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

Foundations of Computational Intelligence Volume 6

Studies in Computational Intelligence, Volume 206

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. 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. 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

Vol. 205. Ajith Abraham, Aboul-Ella Hassanien, and Václav Snášel (Eds.)
Foundations of Computational Intelligence Volume 5, 2009
ISBN 978-3-642-01535-9

Vol. 206. Ajith Abraham, Aboul-Ella Hassanien, André Ponce de Leon F. de Carvalho, and Václav Snášel (Eds.)
Foundations of Computational Intelligence Volume 6, 2009
ISBN 978-3-642-01090-3

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

Foundations of Computational Intelligence Volume 6

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

Václav Snášel
Technical University Ostrava
Dept. Computer Science
Tr. 17. Listopadu 15
708 33 Ostrava
Czech Republic
E-mail: vaclav.snasel@vsb.cz

ISBN 978-3-642-01090-3 e-ISBN 978-3-642-01091-0

DOI 10.1007/978-3-642-01091-0

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 6: Data Mining: Theoretical Foundations and Applications

Finding information hidden in data is as theoretically difficult as it is practically important. With the objective of discovering unknown patterns from data, the methodologies of data mining were derived from statistics, machine learning, and artificial intelligence, and are being used successfully in application areas such as bioinformatics, business, health care, banking, retail, and many others. Advanced representation schemes and computational intelligence techniques such as rough sets, neural networks; decision trees; fuzzy logic; evolutionary algorithms; artificial immune systems; swarm intelligence; reinforcement learning, association rule mining, Web intelligence paradigms etc. have proved valuable when they are applied to Data Mining problems. Computational tools or solutions based on intelligent systems are being used with great success in Data Mining applications. It is also observed that strong scientific advances have been made when issues from different research areas are integrated.

This Volume comprises of 15 chapters including an overview chapter providing an up-to-date and state-of-the research on the applications of Computational Intelligence techniques for Data Mining.

The book is divided into 3 parts:

Part-I: Data Click Streams and Temporal Data Mining
Part-II: Text and Rule Mining
Part-III: Applications

Part I on Data Click Streams and Temporal Data Mining contains four chapters that describe several approaches in Data Click Streams and Temporal Data Mining.

Hannah and Thangavel in Chapter 1, “Mining and Analysis of Clickstream Patterns” propose a Multi-Pass CSD-Means algorithm for clustering web clickstream patterns from web access logs of Microsoft Web site, which is available in the

UCI repository. Using these algorithms can make a meaningful contribution for the clustering analysis of web logs. This algorithm estimates the optimum number of clusters automatically and is capable of manipulating efficiently very large data sets.

In Chapter 2, "An Overview on Mining Data Streams" Gama and Rodrigues discuss the main challenges and issues when learning from data streams. Authors illustrate the most relevant issues in knowledge discovery from data streams: incremental learning, cost-performance management, change detection, and novelty detection. Illustrative algorithms are presented for these learning tasks, and a real-world application illustrating the advantages of stream processing.

Chapter 3, "Data Stream Mining Using Granularity-based Approach" by Gaber presents novel approach to solving the problem of mining data streams in resource constrained environments that are the typical representatives for data stream sources and processing units in many applications. The proposed Algorithm Output Granularity (AOG) approach adapts the output rate of a data-mining algorithm according to available resources and data rate. AOG has been formalized and the main concepts and definitions have been introduced followed by a rigorous discussion.

In Chapter 4, "Time Granularity in Temporal Data Mining" Cotofrei1 and Stoffel, illustrate formalism for a specific temporal data mining task: the discovery of knowledge, represented in the form of general Horn clauses, inferred from databases with a temporal dimension. The theoretical framework proposed by the authors is based on first-order temporal logic, which permits to define the main notions (event, temporal rule, and constraint) in a formal way. The concept of a consistent linear time structure allows us to introduce the notions of general interpretation, of support and of confidence, the lasts two measure being the expression of the two similar concepts used in data mining.

In Chapter 5, "Mining User Preference Model from Utterances" Takama1 and Muto, introduce a method for mining user preference model from user's behavior including utterances. The authors also analyze utterance of user. The proposed approach is applied to TV program recommendation, in which the log of watched TV programs as well as utterances while watching TV is collected. First, user's interest in a TV program is estimated based on fuzzy inference, of which inputs are watching time, utterance frequency, and contents of utterances obtained by sentiment analysis. Then, user profile is generated by identifying features common to user's favorite TV programs.

Part II on Text and Rule Mining contains six chapters discussing many approaches in text and rule mining problem.

One of the most relevant today's problems called information overloading has increased the necessity of more sophisticated and powerful information compression methods or summarizers. Chapter 6, "Text Summarization: an Old Challenge and New Approaches" by Steinberger and Karel, firstly introduce taxonomy of summarization methods and an overview of their principles from classical ones, over corpus based, to knowledge rich approaches. Authors considered various

aspects, which can affect the categorization. A special attention is devoted to application of recent information reduction methods, based on algebraic transformations including Latent Semantic Analysis. Evaluation measures for assessing quality of a summary and taxonomy of evaluation measures is also presented.

Chapter 7, "From Faceted Classification to Knowledge Discovery of Semi-Structured Text Records", by Yee et al. implement a faceted classification approach to enhance the retrieval and knowledge discovery from extensive aerospace in-service records. The retrieval mechanism afforded by faceted classification can expedite responses to urgent in-service issues as well as enable knowledge discovery that could potentially lead to root-cause findings and continuous improvement.

Lui and Chiu in Chapter 8, "Multi-Value Association Patterns and Data Mining" discuss three related multi-value association patterns and their relationships. All of them have shown to be very important for data mining involving discrete valued data. Furthermore, authors generalized sequential data and have a more specific interpretation to multi-variable patterns. Further evaluations with respect to their conceptual mathematical properties are also presented in the Chapter.

Chapter 9, "Clustering Time Series Data: An Evolutionary Approach" by Monica et.al. discusses the state-of-the-art methodology for some mining time series databases and presents a new evolutionary algorithm for times series clustering an input time series data set. The data mining methods presented include techniques for efficient segmentation, indexing, and clustering time series.

Chapter 10, "Support Vector Clustering: From Local Constraint to Global Stability" by Khanloo et. al. Study the unsupervised support vector method for clustering and establishing a reliable framework for automating the clustering procedure and regularizing the complexity of the decision boundaries. The studied method takes advantage of the information obtained from a Mixture of Factor Analyzers (MFA) assuming that lower dimensional non-linear manifolds are locally linearly related and smoothly changing.

Chapter 11, "New algorithms for generation decision trees: Ant-Miner and its modifications" by Boryczka and Kozak propose new modifications in the original Ant-Miner rule producer. Compared to the previous implementations and settings of Ant-Miner, authors experimental studies illustrate how these extensions improve, or sometimes deteriorate, the performance of Ant-Miner.

The final Part of the book deals with the data mining applications. It contains five chapters, which discusses Adaptive Path Planning on Large Road Networks and provide a Framework for Composing Knowledge Discovery Workflows in Grids

Chapter 12, "Automated Incremental Building of Weighted Semantic Web Repository" by Martin and Roman, introduce an incremental algorithm creating a self-organizing repository and it describes the processes needed for updates and inserts into the repository, especially the processes updating estimated structure driving data storage in the repository. The process of building repository is foremost aimed at allowing the well-known Semantic web tools to query data presented by the current web sources.

Chapter 13, "A Data Mining Approach for Adaptive Path Planning on Large Road Networks" by Awasthi et al. presents a statistical approach for approximating fastest paths under stepwise constant input flows and initial states of the arcs on urban networks. Hybrid clustering and canonical correlation analysis have been used to find arc states and input flows that govern the fastest paths on the network.

In Chapter 14,"Linear Models for Visual Data Mining in Medical Images" by Machado proposes an analysis of available methods for data mining for very high-dimensional sets of data obtained from medical imaging modalities. When applied to imaging studies, data reduction methods may be able to minimize data redundancy and reveal subtle or hidden patterns. Author analysis is concentrated on linear transformation models based on unsupervised learning that explores the relationships among morphologic variables, in order to find clusters with strong correlation.

In Chapter 15, "A Framework for Composing Knowledge Discovery Workflows in Grids" by Lackovic et.al., present a framework to support the execution of knowledge discovery workflows in computational grid environments by executing data mining and computation intelligence algorithms on a set of grid nodes. The framework is an extension of Weka, an open-source toolkit for data mining and knowledge discovery, and makes use of Web Service technologies to access Grid resources and distribute the computation. We present the implementation of the framework and show through some applications how it supports the design of knowledge discovery workflows and their execution on a Grid.

In Chapter 16, "Distributed Data Clustering: A Comparative Analysis" Karthikeyani and Thangavel, compare the performance of two distributed clustering algorithms namely, Improved Distributed Combining Algorithm and Distributed K-Means algorithm against traditional Centralized Clustering Algorithms. Both algorithms use cluster centroid to form a cluster ensemble, which is required to perform global clustering. The centroid based partitioned clustering algorithms namely K-Means, Fuzzy K-Means and Rough K-Means are used with each distributed clustering algorithm, in order to analyze the performance of both hard and soft clusters in distributed environment.

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, Egypt
Václav Snášel, Ostrava, Czech Republic

Contents