Unsupervised Classification

Sanghamitra Bandyopadhyay • Sriparna Saha

Unsupervised Classification

Similarity Measures, Classical and Metaheuristic Approaches, and Applications

Springer

Sanghamitra Bandyopadhyay
Machine Intelligence Unit
Indian Statistical Institute
Kolkata, West Bengal, India

Sriparna Saha
Dept. of Computer Science
Indian Institute of Technology
Patna, India

ISBN 978-3-642-32450-5 ISBN 978-3-642-32451-2 (eBook)
DOI 10.1007/978-3-642-32451-2
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012953901

ACM Computing Classification (1998): I.2, J.3, H.2, H.3

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my parents Satyendra Nath Banerjee and Bandana Banerjee, my husband Prof. Ujjwal Maulik, and son Utsav
Sanghamitra Bandyopadhyay

To my husband Dr. Asif Ekbal, my daughter Spandana, and my parents Sankar Prasad Saha and Sila Saha
Sriparna Saha

Preface

Classification is an integral part of any pattern recognition system. Depending on whether a set of labeled training samples is available or not, classification can be either supervised or unsupervised. Clustering is an important unsupervised classification technique where a number of data points are grouped into clusters such that points belonging to the same cluster are similar in some sense and points belonging to different clusters are dissimilar in the same sense. Cluster analysis is a complex problem as a variety of similarity and/or dissimilarity measures exist in the literature without any universal definition. In a crisp clustering technique, each pattern is assigned to exactly one cluster, whereas in the case of fuzzy clustering, each pattern is given a membership degree to each class. Fuzzy clustering is inherently more suitable for handling imprecise and noisy data with overlapping clusters.

For partitioning a data set, one has to define a measure of similarity or proximity based on which cluster assignments are done. The measure of similarity is usually data dependent. It may be noted that, in general, one of the fundamental features of shapes and objects is symmetry, which is considered to be important for enhancing their recognition. Examples of symmetry abound in real life, such as the human face, human body, flowers, leaves, and jellyfish. As symmetry is so common, it may be interesting to exploit this property while clustering a data set. Based on this observation, in recent years a large number of symmetry-based similarity measures have been proposed. This book is focused on different issues related to clustering, with particular emphasis on symmetry-based and metaheuristic approaches.

The aim of a clustering technique is to find a suitable grouping of the input data set so that some criteria are optimized. Hence, the problem of clustering can be posed as an optimization problem. The objective to be optimized may represent different characteristics of the clusters, such as compactness, symmetrical compactness, separation between clusters, connectivity within a cluster, etc. A straightforward way to pose clustering as an optimization problem is to optimize some cluster validity index that reflects the goodness of the clustering solutions. All possible values of the chosen optimization criterion (validity index) define the complete search space. Traditional partitional clustering techniques, such as K-means and fuzzy C-means, employ a greedy search technique over the search space in order to optimize

the compactness of the clusters. Although these algorithms are computationally efficient, they suffer from some drawbacks. They often get stuck at some local optima depending on the choice of the initial cluster centers. They are not able to determine the appropriate number of clusters from data sets and/or are capable of detecting clusters of some specific shape only.

To overcome the problem of getting stuck at local optima, several metaheuristic optimization tools, such as genetic algorithms (GAs), simulated annealing (SA), differential evolution (DE), etc., have been widely used to reach the global optimum value of the chosen validity measure. These techniques perform multimodal search in complex landscapes and provide near-optimal solutions for the objective or fitness function of an optimization problem. They have applications in fields as diverse as pattern recognition, image processing, neural networks, machine learning, job shop scheduling, and very large-scale integration (VLSI design), to mention just a few.

The two fundamental questions that need to be addressed in any typical clustering scenario are: (i) how many clusters are actually present in the data and (ii) how real or good is the clustering itself. That is, whatever the clustering technique, one has to determine the number of clusters and also the validity of the clusters formed. The measure of validity of clusters should be such that it will be able to impose an ordering of the clusters in terms of their goodness. Several cluster validity indices have been proposed in the literature, e.g., the Davies-Bouldin (DB) index, Dunn's index, Xie-Beni (XB) index, I-index, CS-index, etc., to name just a few. In recent years, several symmetry-based cluster validity indices have also been developed which are able to detect any kind of symmetric cluster from data sets. In this book, we discuss in detail several existing well-known cluster validity indices as well as some recently proposed symmetry-based versions.

Conventional GA-based clustering techniques, as well as related approaches, use some validity measure for optimization. Most of the existing clustering algorithms are able to determine hyperspherical/ellipsoidal-shaped clusters depending on the distance norm used. In recent years, some symmetry-based clustering techniques have been developed which can determine the appropriate number of clusters and the appropriate partitioning from data sets having any type of clusters, irrespective of their geometrical shape and overlapping nature, as long as they possess the characteristic of symmetry. A major focus of this book is on GA-based clustering techniques, which use symmetry as a similarity measure. Some of these clustering techniques are also able to detect the number of clusters automatically.

In many real-life situations one may need to optimize several objectives simultaneously. These problems are known as multiobjective optimization problems (MOOPs). In this regard, a multitude of metaheuristic single-objective optimization techniques such as genetic algorithms, simulated annealing, differential evolution, and their multiobjective versions have been developed. In this book we discuss in detail some existing single- and multiobjective optimization techniques. Moreover, a newly developed multiobjective simulated annealing-based technique is elaborately described and its effectiveness for solving several benchmark test problems is shown.

In the realm of clustering, simultaneous optimization of multiple validity indices, capturing different characteristics of the data, is expected to provide improved robustness to different data properties. Hence, it is useful to apply some multiobjective optimization (MOO) technique to solve the clustering problem. In MOO, search is performed over a number of objective functions. In contrast to single-objective optimization, which yields a single best solution, in MOO the final solution set contains a number of Pareto-optimal solutions, none of which can be further improved with regard to any one objective without degrading another. In a part of this book some recently developed multiobjective clustering techniques are discussed in detail.

A major focus of this book is on several real-life applications of clustering techniques in domains such as remote sensing satellite images, magnetic resonance (MR) brain images, and face detection. Analysis of remote sensing satellite images has significant utility in different areas such as climate studies, assessment of forest resources, examining marine environments, etc. An important task in remote sensing applications is the classification of pixels in the images into homogeneous regions, each of which corresponds to some particular land cover type. This problem has often been modeled as a segmentation problem, and clustering methods have been used to solve it. However, since it is difficult to have a priori information about the number of clusters in satellite images, the clustering algorithms should be able to automatically determine this value. Moreover, in satellite images it is often the case that some regions occupy only a few pixels, while the neighboring regions are significantly large. Thus, automatically detecting regions or clusters of such widely varying sizes presents a challenge in designing segmentation algorithms. In this book, we explore the applications of some symmetry-based clustering algorithms to classify remote sensing imagery in order to demonstrate their effectiveness.

Automatically classifying brain tissues from magnetic resonance images (MRI) has a major role in research and clinical study of neurological pathology. Additionally, regional volume calculations may provide even more useful diagnostic information. Automatic segmentation of brain MR images is a complex problem. Clustering approaches have been widely used for segmentation of MR brain images. A part of this book is dedicated to the applications of some symmetry-based clustering algorithms to classify MR brain images in order to demonstrate their effectiveness.

In recent years the problem of human face recognition has gained huge popularity. There are wide applications of face recognition systems including secure access control and financial transactions. The first important step of fully automatic human face recognition is human face detection. Face detection is the technique to automatically determine the locations and sizes of faces in a given input image. In this book a procedure to detect human faces based on symmetry is also described in detail.

This book aims to provide a treatise on clustering in a metaheuristic framework, with extensive applications to real-life data sets. Chapter 1 provides an introduction to pattern recognition, machine learning, classification, clustering, and related areas, along with different applications of pattern recognition. Chapter 2 describes in detail existing metaheuristic-based single- and multiobjective optimization techniques. An elaborate discussion on a recently developed multiobjective simulated annealing-based technique, archived multiobjective simulated annealing (AMOSA), is also

provided in this chapter. The utility of this technique to solve several benchmark test problems is shown. Chapter 3 mainly describes the different types of similarity measures developed in the literature for handling binary, categorical, ordinal, and quantitative variables. It also contains a discussion on different normalization techniques. Chapter 4 gives a broad overview of the existing clustering techniques and their relative advantages and disadvantages. It also provides a detailed discussion of several recently developed single- and multiobjective metaheuristic clustering techniques. Chapter 5 presents symmetry-based distances and a genetic algorithm-based clustering technique that uses this symmetry-based distance for assignment of points to different clusters and for fitness computation. In Chap. 6, some symmetry-based cluster validity indices are described. Elaborate experimental results are also presented in this chapter. Application of these symmetry-based cluster validity indices to segment remote sensing satellite images is also presented. In Chap. 7, some automatic clustering techniques based on symmetry are presented. These techniques are able to determine the appropriate number of clusters and the appropriate partitioning from data sets having symmetric clusters. The effectiveness of these clustering techniques is shown for many artificial and real-life data sets, including MR brain image segmentation. Chapter 8 deals with an extension of the concept of point symmetry to line symmetry-based distances, and genetic algorithm-based clustering techniques using these distances. Results are presented for some artificial and real-life data sets. A technique using the line symmetry-based distance for face detection from images is also discussed in detail. Some multiobjective clustering techniques based on symmetry are described in detail in Chap. 9. Three different clustering techniques are discussed here. The first one assumes the number of clusters a priori. The second and third clustering techniques are able to detect the appropriate number of clusters from data sets automatically. The third one, apart from using the concept of symmetry, also uses the concept of connectivity. A method of measuring connectivity between two points in a cluster is described for this purpose. A connectivity-based cluster validity index is also discussed in this chapter. Extensive experimental results illustrating the greater effectiveness of the three multiobjective clustering techniques over the single-objective approaches are presented for several artificial and real-life data sets.

This book contains an in-depth discussion on clustering and its various facets. In particular, it concentrates on metaheuristic clustering using symmetry as a similarity measure with extensive real-life applications in data mining, satellite remote sensing, MR brain imaging, gene expression data, and face detection. It is, in this sense, a complete book that will be equally useful to the layman and beginner as to an advanced researcher in clustering, being valuable for several reasons.

It includes discussions on traditional as well as some recent symmetry-based similarity measurements. Existing well-known clustering techniques along with metaheuristic approaches are elaborately described. Moreover, some recent clustering techniques based on symmetry are described in detail. Multiobjective clustering is another emerging topic in unsupervised classification. A multiobjective data clustering technique is described elaborately in this book along with extensive experimental results. A chapter of this book is wholly devoted to discussing some existing and

new multiobjective clustering techniques. Extensive real-life applications in remote sensing satellite imaging, MR brain imaging, and face detection are also provided in the book. Theoretical analyses of some recent symmetry-based clustering techniques are included.

The book will be useful to graduate students and researchers in computer science, electrical engineering, system science, and information technology as both text and reference book for some parts of the curriculum. Theoretical and experimental researchers will benefit from the discussions in this book. Researchers and practitioners in industry and R & D laboratories working in fields such as pattern recognition, data mining, soft computing, remote sensing, and biomedical imaging will also benefit.

The authors gratefully acknowledge the initiative and support for the project provided by Mr. Ronan Nugent of Springer. Sanghamitra Bandyopadhyay acknowledges the support provided by the Swarnajayanti Project Grant (No. DST/SJF/ET-02/2006-07) of the Department of Science and Technology, Government of India. The authors acknowledge the help of Mr. Malay Bhattacharyya, senior research fellow of ISI Kolkata, for drawing some of the figures of Chap. 5. Sriparna Saha also acknowledges the help of her students Mridula, Abhay, and Utpal for proofreading the chapters.

Kolkata, India

Sanghamitra Bandyopadhyay
Sriparna Saha

Contents

1 Introduction 1
1.1 Introduction 1
1.2 Data Types with Examples 3
1.3 Pattern Recognition: Preliminaries 4
1.3.1 Data Acquisition 6
1.3.2 Feature Selection 7
1.3.3 Classification 7
1.3.4 Clustering 8
1.3.5 Distance Measures in Clustering 9
1.3.6 Model Selection 10
1.3.7 Model Order Selection 10
1.4 Robustness to Outliers and Missing Values 11
1.5 Fuzzy Set-Theoretic Approach: Relevance and Features 12
1.6 Applications of Pattern Recognition and Learning 13
1.7 Summary and Scope of the Book 14

2 Some Single- and Multiobjective Optimization Techniques 17
2.1 Introduction 17
2.2 Single-Objective Optimization Techniques 18
2.2.1 Overview of Genetic Algorithms 19
2.2.2 Simulated Annealing 23
2.3 Multiobjective Optimization 25
2.3.1 Multiobjective Optimization Problems 26
2.3.2 Various Methods to Solve MOOPs 28
2.3.3 Recent Multiobjective Evolutionary Algorithms 29
2.3.4 MOOPs and SA 33
2.4 An Archive-Based Multiobjective Simulated Annealing Technique: AMOSA 36
2.4.1 Introduction 36
2.4.2 Archived Multiobjective Simulated Annealing (AMOSA) . 37
2.4.3 Archive Initialization 38

2.4.4 Clustering the Archive Solutions . . . 38
2.4.5 Amount of Domination . . . 40
2.4.6 The Main AMOSA Process . . . 41
2.4.7 Complexity Analysis . . . 45
2.5 Simulation Results . . . 46
2.5.1 Comparison Measures . . . 47
2.5.2 Comparison of Binary-Encoded AMOSA with NSGA-II and PAES . . . 48
2.5.3 Comparison of Real-Coded AMOSA with the Algorithm of Smith et al. and Real-Coded NSGA-II . . . 51
2.5.4 Discussion on Annealing Schedule . . . 56
2.6 Discussion and Conclusions . . . 57

3 Similarity Measures . . . 59
3.1 Introduction . . . 59
3.2 Definitions . . . 60
3.2.1 Need for Measuring Similarity . . . 60
3.3 Similarity/Dissimilarity for Binary Variables . . . 61
3.4 Distance for Nominal/Categorical Variable . . . 63
3.4.1 Method 1: Each Category Is Represented by a Single Binary Variable [278] . . . 64
3.4.2 Method 2: Each Category Is Represented by Several Binary Variables [278] . . . 64
3.5 Distance for Ordinal Variables . . . 65
3.5.1 Normalized Rank Transformation . . . 66
3.5.2 Spearman Distance . . . 67
3.5.3 Footrule Distance . . . 67
3.6 Distance for Quantitative Variables . . . 68
3.6.1 Euclidean Distance . . . 68
3.6.2 Minkowski Distance of Order λ . . . 68
3.6.3 City Block Distance . . . 69
3.6.4 Chebyshev Distance . . . 69
3.6.5 Canberra Distance . . . 70
3.6.6 Bray-Curtis Distance . . . 70
3.6.7 Angular Separation . . . 70
3.6.8 Correlation Coefficient . . . 71
3.6.9 Mahalanobis Distance . . . 72
3.7 Normalization Methods . . . 72
3.8 Summary . . . 73

4 Clustering Algorithms . . . 75
4.1 Introduction . . . 75
4.2 Preliminaries . . . 76
4.2.1 Definition of Clustering . . . 76
4.2.2 Some Clustering Techniques . . . 76

4.3 Partitional Clustering Techniques . . . 77
4.3.1 K-Means Clustering Technique . . . 77
4.3.2 K-Medoid Clustering Technique . . . 79
4.3.3 Fuzzy C-Means Clustering Algorithm . . . 79
4.4 Distribution-Based Clustering Approach . . . 80
4.5 Hierarchical Clustering Techniques . . . 82
4.6 Density-Based Clustering Techniques . . . 83
4.7 Grid-Based Clustering Techniques . . . 85
4.8 Some Recent Clustering Techniques . . . 85
4.9 Some Evolutionary Approaches to Clustering . . . 87
4.9.1 Algorithms for a Fixed Value of the Number of Clusters . . . 88
4.9.2 Algorithms with Variable Number of Clusters . . . 89
4.10 MOO and Clustering . . . 90
4.11 Summary . . . 91

5 Point Symmetry-Based Distance Measures and Their Applications to Clustering . . . 93
5.1 Introduction . . . 93
5.2 Some Existing Symmetry-Based Distance Measures . . . 94
5.3 A New Definition of the Point Symmetry Distance [27] . . . 100
5.4 Some Properties of $d_{ps}(\overline{x}, \overline{c})$ . . . 102
5.5 Kd-Tree-Based Nearest Neighbor Computation . . . 103
5.6 GAPS: The Genetic Clustering Scheme with New PS-Distance [27] . . . 104
5.6.1 Chromosome Representation and Population Initialization . . . 104
5.6.2 Fitness Computation . . . 106
5.6.3 Selection . . . 107
5.6.4 Crossover . . . 107
5.6.5 Mutation . . . 108
5.6.6 Termination . . . 109
5.6.7 Complexity Analysis . . . 109
5.7 On the Convergence Property of GAPS . . . 110
5.7.1 Preliminaries . . . 110
5.7.2 Convergence Proof . . . 112
5.8 Experimental Results of GAPS . . . 113
5.8.1 Data Sets Used . . . 113
5.8.2 Implementation Results . . . 114
5.8.3 Summary . . . 121

6 A Validity Index Based on Symmetry: Application to Satellite Image Segmentation . . . 125
6.1 Introduction . . . 125
6.2 Some Existing Cluster Validity Indices . . . 126
6.2.1 BIC Index . . . 126
6.2.2 Calinski-Harabasz Index . . . 127

6.2.3 Silhouette Index . . . 127
6.2.4 DB Index . . . 127
6.2.5 Dunn's Index . . . 128
6.2.6 Xie-Beni Index . . . 128
6.2.7 PS Index . . . 129
6.2.8 I-Index . . . 129
6.2.9 CS-Index . . . 129
6.3 *Sym*-Index: The Symmetry-Based Cluster Validity Index . . . 130
6.3.1 The Cluster Validity Measure . . . 130
6.3.2 Mathematical Justification . . . 134
6.3.3 Interaction Between the Different Components of Sym-Index . . . 136
6.4 Experimental Results . . . 140
6.4.1 Data Sets . . . 140
6.4.2 Comparative Results . . . 140
6.4.3 Analysis of Results . . . 143
6.5 Incorporating d_{ps} in Some Existing Cluster Validity Indices . . . 147
6.6 Point Symmetry-Based Cluster Validity Indices . . . 148
6.6.1 Symmetry-Based Davies-Bouldin Index (*Sym-DB* Index) . . . 148
6.6.2 Symmetry-Based Dunn's Index (*Sym-Dunn* Index) . . . 149
6.6.3 Symmetry-Based Generalized Dunn's Index (*Sym-GDunn* Index) . . . 149
6.6.4 New Symmetry Distance-Based PS-Index (*Sym-PS* Index) . . . 150
6.6.5 Symmetry-Based Xie-Beni Index (*Sym-XB* Index) . . . 151
6.6.6 Symmetry-Based FS-Index (*Sym-FS* Index) . . . 151
6.6.7 Symmetry-Based K-Index (*Sym-K* Index) . . . 151
6.6.8 Symmetry-Based SV-Index (*Sym-SV* Index) . . . 152
6.7 Experimental Results . . . 152
6.7.1 Discussion of Results . . . 152
6.8 Application to Remote Sensing Imagery . . . 155
6.8.1 Simulated Circle Image (SCI) . . . 156
6.8.2 SPOT Image of Kolkata . . . 156
6.9 Discussion and Conclusions . . . 160

7 Symmetry-Based Automatic Clustering . . . 165
7.1 Introduction . . . 165
7.2 Some Existing Genetic Algorithm-Based Automatic Clustering Techniques . . . 166
7.3 Description of VGAPS . . . 168
7.3.1 Chromosome Representation and Population Initialization . . . 168
7.3.2 Fitness Computation . . . 168
7.3.3 Genetic Operations and Terminating Criterion . . . 168

7.4 On the Convergence Property of VGAPS . . . 171
7.4.1 To Check Whether the Mutation Matrix Is Positive . . . 172
7.4.2 Conditions on Selection . . . 173
7.5 Data Sets Used and Implementation Results . . . 174
7.5.1 Data Sets Used . . . 174
7.5.2 Results and Discussions . . . 174
7.6 Extending VGAPS to Fuzzy Clustering . . . 178
7.6.1 Fuzzy Symmetry-Based Cluster Validity Index . . . 178
7.6.2 Fuzzy-VGAPS Clustering . . . 179
7.7 Implementation Results and Comparative Study . . . 182
7.7.1 Discussion of Results . . . 184
7.7.2 Application to MR Brain Image Segmentation . . . 189
7.7.3 Experimental Results . . . 190
7.8 Summary . . . 194

8 Some Line Symmetry Distance-Based Clustering Techniques . . . 197
8.1 Introduction . . . 197
8.2 The Line Symmetry-Based Distance . . . 197
8.2.1 Definition . . . 198
8.3 GALSD: The Genetic Clustering Scheme with Line Symmetry-Based Distance . . . 199
8.3.1 String Representation and Population Initialization . . . 199
8.3.2 Fitness Computation . . . 199
8.3.3 Genetic Operators . . . 200
8.4 Experimental Results . . . 200
8.4.1 Data Sets Used . . . 200
8.4.2 Discussion of Results . . . 202
8.4.3 Computation Time . . . 203
8.5 Application to Face Recognition . . . 204
8.5.1 Human Face Detection Algorithm . . . 205
8.5.2 Experimental Results . . . 207
8.6 A Generalized Line Symmetry-Based Distance and Its Application to Data Clustering . . . 209
8.7 Implementation Results . . . 210
8.7.1 Data Sets Used . . . 211
8.7.2 Discussion of Results . . . 211
8.8 Discussion and Conclusions . . . 215

9 Use of Multiobjective Optimization for Data Clustering . . . 217
9.1 Introduction . . . 217
9.2 MOPS: Multiobjective Clustering Using Point Symmetry Distance . . . 219
9.2.1 Selection of a Solution from the Archive . . . 220
9.2.2 Experimental Results . . . 220
9.3 VAMOSA: Symmetry-Based Multiobjective Clustering Technique for Automatic Evolution of Clusters . . . 221

9.3.1 Data Sets Used for Experiment 222
9.3.2 Experimental Results 223
9.4 A Generalized Automatic Clustering Algorithm in a Multiobjective Framework 227
9.4.1 *GenClustMOO*: Multiobjective Clustering Technique . . . 228
9.4.2 Subcluster Merging for Objective Function Calculation . . 232
9.5 Experimental Results 233
9.5.1 Data Sets Used 233
9.5.2 Discussion of Results 235
9.6 Discussion and Conclusions 242

References 245

Index 259