Ajith Abraham, Aboul-Ella Hassanien, and
André Ponce de Leon F. de Carvalho (Eds.)

# Foundations of Computational Intelligence Volume 4

# Studies in Computational Intelligence, Volume 204

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

Vol. 181. Georgios Miaoulis and Dimitri Plemenos (Eds.)
*Intelligent Scene Modelling Information Systems,* 2009
ISBN 978-3-540-92901-7

Vol. 182. Andrzej Bargiela and Witold Pedrycz (Eds.)
*Human-Centric Information Processing Through Granular Modelling,* 2009
ISBN 978-3-540-92915-4

Vol. 183. Marco A.C. Pacheco and Marley M.B.R. Vellasco (Eds.)
*Intelligent Systems in Oil Field Development under Uncertainty,* 2009
ISBN 978-3-540-92999-4

Vol. 184. Ljupco Kocarev, Zbigniew Galias and Shiguo Lian (Eds.)
*Intelligent Computing Based on Chaos,* 2009
ISBN 978-3-540-95971-7

Vol. 185. Anthony Brabazon and Michael O'Neill (Eds.)
*Natural Computing in Computational Finance,* 2009
ISBN 978-3-540-95973-1

Vol. 186. Chi-Keong Goh and Kay Chen Tan
*Evolutionary Multi-objective Optimization in Uncertain Environments,* 2009
ISBN 978-3-540-95975-5

Vol. 187. Mitsuo Gen, David Green, Osamu Katai, Bob McKay, Akira Namatame, Ruhul A. Sarker and Byoung-Tak Zhang (Eds.)
*Intelligent and Evolutionary Systems,* 2009
ISBN 978-3-540-95977-9

Vol. 188. Agustín Gutiérrez and Santiago Marco (Eds.)
*Biologically Inspired Signal Processing for Chemical Sensing,* 2009
ISBN 978-3-642-00175-8

Vol. 189. Sally McClean, Peter Millard, Elia El-Darzi and Chris Nugent (Eds.)
*Intelligent Patient Management,* 2009
ISBN 978-3-642-00178-9

Vol. 190. K.R. Venugopal, K.G. Srinivasa and L.M. Patnaik
*Soft Computing for Data Mining Applications,* 2009
ISBN 978-3-642-00192-5

Vol. 191. Zong Woo Geem (Ed.)
*Music-Inspired Harmony Search Algorithm,* 2009
ISBN 978-3-642-00184-0

Vol. 192. Agus Budiyono, Bambang Riyanto and Endra Joelianto (Eds.)
*Intelligent Unmanned Systems: Theory and Applications,* 2009
ISBN 978-3-642-00263-2

Vol. 193. Raymond Chiong (Ed.)
*Nature-Inspired Algorithms for Optimisation,* 2009
ISBN 978-3-642-00266-3

Vol. 194. Ian Dempsey, Michael O'Neill and Anthony Brabazon (Eds.)
*Foundations in Grammatical Evolution for Dynamic Environments,* 2009
ISBN 978-3-642-00313-4

Vol. 195. Vivek Bannore and Leszek Swierkowski
*Iterative-Interpolation Super-Resolution Image Reconstruction:*
*A Computationally Efficient Technique,* 2009
ISBN 978-3-642-00384-4

Vol. 196. Valentina Emilia Balas, János Fodor and Annamária R. Várkonyi-Kóczy (Eds.)
*Soft Computing Based Modeling in Intelligent Systems,* 2009
ISBN 978-3-642-00447-6

Vol. 197. Mauro Birattari
*Tuning Metaheuristics,* 2009
ISBN 978-3-642-00482-7

Vol. 198. Efrén Mezura-Montes (Ed.)
*Constraint-Handling in Evolutionary Optimization,* 2009
ISBN 978-3-642-00618-0

Vol. 199. Kazumi Nakamatsu, Gloria Phillips-Wren, Lakhmi C. Jain, and Robert J. Howlett (Eds.)
*New Advances in Intelligent Decision Technologies,* 2009
ISBN 978-3-642-00908-2

Vol. 200. Dimitri Plemenos and Georgios Miaoulis
*Visual Complexity and Intelligent Computer Graphics Techniques Enhancements,* 2009
ISBN 978-3-642-01258-7

Vol. 201. Aboul-Ella Hassanien, Ajith Abraham, Athanasios V. Vasilakos, and Witold Pedrycz (Eds.)
*Foundations of Computational Intelligence Volume 1,* 2009
ISBN 978-3-642-01081-1

Vol. 202. Aboul-Ella Hassanien, Ajith Abraham, and Francisco Herrera (Eds.)
*Foundations of Computational Intelligence Volume 2,* 2009
ISBN 978-3-642-01532-8

Vol. 203. Ajith Abraham, Aboul-Ella Hassanien, Patrick Siarry, and Andries Engelbrecht (Eds.)
*Foundations of Computational Intelligence Volume 3,* 2009
ISBN 978-3-642-01084-2

Vol. 204. Ajith Abraham, Aboul-Ella Hassanien, and André Ponce de Leon F. de Carvalho (Eds.)
*Foundations of Computational Intelligence Volume 4,* 2009
ISBN 978-3-642-01087-3

Ajith Abraham, Aboul-Ella Hassanien, and
André Ponce de Leon F. de Carvalho (Eds.)

# Foundations of Computational Intelligence Volume 4

## Bio-Inspired Data Mining

Springer

Dr. Ajith Abraham
Machine Intelligence Research Labs
(MIR Labs)
Scientific Network for Innovation
and Research Excellence
P.O. Box 2259 Auburn,
Washington 98071-2259
USA
E-mail: ajith.abraham@ieee.org
http://www.mirlabs.org
http://www.softcomputing.net

Prof. Aboul-Ella Hassanien
College of Business Administration
Quantitative and Information System
Department
Kuwait University
P.O. Box 5486
Safat, 13055
Kuwait
E-mail: abo@cba.edu.kw

Prof. André Ponce de Leon F. de
Carvalho
Department of Computer Science
University of São Paulo
SCE - ICMSC - USP
Caixa Postal 668
13560-970 Sao Carlos, SP
Brazil
E-mail: andre@icmc.usp.br

ISBN 978-3-642-01087-3 e-ISBN 978-3-642-01088-0

DOI 10.1007/978-3-642-01088-0

Studies in Computational Intelligence ISSN 1860949X

Library of Congress Control Number: Applied for

*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

# Preface

## Foundations of Computational Intelligence

### Volume 4: Bio-Inspired Data Mining Theoretical Foundations and Applications

Recent advances in the computing and electronics technology, particularly in sensor devices, databases and distributed systems, are leading to an exponential growth in the amount of data stored in databases. It has been estimated that this amount doubles every 20 years. For some applications, this increase is even steeper. Databases storing DNA sequence, for example, are doubling their size every 10 months. This growth is occurring in several applications areas besides bioinformatics, like financial transactions, government data, environmental monitoring, satellite and medical images, security data and web. As large organizations recognize the high value of data stored in their databases and the importance of their data collection to support decision-making, there is a clear demand for sophisticated Data Mining tools. Data mining tools play a key role in the extraction of useful knowledge from databases. They can be used either to confirm a particular hypothesis or to automatically find patterns. In the second case, which is related to this book, the goal may be either to describe the main patterns present in dataset, what is known as descriptive Data Mining or to find patterns able to predict behaviour of specific attributes or features, known as predictive Data Mining. While the first goal is associated with tasks like clustering, summarization and association, the second is found in classification and regression problems.

Computational tools or solutions based on intelligent systems are being used with great success in Data Mining applications. Nature has been very successful in providing clever and efficient solutions to different sorts of challenges and problems posed to organisms by ever-changing and unpredictable environments. It is easy to observe that strong scientific advances have been made when issues from different research areas are integrated. A particularly fertile integration combines biology and computing. Computational tools inspired on biological process can be found in a large number of applications. One of these applications is Data Mining, where computing techniques inspired on nervous systems; swarms, genetics, natural selection, immune systems and molecular biology have provided new efficient alternatives to obtain new, valid, meaningful and useful patterns in large datasets.

This Volume comprises of 16 chapters, including an overview chapter, providing an up-to-date and state-of-the research on the application of Bio-inspired techniques for Data Mining.

The book is divided into 5 parts:

Part-I: Bio-inspired approaches in sequence and data streams  
Part-II: Bio-inspired approaches in classification problem  
Part-III: Evolutionary Fuzzy and Swarm in Clustering Problems  
Part-IV: Genetic and evolutionary algorithms in Bioinformatics  
Part-V: Bio-inspired approaches in information retrieval and visualization

Part I on **Bio-inspired approaches in sequence and data streams** contains four chapters that describe several approaches bio-inspired approaches in sequence and data streams.

**In Chapter 1**, "Adaptive and Self-adaptive Techniques for Evolutionary Forecasting Applications Set in Dynamic and Uncertain Environments," "Adaptive and Self-adaptive Techniques for Evolutionary Forecasting Applications Set in Dynamic and Uncertain Environments," *Wagner and Michalewicz*, present recent studies on evolutionary forecasting, showing how adaptive and self-adaptive algorithms can be efficiently used for the analysis and prediction of dynamic time series. In these time series, the data-generating process can change with time, which is the case in real world time series. Authors point out that previous works usually do not consider this dynamic behaviour and they propose a self-adaptive windowing technique based on Genetic Programming.

**Chapter 2,** "Sequence Pattern Mining," by *Zhou, Shimada, Mabu and Hirasawa*, presents the main aspects of mining datasets where the data assume the format of sequences. Sequence datasets are found in several application areas, like bioinformatics, web and system use logs. The authors analyse the different nature of sequences and models found in the literature for sequence mining. They also propose a new model for sequence mining based on Evolutionary Algorithms.

**In Chapter 3**, "Growing Self-Organizing Map for Online Continuous Clustering," written by *Smith and Alahakoon*, the authors propose a hybrid intelligent learning algorithm based on Self-Organising Maps. The proposed algorithm combines a Growing Self-Organising Map with a Cellular Probabilistic Self-Organising Map. The authors illustrate the advantages of using their algorithm for dynamic clustering in data stream applications. For such, they show the results obtained in experiments using artificial and real world data.

**Chapter 4,** "Synthesis of Spatio-Temporal Models by the Evolution of Non-Uniform Cellular Automata," by *Romano, Villanueva, Zanetti and Von Zuben*, deals with the definition of transition rules for each cell in cellular automata. Motivated by the fact that the search space is very large, the authors employ Evolutionary Algorithms to optimize the definition of the set of transition rules in cellular automata. In the experiments performed, the authors considered one and two-dimensional regular lattices.

Part II on **Bio-inspired approaches in classification problems** contains three chapters discussing many approaches in classification problem.

**Chapter 5**, "Genetic Selection Algorithm and Cloning for Data Mining with GMDH Method," by *Jirina and Jirina, Jr*, is related to Artificial Immune Systems. In this chapter, the authors modify the well known GMDH MIA (Group Method Data Handling Multilayer Iterative Algorithm) neural networks model by employing a selection operation to select the parents of a new neuron. The cloning takes place by small modifications in the parameters present in the copies of the best neuron. The classification accuracy of the new model is compared with the previous models and four other classification techniques using several datasets.

**In Chapter 6,** "Inducing Relational Fuzzy Classification Rules by means of Cooperative Co-evolution," by *Akbarzadeh, Sadeghian and dos Santos*, the induction of fuzzy classification rules using Evolutionary Algorithms is investigated. The Evolutionary Algorithm employs two separate populations. The first population has fuzzy classification rules and is evolved by Genetic Programming. The second population is composed by definitions for the membership function and is evolved by a mutation-based Evolutionary Algorithm. Relational operators are fuzzified by evolutionary methods. The proposed approach is experimentally evaluated and compared with some other evolutionary approaches.

A new Evolutionary Algorithm able to evolve decision trees is presented in **Chapter 7**, "Post-processing Evolved Decision Trees," authored by Johansson, *König, Löfström, Sönströd and Niklasson*. The proposed algorithm iteratively builds a Decision Tree by progressively including new nodes in order to improve the tree accuracy for the training set. In the experiments performed using 22 datasets, Decision Trees have been induced either directly from a training set or from the Neural Networks ensembles.

**Fuzzy and swarm in clustering problems** is the third Part of the book. It contains two chapters discussing the issues of clustering using Evolutionary Fuzzy and Swarm bio-inspired approaches.

**In Chapter 8,** "Evolutionary Fuzzy Clustering: An Overview and Efficiency Issues," by *Horta, Naldi, Campello, Hruschka and de Carvalho*, the authors, after discussing the importance of clustering techniques for data Mining, and presenting a brief description of hybrid clustering techniques, describe the Evolutionary Algorithm for Clustering (EAC) algorithm. They show that EAC provides a efficient combination of Fuzzy clustering and Evolutionary algorithms. After presenting the main features of EAC, the algorithm is experimentally evaluated.

**In Chapter 9**, "Stability-based Model Order Selection for Clustering Using Multiple Cooperative Swarms," by *Ahmadi, Karray and Kamel*, the authors propose a clustering algorithm based on the cooperative work of a multiple swarms. They also investigate a stability analysis model able to define the number of clusters, also known as model order selection, in a given dataset. The Multiple Cooperative Swarm clustering has its performance compared with other clustering algorithms in four datasets. Different clustering validation indexes are used in these comparisons.

**Genetic and evolutionary algorithms in Bioinformatics** are the fourth part in this book. It contains three chapters discussing some Bio-inspired approach in bioinformatics applications.

Another interesting application of Data Mining in Bioinformatics is described in **Chapter 10,** "Data-mining Protein Structure by Clustering, Segmentation and Evolutionary Algorithms," by *Lexa, Snásel and Zelinka*. After a brief introduction to Bioinformatics, the authors discuss how Evolutionary Algorithms can be used to solve problems from Bioinformatics. Later, the authors describe how clustering techniques can group protein fragments and how short fragments can be combined to obtain a larger segment and therefore be able to infer higher level functions for a protein.

**Chapter 11,** "A Clustering Genetic Algorithm for Genomic Data Mining," by *Tapia, Morett and Vallejo*, proposes a new framework based on clustering for the reconstruction of functional modules of proteins. Authors formulate the problem of protein-protein interactions as a multi-objective optimization problem. The framework is evaluated for the analysis of phylogenetic profiles. After presenting the main features of the evolutionary-based clustering algorithm investigated in this chapter, the authors provide a set of experimental evaluations on the prediction of protein-protein functional interactions from different sources of genomic data.

**Chapter 12,** "Detection of Remote Protein Homologs Using Social Programming," by *Ramstein, Beaume and Jacques*, covers an important issue in Bioinformatics, the identification of the function of new, unknown, proteins by looking for the function of homologous proteins. For this application, the authors propose the use of Social Programming, particularly, Grammatical Swarms. In the proposed approach, Support Vector machines are use to identify remote homologs. Experiments are carried out using protein sequences extracted from the SCOP database.

The final Part of the book deals with the **Bio-inspired approaches in information retrieval and visualization.** It contains four chapters, which discusses the Information Retrieval using bio-inspired approaches including the optimizing and clustering information retrieval as well as mining network traffic data.

**Chapter 13,** "Optimizing Information Retrieval Using Evolutionary Algorithms and Fuzzy Inference System," by *Snásel, Abraham, Owais, Platos and Krömer*, investigates the use of two models for information retrieval, one based on a evolutionary algorithms and crisp membership function for document terms, crisp Information Retrieval (IR) framework, named BRIM (Boolean Information Retrieval Model), and another version that combines evolutionary algorithms and fuzzy systems, using fuzzy membership functions, fuzzy IR framework, named EBIRM (Extended BIRM). Experiments compare these two frameworks using different scenarios.

**In Chapter 14,** "**Web Data Clustering,**" by *Húsek, Pokorný, Řezanková and Snášel*, the authors show the benefits of using clustering algorithms for information retrieval, particularly for analysing information from the web. After presenting the fundamentals of cluster analysis, with emphasis on connectionist-based algorithms, they present several applications of clustering in the Web environment.

**In Chapter 15,** "Efficient Construction of Image Feature Extraction Programs by Using Linear Genetic Programming with Fitness Retrieval and Intermediate-result Caching," by *Watchareeruetai, Matsumoto, Takeuchi, Kudo and Ohnishi*, the authors illustrate how bio-inspired algorithms can be used to evolve programs for feature extraction. In their approach, a variation of Linear Genetic Programming uses a population of feature extraction programs, derived from basic image processing operations. The authors show that the computational efficiency is improved by storing intermediate results. The computational efficiency of this approach is assessed by several experiments.

**Chapter 16,** “Mining Network Traffic Data for Attacks through MOVICAB-IDS,” by *Herrero and Corchado* describe an Intrusion Detection System (IDS) called MOVICAB-IDS (MObile VIsualization Connectionist Agent-Based IDS). This system is based on a dynamic multiagent architecture combining case-base reasoning and an unsupervised neural projection model to visualize and analyze the flow of network traffic data. To illustrate the performance of the described IDS, it has been tested in different domains containing several interesting attacks and anomalous situations.

We are very much grateful to the authors of this volume and to the reviewers for their great efforts by reviewing and providing interesting feedback to authors of the chapter. The editors 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 joy and will find it useful!

December 2008

Ajith Abraham Trondheim, Norway
Aboul Ella Hassanien, Cairo University
André Ponce de Leon F. de Carvalho,
Sao Carlos, SP, Brazil

# Contents

## Part I: Bio-Inspired Approaches in Sequence and Data Streams

## Part II: Bio-Inspired Approaches in Classification Problem