the twentieth century. Although the subject has attained a very mature level during the past four decades or so, it remains green to the researchers because of continuous cross-fertilization of ideas from disciplines such as computer science, physics, neurobiology, psychology, engineering, statistics, mathematics, and cognitive science. Depending on the practical need and demand, various modern methodologies have come into being, which often supplement the classical techniques [11].

In recent years, the rapid advances made in computer technology have ensured that large sections of the world population have been able to gain easy access to computers on account of the falling costs worldwide, and their use is now commonplace in all walks of life. Government agencies and scientific, business, and commercial organizations routinely use computers, not only for computational purposes but also for storage, in massive databases, of the immense volumes of data that they routinely generate or require from other sources. Large-scale computer networking has ensured that such data has become accessible to more and more people. In other words, we are in the midst of an information explosion, and there is an urgent need for methodologies that will help us to bring some semblance of order into the phenomenal volumes of data that can readily be accessed by us with a few clicks of the keys of our computer keyboard. Traditional statistical data summarization and database management techniques are just not adequate for handling data on this scale and for intelligently extracting information, or rather, knowledge that may be useful for exploring the domain in question or the phenomena responsible for the data, and providing support to decision-making processes. This quest has thrown up a new phrase, called *data mining* [12–14].

The massive databases are generally characterized by the numeric as well as textual, symbolic, pictorial, and aural data. They may contain redundancy, errors, imprecision, and so on. Data mining is aimed at discovering natural structures

within such massive and often heterogeneous data. It is visualized as being capable of knowledge discovery using generalizations and magnifications of existing and new pattern recognition algorithms. Therefore, pattern recognition plays a significant role in the data mining process. Data mining deals with the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Hence, it can be viewed as applying pattern recognition and machine learning principles in the context of voluminous, possibly heterogeneous data sets [11].

One of the important problems in real-life data analysis is uncertainty management. Some of the sources of this uncertainty include incompleteness and vagueness in class definitions. In this background, the possibility concept introduced by the fuzzy sets theory [15] and rough sets theory [16] have gained popularity in modeling and propagating uncertainty. Both the fuzzy sets and rough sets provide a mathematical framework to capture uncertainties associated with the data [17]. They are complementary in some aspects. The generalized theories of rough-fuzzy sets and fuzzy-rough sets have been applied successfully to feature selection of real-valued data [18, 19], classification [20], image processing [21], data mining [22], information retrieval [23], fuzzy decision rule extraction, and rough-fuzzy clustering [24, 25].

The objective of this book is to provide some results of investigations, both theoretical and experimental, addressing the relevance of rough-fuzzy approaches to pattern recognition with real-life applications. Various methodologies are presented, integrating fuzzy logic and rough sets for clustering, classification, and feature selection. The emphasis of these methodologies is given on (i) handling data sets which are large, both in size and dimension, and involve classes that are overlapping, intractable and/or having nonlinear boundaries; (ii) demonstrating the significance of rough-fuzzy granular computing in soft computing paradigm for dealing with the knowledge discovery aspect; and (iii) demonstrating their success in certain tasks of bioinformatics and medical imaging as an example. Before describing the scope of the book, a brief review of pattern recognition, data mining, and application of pattern recognition algorithms in data mining problems is provided.

The structure of the rest of this chapter is as follows: Section 1.2 briefly presents a description of the basic concept, features, and techniques of pattern recognition. In Section 1.3, the data mining aspect is elaborated, discussing its components, tasks involved, approaches, and application areas. The pattern recognition perspective of data mining is introduced next and related research challenges are mentioned. The role of soft computing in pattern recognition and data mining is described in Section 1.4. Finally, Section 1.5 discusses the scope and organization of the book.

1.2 PATTERN RECOGNITION

A typical pattern recognition system consists of three phases, namely, *data acquisition*, *feature selection or extraction*, and *classification or clustering*. In the data

acquisition phase, depending on the environment within which the objects are to be classified or clustered, data are gathered using a set of sensors. These are then passed on to the feature selection or extraction phase, where the dimensionality of the data is reduced by retaining or measuring only some characteristic features or properties. In a broader perspective, this stage significantly influences the entire recognition process. Finally, in the classification or clustering phase, the selected or extracted features are passed on to the classification or clustering system that evaluates the incoming information and makes a final decision. This phase basically establishes a transformation between the features and the classes or clusters [1, 2, 8].

1.2.1 Data Acquisition

In data acquisition phase, data are gathered via a set of sensors depending on the environment within which the objects are to be classified. Pattern recognition techniques are applicable in a wide domain, where the data may be qualitative, quantitative, or both; they may be numerical, linguistic, pictorial, or any combination thereof. Generally, the data structures that are used in pattern recognition systems are of two types: object data vectors and relational data. Object data, sets of numerical vectors of m features, are represented as $X = \{x_1, \ldots, x_i, \ldots, x_n\}$, a set of n feature vectors in the m-dimensional measurement space $\Re^m$. The ith object observed in the process has vector x_i as its numerical representation; x_{ij} is the jth ($j = 1, \ldots, m$) feature associated with the ith object. On the other hand, relational data are a set of n^2 numerical relationships, say r_{ij}, between pairs of objects. In other words, r_{ij} represents the extent to which objects x_i and x_j are related in the sense of some binary relationship ρ. If the objects that are pairwise related by ρ are called $O = \{o_1, \ldots, o_i, \ldots, o_n\}$, then $\rho : O \times O \rightarrow \Re$.

1.2.2 Feature Selection

Feature selection or extraction is a process of selecting a map by which a sample in an m-dimensional measurement space is transformed into a point in a d-dimensional feature space, where $d < m$ [1, 8]. Mathematically, it finds a mapping of the form $y = f(x)$, by which a sample $x = [x_1, \ldots, x_j, \ldots, x_m]$ in an m-dimensional measurement space $\mathcal{M}$ is transformed into an object $y = [y_1, \ldots, y_j, \ldots, y_d]$ in a d-dimensional feature space $\mathcal{F}$.

The main objective of this task is to retain or generate the optimum salient characteristics necessary for the recognition process and to reduce the dimensionality of the measurement space so that effective and easily computable algorithms can be devised for efficient classification. The problem of feature selection or extraction has two aspects, namely, formulating a suitable criterion to evaluate the goodness of a feature set and searching the optimal set in terms of the criterion. In general, those features are considered to have optimal saliencies for which interclass (respectively, intraclass) distances are maximized (respectively, minimized). The criterion for a good feature is that it should be unchanging with

any other possible variation within a class, while emphasizing differences that are important in discriminating between patterns of different types.

The major mathematical measures so far devised for the estimation of feature quality are mostly statistical in nature, and can be broadly classified into two categories, namely, feature selection in the measurement space and feature selection in a transformed space. The techniques in the first category generally reduce the dimensionality of the measurement space by discarding redundant or least information-carrying features. On the other hand, those in the second category utilize all the information contained in the measurement space to obtain a new transformed space, thereby mapping a higher dimensional pattern to a lower dimensional one. This is referred to as *feature extraction* [1, 2, 8].

1.2.3 Classification and Clustering

The problem of classification and clustering is basically one of partitioning the feature space into regions, one region for each category of input. Hence, it attempts to assign every data object in the entire feature space to one of the possible classes or clusters. In real life, the complete description of the classes is not known. Instead, a finite and usually smaller number of samples are available, which often provide partial information for optimal design of feature selector or extractor or classification or clustering system. Under such circumstances, it is assumed that these samples are representative of the classes or clusters. Such a set of typical patterns is called a *training set*. On the basis of the information gathered from the samples in the training set, the pattern recognition systems are designed. That is, the values of the parameters of various pattern recognition methods are decided.

Design of a classification or clustering scheme can be made with labeled or unlabeled data. When the algorithm is given a set of objects with known classifications, that is, labels, and is asked to classify an unknown object based on the information acquired by it during training, the design scheme is called *supervised learning*; otherwise it is *unsupervised learning*. Supervised learning is used for classifying different objects, while clustering is performed through unsupervised learning. Through cluster analysis, a given data set is divided into a set of clusters in such a way that two objects from the same cluster are as similar as possible and the objects from different clusters are as dissimilar as possible. In effect, it tries to mimic the human ability to group similar objects into classes and categories. A number of clustering algorithms have been proposed to suit different requirements [2, 26, 27].

Pattern classification or clustering, by its nature, admits many approaches, sometimes complementary, sometimes competing, to provide the solution to a given problem. These include decision theoretic approach (both deterministic and probabilistic), syntactic approach, connectionist approach, fuzzy and rough set theoretic approaches and hybrid or soft computing approach. Let $\beta = \{\beta_1, \ldots, \beta_i, \ldots, \beta_c\}$ represent the c possible classes or clusters in a d-dimensional feature space $\mathcal{F}$, and $y = [y_1, \ldots, y_j, \ldots, y_d]$ be an unknown

pattern vector whose class is to be identified. In deterministic classification or clustering approach, the object is assigned to only one unambiguous pattern class or cluster β_i if the decision function D_i associated with the class β_i satisfies the following relation:

$$D_i(y) > D_j(y), \qquad j = 1, \ldots, c, \text{ and } j \neq i. \tag{1.2}$$

In the decision theoretic approach, once a pattern is transformed through feature evaluation to a vector in the feature space, its characteristics are expressed only by a set of numerical values. Classification can be done by using deterministic or probabilistic techniques [1, 2, 8]. The nearest neighbor classifier [2] is an example of deterministic classification approach, where it is assumed that there exists only one unambiguous pattern class corresponding to each of the unknown pattern vectors. In most of the practical problems, the features are often noisy and the classes in the feature space are overlapping. In order to model such systems, the features are considered as random variables in the probabilistic approach. The most commonly used classifier in such probabilistic systems is the Bayes maximum likelihood classifier [2].

When a pattern is rich in structural information such as picture recognition, character recognition, scene analysis, that is, the structural information plays an important role in describing and recognizing the patterns, it is convenient to use the syntactic approach [3]. It deals with the representation of structures via sentences, grammars, and automata. In the syntactic method [3], the ability of selecting and classifying the simple pattern primitives and their relationships represented by the composition operations is the vital criterion for making a system effective. Since the techniques of composition of primitives into patterns are usually governed by the formal language theory, the approach is often referred to as a *linguistic approach*. An introduction to a variety of approaches based on this idea can be found in Fu [3]. Other approaches to pattern recognition are discussed in Section 1.4 under soft computing methods.

1.3 DATA MINING

Data mining involves fitting models to or determining patterns from observed data. The fitted models play the role of inferred knowledge. Typically, a data mining algorithm constitutes some combination of three components, namely, model, preference criterion, and search algorithm [13].

The model represents its function (e.g., classification, clustering) and its representational form (e.g., linear discriminants, neural networks). A model contains parameters that are to be determined from the data. The preference criterion is a basis to decide the preference of one model or a set of parameters over another, depending on the given data. The criterion is usually some form of goodness of fit function of the model to the data, perhaps tempered by a smoothing term to avoid overfitting, or generating a model with too many degrees of freedom to