Studies in Computational Intelligence, Volume 126

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. 105. Wolfgang Guenthner
Enhancing Cognitive Assistance Systems with Inertial Measurement Units, 2008
ISBN 978-3-540-76996-5

Vol. 106. Jacqueline Jarvis, Dennis Jarvis, Ralph Rönnquist and Lakhmi C. Jain (Eds.)
Holonic Execution: A BDI Approach, 2008
ISBN 978-3-540-77478-5

Vol. 107. Margarita Sordo, Sachin Vaidya and Lakhmi C. Jain (Eds.)
Advanced Computational Intelligence Paradigms in Healthcare - 3, 2008
ISBN 978-3-540-77661-1

Vol. 108. Vito Trianni
Evolutionary Swarm Robotics, 2008
ISBN 978-3-540-77611-6

Vol. 109. Panagiotis Chountas, Ilias Petrounias and Janusz Kacprzyk (Eds.)
Intelligent Techniques and Tools for Novel System Architectures, 2008
ISBN 978-3-540-77621-5

Vol. 110. Makoto Yokoo, Takayuki Ito, Minjie Zhang, Juhnyoung Lee and Tokuro Matsuo (Eds.)
Electronic Commerce, 2008
ISBN 978-3-540-77808-0

Vol. 111. David Elmakias (Ed.)
New Computational Methods in Power System Reliability, 2008
ISBN 978-3-540-77810-3

Vol. 112. Edgar N. Sanchez, Alma Y. Alanís and Alexander G. Loukianov
Discrete-Time High Order Neural Control: Trained with Kalman Filtering, 2008
ISBN 978-3-540-78288-9

Vol. 113. Gemma Bel-Enguix, M. Dolores Jimenez-Lopez and Carlos Martín-Vide (Eds.)
New Developments in Formal Languages and Applications, 2008
ISBN 978-3-540-78290-2

Vol. 114. Christian Blum, Maria José Blesa Aguilera, Andrea Roli and Michael Sampels (Eds.)
Hybrid Metaheuristics, 2008
ISBN 978-3-540-78294-0

Vol. 115. John Fulcher and Lakhmi C. Jain (Eds.)
Computational Intelligence: A Compendium, 2008
ISBN 978-3-540-78292-6

Vol. 116. Ying Liu, Aixin Sun, Han Tong Loh, Wen Feng Lu and Ee-Peng Lim (Eds.)
Advances of Computational Intelligence in Industrial Systems, 2008
ISBN 978-3-540-78296-4

Vol. 117. Da Ruan, Frank Hardeman and Klaas van der Meer (Eds.)
Intelligent Decision and Policy Making Support Systems, 2008
ISBN 978-3-540-78306-0

Vol. 118. Tsau Young Lin, Ying Xie, Anita Wasilewska and Churn-Jung Liau (Eds.)
Data Mining: Foundations and Practice, 2008
ISBN 978-3-540-78487-6

Vol. 119. Slawomir Wiak, Andrzej Krawczyk and Ivo Dolezel (Eds.)
Intelligent Computer Techniques in Applied Electromagnetics, 2008
ISBN 978-3-540-78489-0

Vol. 120. George A. Tsihrintzis and Lakhmi C. Jain (Eds.)
Multimedia Interactive Services in Intelligent Environments, 2008
ISBN 978-3-540-78491-3

Vol. 121. Nadia Nedjah, Leandro dos Santos Coelho and Luiza de Macedo Mourelle (Eds.)
Quantum Inspired Intelligent Systems, 2008
ISBN 978-3-540-78531-6

Vol. 122. Tomasz G. Smolinski, Mariofanna G. Milanova and Aboul-Ella Hassanien (Eds.)
Applications of Computational Intelligence in Biology, 2008
ISBN 978-3-540-78533-0

Vol. 123. Shuichi Iwata, Yukio Ohsawa, Shusaku Tsumoto, Ning Zhong, Yong Shi and Lorenzo Magnani (Eds.)
Communications and Discoveries from Multidisciplinary Data, 2008
ISBN 978-3-540-78732-7

Vol. 124. Ricardo Zavala Yoe
Modelling and Control of Dynamical Systems: Numerical Implementation in a Behavioral Framework, 2008
ISBN 978-3-540-78734-1

Vol. 125. Larry Bull, Bernadó-Mansilla Ester and John Holmes (Eds.)
Learning Classifier Systems in Data Mining, 2008
ISBN 978-3-540-78978-9

Vol. 126. Oleg Okun and Giorgio Valentini (Eds.)
Supervised and Unsupervised Ensemble Methods and their Applications, 2008
ISBN 978-3-540-78980-2

Oleg Okun
Giorgio Valentini
(Eds.)

Supervised and Unsupervised Ensemble Methods and their Applications

With 50 Figures and 46 Tables

Springer

Dr. Oleg Okun
Machine Vision Group
Infotech Oulu
&
Department of Electrical and Information
Engineering
University of Oulu
P.O. Box 4500
FI-90014 Oulu
Finland
oleg@ee.oulu.fi

Giorgio Valentini
Dipartimento di Scienze dell'Informazione
Universita degli Studi di Milano
Via Comelico 39
20135 Milano
Italy
valentini@dsi.unimi.it

ISBN 978-3-540-78980-2 e-ISBN 978-3-540-78981-9

Studies in Computational Intelligence ISSN 1860-949X

Library of Congress Control Number: 2008924367

Cover design: Deblik, Berlin, Germany

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To my parents, Raisa and Gregory, and to my wife, Helen
– Oleg Okun

Al mio caro topolino, che amo tanto
– Giorgio Valentini

Preface

The rapidly growing amount of data, available from different technologies in the field of bio-sciences, high-energy physics, economy, climate analysis, and in several other scientific disciplines, requires a new generation of machine learning and statistical methods to deal with their complexity and heterogeneity.

As data collections becomes easier, data analysis is required to be more sophisticated in order to extract useful information from the available data. Even if data can be represented in several ways, according to their structural characteristics, ranging from strings, lists, trees to graphs and other more complex data structures, in most applications they are typically represented as a matrix whose rows correspond to measurable characteristics called features, attributes, variables, depending on the considered discipline and whose columns correspond to examples (cases, samples, patterns). In order to avoid confusion, we will talk about features and examples. In real-world tasks, there can be many more features than examples (cancer classification based on gene expression levels in bioinformatics) or there can be many more examples than features (intrusion detection in computer/network security). In addition, each example can be either labeled or not. Attaching labels allows to distinguish members of the same class or group from members of other classes or groups. Hence, one can talk about supervised and unsupervised tasks that can be solved by machine learning methods.

Since it is widely accepted that no single classifier or clustering algorithm can be superior to the others, ensembles of supervised and unsupervised methods are gaining popularity. A typical ensemble includes a number of classifiers/clusterers whose predictions are combined together according to a certain rule, e.g. majority vote.

Statistical, algorithmical, representational, computational and practical reasons can explain the success of ensemble methods. In particular several empirical results have demonstrated that ensembles often provide a better solution to the problem than any single method.

This book was inspired by the last argument and resulted from the workshop on Supervised and Unsupervised Ensemble Methods and their Applications (briefly, SUEMA) organized on June 4, 2007 in Girona, Spain. This workshop was held in conjunction with the 3rd Iberian Conference on Pattern Recognition and Image Analysis and was intended to encompass the progress in the ensemble applications made by the Iberian and international scholars. Despite its small format, SUEMA attracted researchers from Spain, Portugal, France, USA, Italy, and Finland, who presented interesting ideas about using the ensembles in various practical cases. Encouraged by this enthusiastic reply, we decided to publish workshop papers in an edited book, since CD proceedings were the only media distributed among the workshop participants at that time.

The book includes nine chapters divided into two parts, assembling contributions to the applications of supervised and unsupervised ensembles. Chapter 1 serves the tutorial purpose as an introduction to unsupervised ensemble methods. Chapter 2 concerns ensemble clustering of categorical data where symbolic names are assigned to examples as labels. Chapter 3 describes fuzzy ensemble clustering applied to gene expression data for cancer classification and discovery of subclasses of pathologies at bio-molecular level. Chapter 4 introduces collaborative multi-strategical clustering where individual algorithms attempt to find a consensus with other ensemble members while grouping the data into clusters on remote sensed images of urban and coastal areas. Chapter 5 presents the application of ensembles combining one-class classifiers to computer network intrusion detection. Chapter 6 deals with ensembles of nearest neighbor classifiers for gene expression based cancer classification. Chapter 7 applies the two-level ensemble scheme called stacking to multivariate time series classification for industrial process diagnosis and speaker recognition. Chapter 8 concentrates on the analysis of heteroskedastic financial time series by means of boosting-like ensembles utilizing neural networks as the base classifiers. Chapter 9 explores three two-level ensemble schemes – stacking, grading, and cascading – when working with nominal data.

The book is intended to be primarily a reference work. It could be a good complement to two excellent books on ensemble methodology – "*Combining pattern classifiers: methods and algorithms*" by Ludmila Kuncheva (John Wiley & Sons, 2004) and "*Decomposition methodology for knowledge discovery and data mining: theory and applications*" by Oded Maimon and Lior Rokach (World Scientific, 2005). Extra primal sources of information are proceedings of the biannual international workshop on Multiple Classifier Systems (MCS) published by Springer-Verlag, and proceedings of the International Conference on Information Fusion (FUSION) organized by the International Society of Information Fusion (http://www.isif.org/). Among other conferences of interest are International Conference on Machine Learning (ICML), European Conference on Machine Learning (ECML), and International Conference on Machine Learning and Data Mining (MLDM) (proceedings of the two latter are published by Springer-Verlag). Two international journals are largely devoted to

the topic of our book are Information Fusion published by Elsevier and Journal of Advances in Information Fusion published by The International Society of Information Fusion, but most machine learning journals such as Machine Learning, the Journal of Machine Learning Research and the IEEE Transactions on Pattern Analysis and Machine Intelligence dedicate large room to papers on ensemble methods. These recommended sources, of course, do not constitute the complete list to look in, since nowadays ensemble methods gain increasing popularity and thus, they are among topics of many scientific meetings and journal issues. Our book would dramatically increase in size if we tried to list all events and publications related to ensemble methods. Hence, we use this argument to apologize to all researchers and organizations whose valuable contribution to the exciting field of ensembles we unintentionally missed.

We are grateful to many people who helped this book to appear. We would like to thank Prof. Joan Martí and Dr. Joaquim Salvi for providing us with great opportunity to hold the abovementioned workshop in Girona. We are also thankful to all authors who spent their time and efforts to contribute to this book. Prof. Janusz Kacprzyk and Dr. Thomas Ditzinger from Springer-Verlag deserved our special acknowledgment for warm welcome to our book and their support and a great deal of encouragement.

Oulu (Finland) and Milan (Italy),
January 2008

Oleg Okun
Giorgio Valentini

Contents

List of Contributors

Carlos Alonso
University of Valladolid, Spain
calonso@infor.uva.es

Muna Al-Razgan
George Mason University, USA
malrazda@gmu.edu

Roberto Avogadri
University of Milan, Italy
avogadri@dsi.unimi.it

Daniel Barbará
George Mason University, USA
dbarbara@gmu.edu

Aníbal Bregón
University of Valladolid, Spain
anibal@infor.uva.es

Igino Corona
University of Cagliari, Italy
igino.corona@diee.unica.it

Carlotta Domeniconi
George Mason University, USA
carlotta@ise.gmu.edu

Manuel Febrero
University of Santiago de Compostela, Spain
mfebrero@usc.es

Germain Forestier
University of Strasburg, France
forestier@lsiit.u-strasbg.fr

Ana Fred
Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal
afred@lx.it.pt

Pierre Gançarski
University of Strasburg, France
gancarski@lsiit.u-strasbg.fr

César García Osorio
University of Burgos, Spain
cgosorio@ubu.es

Giorgio Giacinto
University of Cagliari, Italy
giacinto@diee.unica.it

Wenceslao González-Manteiga
University of Santiago de Compostela, Spain
wences@usc.es

André Lourenço
Instituto de Telecomunicações, Instituto Superior de Engenharia de Lisboa, Portugal
arlourenco@lx.it.pt

José Matías
University of Vigo, Spain
jmmatias@uvigo.es

Jesús Maudes
University of Burgos, Spain
jmaudes@ubu.es

Oleg Okun
University of Oulu, Finland
oleg@ee.oulu.fi

Óscar Prieto
University of Valladolid, Spain
oscapri@infor.uva.es

Helen Priisalu
Teradata, Finland
hp185014@ncr.com

Juan Reboredo
University of Santiago de Compostela, Spain
juanca@usc.es

Juan José Rodríguez
University of Burgos, Spain
jjrodriguez@ubu.es

Fabio Roli
University of Cagliari, Italy
roli@diee.unica.it

Giorgio Valentini
University of Milan, Italy
valentini@dsi.unimi.it

Cédric Wemmert
University of Strasburg, France
wemmert@lsiit.u-strasbg.fr