Swagatam Das, Ajith Abraham and Amit Konar

Metaheuristic Clustering

Studies in Computational Intelligence, Volume 178

Editor-in-Chief
Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Vol. 156. Dawn E. Holmes and Lakhmi C. Jain (Eds.)
Innovations in Bayesian Networks, 2008
ISBN 978-3-540-85065-6

Vol. 157. Ying-ping Chen and Meng-Hiot Lim (Eds.)
Linkage in Evolutionary Computation, 2008
ISBN 978-3-540-85067-0

Vol. 158. Marina Gavrilova (Ed.)
Generalized Voronoi Diagram: A Geometry-Based Approach to Computational Intelligence, 2009
ISBN 978-3-540-85125-7

Vol. 159. Dimitri Plemenos and Georgios Miaoulis (Eds.)
Artificial Intelligence Techniques for Computer Graphics, 2009
ISBN 978-3-540-85127-1

Vol. 160. P. Rajasekaran and Vasantha Kalyani David
Pattern Recognition using Neural and Functional Networks, 2009
ISBN 978-3-540-85129-5

Vol. 161. Francisco Baptista Pereira and Jorge Tavares (Eds.)
Bio-inspired Algorithms for the Vehicle Routing Problem, 2009
ISBN 978-3-540-85151-6

Vol. 162. Costin Badica, Giuseppe Mangioni, Vincenza Carchiolo and Dumitru Dan Burdescu (Eds.)
Intelligent Distributed Computing, Systems and Applications, 2008
ISBN 978-3-540-85256-8

Vol. 163. Pawel Delimata, Mikhail Ju. Moshkov, Andrzej Skowron and Zbigniew Suraj
Inhibitory Rules in Data Analysis, 2009
ISBN 978-3-540-85637-5

Vol. 164. Nadia Nedjah, Luiza de Macedo Mourelle, Janusz Kacprzyk, Felipe M.G. França and Alberto Ferreira de Souza (Eds.)
Intelligent Text Categorization and Clustering, 2009
ISBN 978-3-540-85643-6

Vol. 165. Djamel A. Zighed, Shusaku Tsumoto, Zbigniew W. Ras and Hakim Hacid (Eds.)
Mining Complex Data, 2009
ISBN 978-3-540-88066-0

Vol. 166. Constantinos Koutsojannis and Spiros Sirmakessis (Eds.)
Tools and Applications with Artificial Intelligence, 2009
ISBN 978-3-540-88068-4

Vol. 167. Ngoc Thanh Nguyen and Lakhmi C. Jain (Eds.)
Intelligent Agents in the Evolution of Web and Applications, 2009
ISBN 978-3-540-88070-7

Vol. 168. Andreas Tolk and Lakhmi C. Jain (Eds.)
Complex Systems in Knowledge-based Environments: Theory, Models and Applications, 2009
ISBN 978-3-540-88074-5

Vol. 169. Nadia Nedjah, Luiza de Macedo Mourelle and Janusz Kacprzyk (Eds.)
Innovative Applications in Data Mining, 2009
ISBN 978-3-540-88044-8

Vol. 170. Lakhmi C. Jain and Ngoc Thanh Nguyen (Eds.)
Knowledge Processing and Decision Making in Agent-Based Systems, 2009
ISBN 978-3-540-88048-6

Vol. 171. Chi-Keong Goh, Yew-Soon Ong and Kay Chen Tan (Eds.)
Multi-Objective Memetic Algorithms, 2009
ISBN 978-3-540-88050-9

Vol. 172. I-Hsien Ting and Hui-Ju Wu (Eds.)
Web Mining Applications in E-Commerce and E-Services, 2009
ISBN 978-3-540-88080-6

Vol. 173. Tobias Grosche
Computational Intelligence in Integrated Airline Scheduling, 2009
ISBN 978-3-540-89886-3

Vol. 174. Ajith Abraham, Rafael Falcón and Rafael Bello (Eds.)
Rough Set Theory: A True Landmark in Data Analysis, 2009
ISBN 978-3-540-89886-3

Vol. 175. Godfrey C. Onwubolu and Donald Davendra (Eds.)
Differential Evolution: A Handbook for Global Permutation-Based Combinatorial Optimization, 2009
ISBN 978-3-540-92150-9

Vol. 176. Beniamino Murgante, Giuseppe Borruso and Alessandra Lapucci (Eds.)
Geocomputation and Urban Planning, 2009
ISBN 978-3-540-89929-7

Vol. 177. Dikai Liu, Lingfeng Wang and Kay Chen Tan (Eds.)
Design and Control of Intelligent Robotic Systems, 2009
ISBN 978-3-540-89932-7

Vol. 178. Swagatam Das, Ajith Abraham and Amit Konar
Metaheuristic Clustering, 2009
ISBN 978-3-540-92172-1

Swagatam Das
Ajith Abraham
Amit Konar

Metaheuristic Clustering

Springer

Swagatam Das
Department of Electronics and
Telecommunication Engineering (ETCE)
Jadavpur University
Raja S. C. Mullick Road
Jadavpur, Calcutta - 700032
India

Amit Konar
Department of Electronics and
Telecommunication Engineering (ETCE)
Jadavpur University
Raja S. C. Mullick Road
Jadavpur, Calcutta - 700032
India

Ajith Abraham
Norwegian Center of Excellence
Center of Excellence for Quantifiable
Quality of Service
Norwegian University of Science
and Technology
O.S. Bragstads plass 2E
NO-7491 Trondheim
Norway

ISBN 978-3-540-92172-1 e-ISBN 978-3-540-93964-1

DOI 10.1007/978-3-540-93964-1

Studies in Computational Intelligence ISSN 1860949X

Library of Congress Control Number: 2008942042

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Foreword

Indisputably in the oceans of data surrounding us, clustering has gained a central position as a conceptual and algorithmic framework that helps the user make sense of data and reveal some underlying structure that is hidden behind overwhelming streams of numbers.

There are thousands of clustering techniques one can encounter in the literature. We will be seeing far more methods arising over the passage of time. Just the recent search using Google Scholar (dated November 18, 2008) has returned about 1,510,000 hits. This number speaks to the dynamics and omnipresence of the clustering paradigm and its numerous applications.

What we have started seeing more vividly are the two fundamental clustering challenges one has to deal with in an effective manner. First, it becomes apparent that clustering is a processes guided by several objectives (objective functions) rather than a single and somewhat isolated goal. This has led us to the concept of multiobjective clustering. Likewise we have started to realize that to make clustering more user-centric, one needs to fully accommodate some prior domain knowledge and this line of pursuit has resulted in a so-called knowledge-based clustering. Second, there is an acute need for optimization tools that are of *global* nature and in this way may help realize a comprehensive search which is of structural as well as of parametric character. The role evolutionary computing has been already acknowledged in this particular context yet there is a large unexplored research territory where we can anticipate a great deal of interesting findings.

The treatise authored by Professors Das, Abraham, and Konar tackles a very fundamental and practically highly relevant research topic: how to make clustering more efficient and very much in rapport with the reality of multifaceted data and diversified needs of the end users. The notion of metaheuristics used in the title of the book is very much reflective of its very content.

The reader is carefully navigated through the efficacies of clustering, evolutionary optimization and a hybridization of the both. The exposure of the material is lucid. Quite complicated concepts are presented in a clear and convincing way which can be attributed to the expertise of Professors Das, Abraham, and Konar. While Evolutionary Computing has been recognized as a viable optimization platform, it has been noted quite early that a number of well-known techniques such as e.g., Genetic Algorithms and Evolutionary Algorithms come with a substantial computational overhead which becomes difficult to accept in case of problems of higher dimensionality. From this standpoint, the alternative of Differential Evolution (DE) pursued by the authors is indeed a very fortunate choice.

In the exposure of the material, the authors have achieved a sound balance between the theory and practice. We witness a wealth of fundamental and far reaching results, especially when it comes to the analysis of the dynamics of Differential Evolution. We can appreciate the applied facets of the monograph where the algorithmic setting established in the book stresses applicability or leads directly to interesting and well-rounded applications in data analysis.

All in all, this is not only a very timely and badly needed volume but also an outstanding, comprehensive and authoritative treatise of the important subject of metaheuristics clustering.

Professor, Canada Research Chair, IEEE Fellow
University of Alberta, Canada
November 2008

Witold Pedrycz

Preface

Cluster analysis means the organization of an unlabeled collection of objects (or *patterns*) into separate groups based on their similarity. Each valid group, called a 'cluster', should consist of objects that are similar among themselves and dissimilar to objects of other groups. As human beings, we resort to clustering as one of our most primitive mental activities for organizing the data we receive every day, so that we may draw important conclusions from them. It is well nigh impossible to process every piece of such data as a single entity. Thus, humans tend to categorize entities (i.e. objects, persons, events) into clusters. Each cluster is then characterized by the common attributes (features) of the entities that belong to that cluster.

Human beings possess the natural ability of clustering objects. Given a box full of marbles of four different colors say red, green, blue, and yellow, even a child may separate these marbles into four clusters based on their colors. However, making a computer solve this type of problems is quite difficult and demands the attention of computer scientists and engineers all over the world till date. The major hurdle in this task is that the functioning of the brain is much less understood. The mechanisms, with which it stores huge amounts of information, processes them at lightning speeds and infers meaningful rules, and retrieves information as and when necessary have till now eluded the scientists. A question that naturally comes up is: what is the point in making a computer perform clustering when people can do this so easily? The answer is far from trivial. The most important characteristic of this information age is the abundance of data. Advances in computer technology, in particular the Internet, have led to what some people call "data explosion": the amount of data available to any person has increased so much that it is more than he or she can handle. In reality the amount of data is vast and in addition, each data item (an abstraction of a real-life object) may be characterized by a large number of attributes (or *features*), which are based on certain measurements taken on the real-life objects and may be numerical or non-numerical. Mathematically we may think of a mapping of each data item into a point in the multi-dimensional feature space (each dimension corresponding to one feature) that is beyond our perception when number of features exceed just 3. Thus it is nearly impossible for human beings to partition tens of thousands of data items, each coming with several features (usually much greater than 3), into meaningful clusters within a short interval of time. Nonetheless, the task is of paramount importance for organizing and summarizing huge piles of data and discovering useful knowledge from them. So, can we devise some means to generalize to arbitrary dimensions of what humans perceive in two or three dimensions, as densely connected "patches" or "clouds" within data space? The entire research on cluster analysis may be considered as an effort to find satisfactory answers to this fundamental question.

The task of computerized data clustering has been approached from diverse domains of knowledge like graph theory, statistics (multivariate analysis), artificial neural networks, fuzzy set theory, and so on. One of the most popular approaches in this direction has been the formulation of clustering as an optimization problem, where the best partitioning of a given dataset is achieved by minimizing/maximizing one (single-objective clustering) or more (multi-objective clustering) objective functions. The objective functions are usually formed capturing certain statistical-mathematical relationship among the individual data items and the candidate set of representatives of each cluster (also known as cluster-centroids). The clusters are either hard, that is each sample point is unequivocally assigned to a cluster and is considered to bear no similarity to members of other clusters, or fuzzy, in which case a membership function expresses the degree of belongingness of a data item to each cluster.

Most of the classical optimization-based clustering algorithms (including the celebrated hard c-means and fuzzy c-means algorithms) rely on local search techniques (like iterative function optimization, Lagrange's multiplier, Picard's iterations etc.) for optimizing the clustering criterion functions. The local search methods, however, suffer from two great disadvantages. Firstly they are prone to getting trapped in some local optima of the multi-dimensional and usually multimodal landscape of the objective function. Secondly performances of these methods are usually very sensitive to the initial values of the search variables.

Although many respected texts of pattern recognition describe clustering as an unsupervised learning method, most of the traditional clustering algorithms require a prior specification of the number of clusters in the data for guiding the partitioning process, thus making it not completely unsupervised. On the other hand, in many practical situations, it is impossible to provide even an estimation of the number of naturally occurring clusters in a previously unhandled dataset. For example, while attempting to classify a large database of handwritten characters in an unknown language; it is not possible to determine the correct number of distinct letters beforehand. Again, while clustering a set of documents arising from the query to a search engine, the number of classes can change for each set of documents that result from an interaction with the search engine. Data mining tools that predict future trends and behaviors for allowing businesses to make proactive and knowledge-driven decisions, demand fast and fully automatic clustering of very large datasets with minimal or no user intervention. Thus it is evident that the complexity of the data analysis tasks in recent times has posed severe challenges before the classical clustering techniques.

Starting from early 1960s, a keen observation of the underlying relation between optimization and biological evolution has led to the development of an important paradigm of computational intelligence – the evolutionary computing

(EC) - for performing very complex search and optimization. Evolutionary computing harnesses the power of natural selection to turn computers into automatic optimisation and design tools. This volume investigates the application of a recently developed evolutionary computing algorithm, well-known as the Differential Evolution (DE), to develop robust, fast and fully automatic clustering techniques that can circumvent the problems with several classical clustering schemes, as illustrated earlier.

Since its advent in 1995, DE has drawn the attention of the practitioners in optimization all over the globe due to its high degrees of robustness, convergence speed, and accuracy in real parameter optimization problems. A very simple algorithm to code with so few (typically 3 in classical DE) adjustable control-parameters, DE has been shown to outperform several veteran members of the EC family like the Genetic Algorithms (GA), Evolutionary Strategies (ES), and Memetic Algorithms (MA) over both benchmark and real-world problems. Unlike GAs, however, the application of DE to clustering problems has not been much investigated.

In this Volume, we illustrate the performance of DE, when applied to both single and multi-objective clustering problems, where the number of clusters is not known beforehand and must be determined on the run. We first undertake a statistical analysis of the search operators and the convergence behaviour of DE near an isolated equilibrium point in the search space. Taking a cue from the analysis mentioned earlier, we propose a few parameter automation strategies that improve the performance of classical DE without imposing any serious additional computational burden. Next we develop a new DE-based crisp clustering algorithm, which can not only correctly partition the data in appropriate clusters but also find the optimal number of clusters automatically. The proposed algorithm incorporates a new real-coded scheme for search variable representation that makes room for several possible choices of the number of clusters in the dataset. An extensive comparison with several other state-of-the-art clustering algorithms over many synthetic and real-life datasets reflects the statistically superior performance of the proposed scheme in terms of final accuracy, speed and robustness. We also applied the proposed clustering method to an interesting problem of automatic image pixel clustering and land cover study from satellite images. The proposed clustering technique is next extended to the fuzzy clustering in kernel induced feature space, for tackling more complex clusters, which are linearly non-separable and overlapping in nature. A new DE-variant with balanced exploration and exploitation abilities has been proposed for optimizing the clustering objectives in higher dimensional kernel space. The new DE variant is shown to perform better than the classical DE and many other recently developed algorithms for kernel-based clustering in a statistically significant fashion. Finally the Volume compares four most recently proposed multi-objective (MO) variants of the DE with two other state-of-the-art MO clustering methods over ten datasets of widely varying ranges of complexity. A novel framework for multi-objective automatic clustering is proposed for the multi-objective DE variants, one or more of which is always seen to find statistically better result than their other state-of-the-art contestants. An interesting application of the multi-objective DE based clustering to gene expression data of yeast is also investigated in this context.

The most important characteristics of the algorithms proposed in the Volume are:

1) They can optimally cluster a previously unhandled dataset (with numerical features) into correct number of clusters through one shot of optimization. As opposed to the classical local search based optimization techniques, they are able to locate the global optima of the multi-modal landscape of clustering objective function quickly.

2) Their computational speeds are faster than those of the clustering techniques based on other evolutionary and swarm intelligence algorithms.

3) They are fairly robust against different initial conditions and can produce nearly similar results (with small standard deviations) over repeated runs,

4) Owing to the characteristics of DE, they have very few control parameters and can yield good final accuracy over a large variety of clustering problems with minimal or no hand tuning.

The Volume is organized in 7 Chapters. The first Chapter presents a detailed review of the evolutionary clustering algorithms. The Chapter begins with a formal overview of the clustering problem, similarity and dissimilarity measures between patterns and the various methods of clustering. It then addresses a few classical clustering algorithms, pertinent to the present work. Next the Chapter discusses the relevance of evolutionary computing techniques in pattern clustering and outlines the most promising evolutionary clustering methods. The Chapter ends with a discussion on automatic clustering problem, which remains largely unsolved by most of the traditional clustering algorithms.

Chapter 2 presents a conceptual outline of the DE algorithm in sufficient details. It then reviews six prominent variants of DE, including DE with trigonometric mutation, DE with arithmetic recombination, DE/rand/1/either-or, self-adaptive DE, opposition-based DE, binary DE, DE with adaptive local search and finally a new family of DE-variants based on neighborhood-based mutation. An interesting algorithm resulting from the synergy of DE with an important swarm intelligence algorithm, well known as Particle Swarm Optimization (PSO) is also addressed in the Chapter.

Chapter 3 investigates the dynamics of a canonical DE algorithm with DE/rand/1 type mutation and binomial crossover. The Chapter develops a simple mathematical model of the underlying evolutionary dynamics of a one-dimensional DE-population. The model relates the search mechanism of DE to that of the classical gradient descent search. The stability and convergence-behavior of the proposed dynamics is then analyzed with the help of Lyapunov's stability theorems. The mathematical model, developed in this Chapter, provides important insights into the search mechanism of DE in a near neighborhood of an isolated optimum. The Chapter also presents empirical simulation results over simple objective functions to validate the theoretical analyses.

Chapter 4 describes a DE-based algorithm for the automatic crisp clustering of large unlabeled datasets. In contrast to most of the existing clustering techniques, the algorithm, proposed by the chapter, requires no prior knowledge of the data to be classified. Rather, it determines the optimal number of clusters in the data 'on the run'. Superiority of the new method has been demonstrated by comparing it with two recently developed partitional clustering techniques and one popular hierarchical clustering algorithm. The partitional algorithms are based on two powerful optimization algorithms well-known as the Genetic Algorithm (GA) and the Particle Swarm Optimization (PSO). The Chapter also reports an interesting practical application of the proposed method to automatic segmentation of gray-scale images in their intensity space.

Chapter 5 extends the work reported in Chapter 4 to the fuzzy clustering of complex and linearly non-separable datasets in kernel-induced feature space. The proposed method is based on a modified version of the classical DE algorithm, which uses a novel neighborhood-based mutation strategy. It also employs a kernel-induced similarity measure instead of the conventional sum-of-squares distance. Use of the kernel function makes it possible to cluster data that is linearly non-separable in the original input space into homogeneous groups in a transformed high-dimensional feature space. The vector representation scheme remains identical to that described in Chapter 4. The performance of the proposed method has been extensively compared with a few state of the art clustering techniques over a test-suite of several artificial and real life datasets. Based on experiment results, the Chapter also provides some empirical guidelines for selecting the suitable parameters of the modified DE algorithm

Chapter 6 considers the task of fuzzy clustering in a multi-objective optimization (MO) framework. It compares the performances of four recently developed multi-objective variants of DE over the fuzzy clustering problem, where two conflicting fuzzy validity indices are simultaneously optimized. The resultant Pareto optimal set of solutions from each algorithm consists of a number of non-dominated solutions, from which the user can choose the most promising ones according to the problem specifications. A real-coded representation of the search variables, proposed in Chapter 4, is used for DE to accommodate variable number of cluster centers. The performances of four DE variants have also been contrasted to that of two most well-known schemes of MO clustering namely the NSGA II (Non Dominated Sorting GA) and MOCK (Multi-Objective Clustering with an unknown number of clusters K). Experimental results over four artificial and four real life datasets (including a gene expression dataset of yeast sporulation) of varying range of complexities indicates that DE holds immense promise as a candidate algorithm for devising MO clustering schemes.

Finally Chapter 7 concludes the Volume with a discussion on the possible extensions of the works undertaken and projects a possible direction of future research.

We are very much grateful to our friends and colleagues for reviewing the different parts of the manuscript and for providing us valuable feedback. The first author would like to thank the swarm of his undergraduate and graduate students including Mr. Sambarta Dasgupta, Mr. Arijit Biswas, and Ms. Sudeshna Sil for helping him in performing several simulation experiments reported in the book.

Authors would like to thank Dr. Thomas Ditzinger, Springer Engineering Inhouse Editor, Studies in Computational Intelligence Series, Professor Janusz Kacprzyk (Editor-in-Chief, Springer Studies in Computational Intelligence Series) and Ms. Heather King (Editorial Assistant, Springer Verlag, Heidelberg) for the editorial assistance and excellent cooperative collaboration to produce this important scientific work. We hope that the reader will share our excitement to present this volume on 'Metaheuristic Clustering' and will find it useful.

Swagatam Das, Ajith Abraham* and Amit Konar

Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata 700032, India
**Center of Excellence for Quantifiable Quality of Service (Q2S), Norwegian University of Science and Technology, Trondheim, Norway*

Contents