

LNC8 8206

Hujun Yin Ke Tang Yang Gao
Frank Klawonn Minhoo Lee Bin Li
Thomas Weise Xin Yao (Eds.)

Intelligent Data Engineering and Automated Learning – IDEAL 2013

14th International Conference, IDEAL 2013
Hefei, China, October 2013
Proceedings



Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Hujun Yin Ke Tang Yang Gao
Frank Klawonn Minhoo Lee Bin Li
Thomas Weise Xin Yao (Eds.)

Intelligent Data Engineering and Automated Learning – IDEAL 2013

14th International Conference, IDEAL 2013
Hefei, China, October 20-23, 2013
Proceedings

Volume Editors

Hujun Yin

The University of Manchester, UK; E-mail: hujun.yin@manchester.ac.uk

Ke Tang

Bin Li

Thomas Weise

University of Science and Technology of China, Hefei, China

E-mail: {ketang, twiese, binli}@ustc.edu.cn

Yang Gao

Nanjing University, China; E-mail: gaoy@nju.edu.cn

Frank Klawonn

Ostfalia University of Applied Sciences, Wolfenbüttel, Germany

E-mail: f.klawonn@ostfalia.de

Minho Lee

Kyungpook National University, Daegu, Korea; E-mail: mholee@knu.ac.kr

Xin Yao

University of Birmingham, UK; E-mail: x.yao@cs.bham.ac.uk

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-41277-6

e-ISBN 978-3-642-41278-3

DOI 10.1007/978-3-642-41278-3

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013948928

CR Subject Classification (1998): I.2.6-7, H.2.8, I.5, I.2.10-11, J.3, F.2.2, H.3, F.1, H.4

LNCS Sublibrary: SL 3 – Information Systems and Application,
incl. Internet/Web and HCI

© 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.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

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

Preface

In this digital era and subsequent “data rich but information poor” quandary, the IDEAL conference serves its purposes perfectly – making sense of huge volumes of data, evaluating the complexity of real-world problems, and turning data into information and knowledge. The IDEAL conference attracts international experts, researchers, leading academics, practitioners, and industrialists from communities of machine learning, computational intelligence, data mining, knowledge management, biology, neuroscience, bio-inspired systems and agents, and distributed systems. It has enjoyed a vibrant and successful history in the last 15 years, having been held in over 11 locations in 7 different countries. It continues to evolve to embrace emerging topics and exciting trends. This year IDEAL set foot in mainland China, the fastest growing economy in the world. The conference received about 130 submissions, which were rigorously peer-reviewed by Program Committee members. Only the papers judged to be of highest quality were accepted and included in these proceedings.

This volume contains 76 papers accepted and presented at the 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2013), held on 20–23 October 2013 in Hefei, China. These papers provided a valuable collection of recent research outcomes in data engineering and automated learning, from methodologies, frameworks, and techniques to applications. In addition to various topics such as evolutionary algorithms; neural networks; probabilistic modelling; swarm intelligent; multi-objective optimization and practical applications in regression, classification, clustering, biological data processing, text processing, and video analysis, IDEAL 2013 also featured a number of special sessions on emerging topics such as adaptation and learning multi-agent systems, big data, swarm intelligence and data mining, and combining learning and optimization in intelligent data engineering.

We would like to thank all the people who devoted so much time and effort to the successful running of the conference, in particular the members of the Program Committee and reviewers, as well as the authors who contributed to the conference. We are also very grateful for the hard work of the local organizing team at the University of Science and Technology of China (USTC), especially Prof. Bin Li, in local arrangements, as well as the help provided by the University

of Manchester in checking through all the camera-ready files. Continued support and collaboration from Springer, in particular from the LNCS editor, Alfred Hoffman and Anna Kramer, are also greatly appreciated.

August 2013

Hujun Yin
Ke Tang
Yang Gao
Frank Klawonn
Minho Lee
Bin Li
Thomas Weise
Xin Yao

Organization

General Chair

Xin Yao	University of Birmingham, UK
---------	------------------------------

Program Chair

Hujun Yin	University of Manchester, UK
-----------	------------------------------

Program Co-chairs

Ke Tang	University of Science and Technology of China (USTC), China
Yang Gao	Nanjing University, China
Frank Klawonn	Ostfalia University of Applied Sciences, Germany
Minho Lee	Kyungpook National University, South Korea

International Advisory Committee

Lei Xu (Chair)	Chinese University of Hong Kong, Hong Kong
Yaser Abu-Mostafa	CALTECH, USA
Shun-ichi Amari	RIKEN, Japan
Michael Dempster	University of Cambridge, UK
José R. Dorronsoro	Autonomous University of Madrid, Spain
Nick Jennings	University of Southampton, UK
Soo-Young Lee	KAIST, South Korea
Erkki Oja	Helsinki University of Technology, Finland
Latit M. Patnaik	Indian Institute of Science, India
Burkhard Rost	Columbia University, USA
Xin Yao	University of Birmingham, UK

Steering Committee

Hujun Yin (Chair)	University of Manchester, UK
Laiwan Chan (Chair)	Chinese University of Hong Kong, Hong Kong
Guilherme Barreto	Federal University of Ceará, Brazil
Yiu-ming Cheung	Hong Kong Baptist University, Hong Kong

VIII Organization

Emilio Corchado	University of Burgos, Spain
Jose A. Costa	Federal University of Rio Grande do Norte, Brazil
Colin Fyfe	University of the West of Scotland, UK
Marc van Hulle	K.U. Leuven, Belgium
Samuel Kaski	Helsinki University of Technology, Finland
John Keane	University of Manchester, UK
Jimmy Lee	Chinese University of Hong Kong, Hong Kong
Malik Magdon-Ismail	Rensselaer Polytechnic Inst., USA
Vic Rayward-Smith	University of East Anglia, UK
Peter Tino	University of Birmingham, UK
Zheng Rong Yang	University of Exeter, UK
Ning Zhong Maebashi	Inst. of Technology, Japan

Publicity Co-chairs

Emilio Corchado	University of Salamanca, Spain
Jose A. Costa	Federal University of Rio Grande do Norte, Brazil
Thomas Weise	USTC, China

International Liaison

Jinlong Li	USTC, China
David Camacho	Universidad Autónoma de Madrid, Spain
Guilherme Barreto	Federal University of Ceará, Brazil
Brijesh Verma	Central Queensland University, Australia

Local Organizing Committee

Bin Li (Chair)	University of Science and Technology of China
Kaiming Chen	University of Science and Technology of China
Jinlong Li	University of Science and Technology of China
Thomas Weise	University of Science and Technology of China
Rui Xu	University of Science and Technology of China
Ke Tang	University of Science and Technology of China

Program Committee

Ajith Abraham	Romis Attux
Davide Anguita	Javier Bajo Pérez
Bruno Apolloni	Bruno Baruaque
Francisco Assis	Carmelo J.A. Bastos Filho

Vicent Botti
 Antonio Braga
 Fernando Buarque
 Andrés Bustillo
 José Luis Calvo Rolle
 David Camacho
 Heloisa Camargo
 Anne Canuto
 Jaime Cardoso
 Andre Carvalho
 Matthew Casey
 Darryl Charles
 Richard Chbeir
 Chunlin Chen
 Ling Chen
 Songcan Chen
 Tianshi Chen
 Sung-Bae Cho
 Seungjin Choi
 Stelvio Cimato
 David Clifton
 André Coelho
 Leandro Coelho
 Francesco Corona
 Paulo Cortez
 Marcelo A. Costa
 Raúl Cruz-Barbosa
 Ernesto Cuadros-Vargas
 Alfredo Cuzzocrea
 Ernesto Damiani
 Nicoletta Dessì
 Fernando Díaz
 Weishan Dong
 Jose Dorronsoro
 Adrião Duarte
 Jochen Einbeck
 Pablo Estevez
 Igor Farkas
 Jan Feyereisl
 Felipe M.G. França
 Richard Freeman
 Kunihiko Fukushima
 Colin Fyfe
 Marcus Gallagher
 Laura Garcia

Ana Belén Gil
 Fernando Gomide
 Petro Gopych
 Marcin Gorawski
 Juan Manuel Górriz
 Lars Graening
 Manuel Grana
 Maciej Grzenda
 Arkadiusz Grzybowski
 Jerzy Grzymala-Busse
 Alberto Guillen
 Barbara Hammer
 Ioannis Hatzilygeroudis
 Francisco Herrera
 Álvaro Herrero
 J. Michael Herrmann
 James Hogan
 Jaakko Hollmén
 Vasant Honavar
 Wei-Chiang Samuelson Hong
 David Hoyle
 Estevam Rafael Hruschka Junior
 Naohiro Ishii
 Konrad Jackowski
 Gil-Jin Jang
 He Jiang
 Ata Kaban
 Juha Karhunen
 Rheeman Kil
 Dae-Shik Kim
 Dong-Joo Kim
 Kee-Eung Kim
 Kyunghwan Kim
 Kyung-Joong Kim
 Mario Koeppen
 Andreas König
 Bartosz Krawczyk
 Pavel Kromer
 Rudolf Kruse
 Sungoh Kwon
 Pei Ling Lai
 Jonghwan Lee
 Seong-Whan Lee
 Soo-Young Lee
 Ming Li

Jongjin Lim
Paulo Lisboa
Jing Liu
Wenjian Luo
Jiancheng Lv
José Everardo B. Maia
Roque Marin Morales
Urszula Markowska Kaczmar
Jan Martinovic
Giancarlo Mauri
José Ramón Méndez
Yoan Miche
Byoung-Kyong Min
José Manuel Molina
Petr Musilek
Tim Nattkemper
Antonio Neme
He Ni
Yusuke Nojima
Erkki Oja
Chung-Ming Ou
Seiichi Ozawa
Vasile Palade
Hyeyoung Park
Hyung-Min Park
Juan Pavón
Carlos Pedreira
Jan Platos
Jose Principe
Héctor Quintian
Yacine Rebahi
Raquel Redondo
Agnaldo Reis
Bernardete Ribeiro
Ignacio Rojas
Jesús Ángel Román
Fabrice Rossi
Sherif Sakr
Luciano Sánchez Ramos
Javier Sedano
Jose Seixas
Yinghuan Shi
Catarina Silva

Hernando Silva
Ivan Silva
Dragan Simic
Olli Simula
Michael Small
Piotr Sobolewski
Alessandro Sperduti
Ilhong Suh
Ying Tan
Ricardo Tanscheit
Dante Tapia
Chuan-Kang Ting
Peter Tino
Renato Tinós
Alicia Troncoso
Eiji Uchino
Marc van Hulle
Belen Vaquerizo
Marley Vellasco
Alfredo Vellido Alcacena
Vicente Vera
Michel Verleysen
Fernando Von Zuben
Jian-Wu Wan
Jiahai Wang
Lipo Wang
Wenjia Wang
Yong Wang
Katarzyna Wegrzyn-Wolska
Dongqing Wei
Thomas Weise
Michal Wozniak
Zhenyu Yang
Takashi Yoneyama
Yang Yu
Du Zhang
Min-Ling Zhang
Aimin Zhou
Huiyu Zhou
Zexuan Zhu
Rodolfo Zunino

Special Session on Adaptive and Learning Multi-agent Systems

Organizers

Hongbin Dong	Harbin Engineering University, China
Jun He	Aberystwyth University, UK
Xinjun Mao	National University of Defense Technology ,China
Xiangrong Tong	Yantai University, China

Special Session on Big Data

Organizers

Hui Xiong	Rutgers University, USA
Wenjun Zhou	University of Tennessee, USA

Special Session on Soft-Computing Algorithms in Renewable Energy Problems

Organizers

Sancho Salcedo Sanz	Universidad de Alcalá, Madrid, Spain
Jose Antonio Portilla-Figueras	Universidad de Alcalá, Madrid, Spain

Special Session on Swarm Intelligence and Data Mining

Organizers

Jing Liang	Zhengzhou University, China
Chuan-Kang Ting	National Chung Cheng University, Taiwan
Ying-ping Chen	National Chiao Tung University, Taiwan

Special Session on Text Data Learning

Organizers

Baoli Li	Henan University of Technology, China
Carl Vogel	Trinity College Dublin, Ireland

Special Session on Coevolution

Organizers

Siang Yew Chong	University of Nottingham, Malaysia Campus
Zhenyu Yang	National University of Defense Technology, China
Xiaodong Li	Royal Melbourne Institute of Technology, Australia

Special Session on Combining Learning and Optimization for Intelligent Data Engineering

Organizers

Ata Kaban	The University of Birmingham, UK
Jakramate Bootkrajang	The University of Birmingham, UK

Table of Contents

A Pruning Algorithm for Extreme Learning Machine	1
<i>Li Ying and Li Fan-jun</i>	
Measuring Stability and Discrimination Power of Metrics in Information Retrieval Evaluation	8
<i>Huaji Shi, Yanzhi Tan, Xiaolong Zhu, and Shengli Wu</i>	
System for Monitoring and Optimization of Micro- and Nano-Machining Processes Using Intelligent Voice and Visual Communication	16
<i>Dariusz Lipinski and Maciej Majewski</i>	
Racing for Unbalanced Methods Selection	24
<i>Andrea Dal Pozzolo, Olivier Caelen, Serge Waterschoot, and Gianluca Bontempi</i>	
Super-Resolution from One Single Low-Resolution Image Based on R-KSVD and Example-Based Algorithm	32
<i>FangMei Fu, Jiong Jia, ZhongLong Zheng, Fan Yang, Li Guo, HaiXin Zhang, and MuDan Yu</i>	
Bilateral Multi-issue Parallel Negotiation Model Based on Reinforcement Learning	40
<i>Lihong Chen, Hongbin Dong, Qilong Han, and Guangzhe Cui</i>	
Learning to Detect the Subway Station Arrival for Mobile Users	49
<i>Kuifei Yu, Hengshu Zhu, Huanhuan Cao, Baoxian Zhang, Enhong Chen, Jilei Tian, and Jinghai Rao</i>	
Vision Based Multi-pedestrian Tracking Using Adaptive Detection and Clustering	58
<i>Zhibo Yang and Bo Yuan</i>	
Drilling Cost Prediction Based on Self-adaptive Differential Evolution and Support Vector Regression	67
<i>Huaxian Pan, Guojian Cheng, and Jian Ding</i>	
Web Service Evaluation Method Based on Time-aware Collaborative Filtering	76
<i>Guisheng Yin, Xiaohui Cui, Hongbin Dong, and Yuxin Dong</i>	
An Improved PBIL Algorithm for Path Planning Problem of Mobile Robots	85
<i>Qingbin Zhang, Manjun Cai, Fajun Zhou, and Hairong Nie</i>	
An Initialized ACO for the VRPTW	93
<i>Wei Shi and Thomas Weise</i>	

Deadline-Aware Event Scheduling for Complex Event Processing Systems	101
<i>Na Li and Qiang Guan</i>	
A Discrete Hybrid Bees Algorithm for Service Aggregation Optimal Selection in Cloud Manufacturing	110
<i>Sisi Tian, Quan Liu, Wenjun Xu, and Junwei Yan</i>	
Continuous Motion Recognition Using Multiple Time Constant Recurrent Neural Network with a Deep Network Model	118
<i>Zhibin Yu and Minho Lee</i>	
An Extended Version of the LVA-Index	126
<i>Piotr Lasek</i>	
Anomaly Monitoring Framework Based on Intelligent Data Analysis	134
<i>Prapa Rattadilok, Andrei Petrovski, and Sergei Petrovski</i>	
Customer Unification in E-Commerce.....	142
<i>Marcin Gorawski, Aleksander Chrószcz, and Anna Gorawska</i>	
Network Management Based on Domain Partition for Mobile Agents ...	153
<i>Yonghui Liu and Weidong Min</i>	
Multi-objective Quantum Cultural Algorithm and Its Application in the Wireless Sensor Networks' Energy-Efficient Coverage Optimization	161
<i>Yi-nan Guo, Meirong Chen, and Chun Wang</i>	
Image Super Resolution via Visual Prior Based Digital Image Characteristics	168
<i>Yusheng Jia, Wanqi Yang, Yang Gao, Hujun Yin, and Yinghuan Shi</i>	
Deep Learning on Natural Viewing Behaviors to Differentiate Children with Fetal Alcohol Spectrum Disorder	178
<i>Po-He Tseng, Angelina Paolozza, Douglas P. Munoz, James N. Reynolds, and Laurent Itti</i>	
Prevailing Trends Detection of Public Opinions Based on Tianya Forum	186
<i>Lina Cao and Xijin Tang</i>	
Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model	194
<i>Vivek Narayanan, Ishan Arora, and Arjun Bhatia</i>	
A Scale-Free Based Memetic Algorithm for Resource-Constrained Project Scheduling Problems	202
<i>Lixia Wang and Jing Liu</i>	

A Direction based Multi-Objective Agent Genetic Algorithm	210
<i>Chen Zhu and Jing Liu</i>	
A Study of Representations for Resource Constrained Project Scheduling Problems Using Fitness Distance Correlation	218
<i>Bingqi Cai and Jing Liu</i>	
Adapt a Text-Oriented Chunker for Oral Data: How Much Manual Effort Is Necessary?	226
<i>Isabelle Tellier, Yoann Dupont, Iris Eshkol, and Ilaine Wang</i>	
SVD Based Graph Regularized Matrix Factorization	234
<i>Ephrime A. Vidar and Sweedy K. Alvindia</i>	
Clustering, Noise Reduction and Visualization Using Features Extracted from the Self-Organizing Map	242
<i>Leonardo Enzo Brito da Silva and José Alfredo Ferreira Costa</i>	
Efficient Service Deployment by Image-Aware VM Allocation Strategy	252
<i>Binbin Huang, Rongheng Lin, Kai Peng, Hua Zou, and Fangchun Yang</i>	
Forecasting Financial Time Series Using a Hybrid Self-Organising Neural Model	262
<i>Yicun Ouyang and Hujun Yin</i>	
A Novel Diversity Maintenance Scheme for Evolutionary Multi-objective Optimization	270
<i>Sen Bong Gee, Xin Qiu, and Kay Chen Tan</i>	
Adaptive Differential Evolution Fuzzy Clustering Algorithm with Spatial Information and Kernel Metric for Remote Sensing Imagery	278
<i>Ailong Ma, Yanfei Zhong, and Liangpei Zhang</i>	
Dynamic EM in Neologism Evolution	286
<i>Martin Emms</i>	
Estimation of the Regularisation Parameter in Huber-MRF for Image Resolution Enhancement	294
<i>Sakinah Ali Pitchay and Ata Kabán</i>	
Sparse Prototype Representation by Core Sets	302
<i>Frank-Michael Schleif, Xibin Zhu, and Barbara Hammer</i>	
Reconstruction of Wind Speed Based on Synoptic Pressure Values and Support Vector Regression	310
<i>B. Saavedra-Moreno, S. Salcedo-Sanz, L. Carro-Calvo, A. Portilla-Figueras, and J. Magdalena-Saiz</i>	

Direct Solar Radiation Prediction Based on Soft-Computing Algorithms Including Novel Predictive Atmospheric Variables	318
<i>S. Salcedo-Sanz, C. Casanova-Mateo, A. Pastor-Sánchez, D. Gallo-Marazuela, A. Labajo-Salazar, and A. Portilla-Figueras</i>	
A Novel Coral Reefs Optimization Algorithm for Multi-objective Problems	326
<i>S. Salcedo-Sanz, A. Pastor-Sánchez, D. Gallo-Marazuela, and A. Portilla-Figueras</i>	
Fuzzy Clustering with Grouping Genetic Algorithms	334
<i>S. Salcedo-Sanz, L. Carro-Calvo, A. Portilla-Figueras, L. Cuadra, and D. Camacho</i>	
Graph-Based Substructure Pattern Mining Using CUDA Dynamic Parallelism	342
<i>Fei Wang, Jianqiang Dong, and Bo Yuan</i>	
Scaling Up Covariance Matrix Adaptation Evolution Strategy Using Cooperative Coevolution	350
<i>Jinpeng Liu and Ke Tang</i>	
Gradient Boosting-Based Negative Correlation Learning	358
<i>Lunjun Wan, Ke Tang, and Rui Wang</i>	
Metamodel Assisted Mixed-Integer Evolution Strategies Based on Kendall Rank Correlation Coefficient	366
<i>Lili Zhuang, Ke Tang, and Yaochu Jin</i>	
Semi-supervised Ranking via List-Wise Approach	376
<i>Zhigao Miao and Ke Tang</i>	
Gaussian Process for Transfer Learning through Minimum Encoding . . .	384
<i>Hao Shao, Rui Xu, and Feng Tao</i>	
Kernel Based Manifold Learning for Complex Industry Fault Detection	392
<i>Jian Cheng and Yi-nan Guo</i>	
An Estimation of Distribution Algorithm for the 3D Bin Packing Problem with Various Bin Sizes	401
<i>Yaxiong Cai, Huaping Chen, Rui Xu, Hao Shao, and Xueping Li</i>	
Accelerating BIRCH for Clustering Large Scale Streaming Data Using CUDA Dynamic Parallelism	409
<i>Jianqiang Dong, Fei Wang, and Bo Yuan</i>	
Swarm Intelligence in Big Data Analytics	417
<i>Shi Cheng, Yuhui Shi, Quande Qin, and Ruibin Bai</i>	

Multidimensional Dynamic Trust Measurement Model with Incentive Mechanism for Internetwork	427
<i>Guisheng Yin, Yingjie Wang, Jianguo Zhang, and Hongbin Dong</i>	
Global Path Planning of Wheeled Robots Using a Multi-Objective Memetic Algorithm	437
<i>Fangxiao Wang and Zexuan Zhu</i>	
Quantifying Flow Field Distances Based on a Compact Streamline Representation	445
<i>Lars Graening and Thomas Ramsay</i>	
MCGA: A Multiobjective Cellular Genetic Algorithm Based on a 3D Grid	455
<i>Hu Zhang, Shenming Song, and Aimin Zhou</i>	
Multi-Objective Particle Swarm Optimization Algorithm Based on Population Decomposition	463
<i>Yuan Zhao and Hai-Lin Liu</i>	
An Effective Ant Colony Approach for Scheduling Parallel Batch-Processing Machines	471
<i>Rui Xu, Huaping Chen, and Hao Shao</i>	
Understanding Instance Complexity in the Linear Ordering Problem . . .	479
<i>Josu Ceberio, Leticia Hernando, Alexander Mendiburu, and Jose A. Lozano</i>	
Multi-Objective Evolutionary Algorithm Based on Decomposition for Air Traffic Flow Network Rerouting Problem	487
<i>Xiao Zhang, Mingming Xiao, and Miao Zhang</i>	
Temporal Dependence in Legal Documents	497
<i>Daniel Iseemann, Khurshid Ahmad, Tim Fernando, and Carl Vogel</i>	
Learning-Guided Exploration in Airfoil Optimization	505
<i>Edgar Reehuis, Markus Olhofer, Bernhard Sendhoff, and Thomas Bäck</i>	
A Trigram Language Model to Predict Part of Speech Tags Using Neural Network	513
<i>Dinesh Kumar Kashyap and Gurpreet Singh Josan</i>	
Handling Different Levels of Granularity within Naive Bayes Classifiers	521
<i>Kemal Ince and Frank Klawonn</i>	
Genetic Algorithm on GPU Performance Optimization Issues	529
<i>Andrius Paukštė</i>	

Mutual Information for Performance Assessment of Multi Objective Optimisers: Preliminary Results	537
<i>Noura Al Moubayed, Andrei Petrovski, and John McCall</i>	
Velocity Divergence of CCPSO in Large Scale Global Optimization	545
<i>Shanqing Hu and Bin Li</i>	
Machine Learning Enhanced Multi-Objective Evolutionary Algorithm Based on Decomposition	553
<i>Yung Siang Liao, Kay Chen Tan, Jun Hu, Xin Qiu, and Sen Bong Gee</i>	
High against Low Quantile Comparison for Biomarker and Classifier Evaluation	561
<i>Katharina Tschumitschew and Frank Klawonn</i>	
Learning a Label-Noise Robust Logistic Regression: Analysis and Experiments	569
<i>Jakramate Bootkrajang and Ata Kabán</i>	
Hybrid Bacterial Foraging Algorithm for Data Clustering	577
<i>Ben Niu, Qiqi Duan, and Jing Liang</i>	
Swarm Intelligence with Clustering for Solving SAT	585
<i>Habiba Drias, Ameer Douib, and Célia Hirèche</i>	
Multilevel Bee Swarm Optimization for Large Satisfiability Problem Instances	594
<i>Marwa Djeflal and Habiba Drias</i>	
Voting-XCSc: A Consensus Clustering Method via Learning Classifier System	603
<i>Liqiang Qian, Yinghuan Shi, Yang Gao, and Hujun Yin</i>	
Distance Weighted Cosine Similarity Measure for Text Classification	611
<i>Baoli Li and Liping Han</i>	
A Survey on Benchmarks for Big Data and Some More Considerations	619
<i>Xiongpai Qin and Xiaoyun Zhou</i>	
Spectral Clustering Algorithm Based on Local Sparse Representation	628
<i>Sen Wu, Min Quan, and Xiaodong Feng</i>	
Author Index	637

Distance Weighted Cosine Similarity Measure for Text Classification

Baoli Li and Liping Han

Department of Computer Science
Henan University of Technology
1 Lotus Street, High & New Industrial Development Zone
Zhengzhou, Henan 450001, China
csb11i@gmail.com

Abstract. In Vector Space Model, Cosine is widely used to measure the similarity between two vectors. Its calculation is very efficient, especially for sparse vectors, as only the non-zero dimensions need to be considered. As a fundamental component, cosine similarity has been applied in solving different text mining problems, such as text classification, text summarization, information retrieval, question answering, and so on. Although it is popular, the cosine similarity does have some problems. Starting with a few synthetic samples, we demonstrate some problems of cosine similarity: it is overly biased by features of higher values and does not care much about how many features two vectors share. A distance weighted cosine similarity metric is thus proposed. Extensive experiments on text classification exhibit the effectiveness of the proposed metric.

1 Introduction

Similarity calculation is a basic component for many text mining applications. For example, if we have a perfect method to assess how two text segments are similar, we could build an ideal information retrieval system. In the past years, a lot of metrics [1,2], such as Euclidean distance based metric, Cosine, Jaccard, Dice, Jensen-Shannon Divergence based metric, have been proposed to deal with different kinds of information retrieval and natural language processing problems. Among the existing metrics, Cosine, which measures the angle between two vectors, is the most popular one. It is effectively calculated as dot-product of two normalized vectors.

Given two N dimension vectors \vec{v} and \vec{w} , the cosine similarity between them is calculated as follows:

$$\text{Cosine}(\vec{v}, \vec{w}) = \frac{\vec{v} \bullet \vec{w}}{|\vec{v}| |\vec{w}|} = \frac{\sum_{i=1}^N v_i \times w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}}$$

In mathematics perspective, Cosine similarity is perfect. However, if we check it in text mining perspective, it may not always be reasonable. Let's consider a few example vectors shown in figure 1. Suppose these 3-D vectors are derived from five text segments

A, B, C, D, and E. The cosine similarities between segment A and the rest are given in the figure. From the values, we can conclude that segment B is the most similar one of A, as it has the highest cosine. However, is it reasonable? Intuitively, text segments C and E, which both have two common terms with segment A, are more relevant to A than B, which contains only one term. Moreover, the segment E has one more term than segment A, but Cosine(A,E) is much lower than Cosine(A,B). If we regard the additional term as a noise and neglect it, E will have the same vector as A.

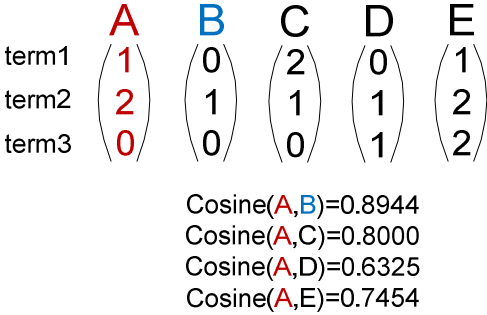


Fig. 1. Cosine similarities between five synthetic vectors

It can thus be derived from the above figure that cosine similarity tends to be overly biased by the features of higher values, but it doesn't care much about how many features two vectors share. In text mining perspective, more features two text segments share, more similar they are. If a part of a text segment is much similar to another segment as whole, the former one is usually thought to be relevant to the latter in Information Retrieval. It is this observation that motivates us to explore more effective similarity metrics than Cosine for text mining.

Because of the proven effectiveness of cosine similarity, we decide to derive new metrics by slightly modifying it. Several distance weighted versions are explored, where distance tends to capture how many features two text segments share. With extensive experiments on a classical text mining problem, i.e. text classification, we obtain a distance weighted cosine metric that performs better than the original cosine metric in most cases. It is also demonstrated with experiments that the similarity metric does have important effects in text mining applications.

The rest of this paper is organized as follows: section 2 introduces the explored distance weighted cosine metrics; section 3 presents extensive experiments on three text classification problems and discussion on the results; Section 4 concludes the paper.

2 Distance Weighted Cosine Similarity Measures

We explore different new similarity metrics with the following evidences or assumptions: 1. cosine similarity is good enough for most text mining applications; 2. more features two text segments share, more similar they are. Therefore, all of the designed metrics have two key components: cosine similarity and distance measure, but they are different in applying different distance measures and assembling strategies.

The explored distance measures, which are expected to capture how many features two vectors share or not, include:

Hamming Distance: it counts how many features two vectors do not share. As vectors may have quite different numbers of valid features, a normalized version, which stands for the percentage of distinct features, is used. Given two N-dimension vectors \vec{v} and \vec{w} , a possible formula to calculate this measure is as follows:

$$HD(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^N (\text{sgn}(v_i) - \text{sgn}(v_i * w_i))}{\sum_{i=1}^N \text{sgn}(v_i)} + \frac{\sum_{i=1}^N (\text{sgn}(w_i) - \text{sgn}(w_i * v_i))}{\sum_{i=1}^N \text{sgn}(w_i)},$$

while $\text{sgn}(x)$ is the sign function.

Weighted Hamming Distance: the former measure takes each feature equally important, which is not ideal. A simple improvement is to weight counts with features' values as follows:

$$WHD(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^N v_i * (1 - \text{sgn}(v_i * w_i))}{\sum_{i=1}^N v_i} + \frac{\sum_{i=1}^N w_i * (1 - \text{sgn}(w_i * v_i))}{\sum_{i=1}^N w_i}.$$

To assemble **Cosine** similarity (dubbed **Cosine**) and **Distance** measure (dubbed **Dist**), several strategies can be applied. As **Cosine** and **Dist** are inverse, we firstly turn the **Dist** into some similarity format, e.g. $\frac{1}{\text{Dist} + 1}$, and then assemble them together with multiplying, averaging, or other strategies. Some possible strategies include:

Multiplying: $\frac{\text{Cosine}}{\text{Dist} + 1}$.

Averaging: $\frac{1}{2} * (\text{Cosine} + \frac{1}{\text{Dist} + 1})$ or $\frac{2 * \frac{\text{Cosine}}{\text{Dist} + 1}}{(\text{Cosine} + \frac{1}{\text{Dist} + 1})}$.

All the above formulas can have different variants. For example, we can add some constants like 1/2, give different weights for the two similarities, and use power or log function to reduce the influence of one component. With extensive experiments on text classification problems, we find the following distance weighted cosine metric could achieve better performance than the traditional cosine similarity in almost all the cases.

$$Dw - \text{Cosine} = \frac{1}{\frac{\text{Dist}^2}{\text{Cosine}} + 1} = \frac{\text{Cosine}}{\text{Dist}^2 + \text{Cosine}}$$

The weighted hamming distance measure is used in the above formula.

3 Experiments and Discussion

In order to evaluate the performance of different distance weighted cosine metrics, we conduct extensive experiments on three single-label text classification problems[3].

3.1 Datasets

We experiment with the following three datasets:

20 Newsgroups: this dataset is evenly partitioned into 20 different newsgroups, each corresponding to a specific topic. Its "bydate" version is widely used in literature, as it has a standard training and test split. The training set has 11,293 samples and the test set 7,528 samples.

Reuters52c: it is a single-label dataset derived from Reuters-21578 with 90 classes by Ana Cardoso-Cachopo during her Ph.D. study [4]. Documents with multiple labels in the original Reuters-21578 (90 classes) dataset are discarded and finally the Reuters52c dataset contains 52 categories, 6,532 documents for training, and 2,568 documents for test. The dataset is imbalanced and some categories only have a few documents, e.g. classes *cpu* and *potato*. We use the all-terms version without stemming.

Sector: this dataset is a collection of web pages belonging to companies from various economic sectors. It has 104 categories, 6,412 training samples, and 3,207 test samples.

3.2 Experimental Settings

We use the vector space model (VSM) for data representation, in which the dimension is determined by the size of the dataset's vocabulary. Each document is then represented as a space vector where the words in the document are mapped onto the corresponding coordinates. In the feature-selection phase of the experiments, we removed words that occur only once [5]. The weight of a feature is given as follows:

$$x_i = \frac{(1 + \log(TF(w_i, d))) \cdot \log(\frac{|D|}{DF(w_i)})}{\sqrt{\sum_j ((1 + \log(TF(w_j, d))) \cdot \log(\frac{|D|}{DF(w_j)}))^2}},$$

which is the same as the standard representation "lfc" in Manning and Schutze [6]. Here, D is the document collection, and TF and DF are a term's frequency in a document d and its document frequency in the collection D respectively.

In classification, we use two widely used algorithms: Centroid and k-Nearest Neighbor [7, 8]. They are all heavily dependent on similarity metrics. With Centroid algorithm, each category is represented by a centroid vector, and a test sample is then classified to the category that has the highest similarity value between its centroid vector and the test sample's vector. With k-Nearest Neighbor algorithm, the category

prediction of a test sample is made according to the category distribution among the top k most similar samples in the training set, where a similarity metric is used to find these k Nearest Neighbors.

We evaluate different similarity metrics, including our proposed distance weighted cosine similarity, the original cosine similarity, Jaccard, and others.

3.3 Evaluation Metric

To evaluate the effectiveness of category assignments to documents by classifiers, the harmonic average of the standard precision and recall, **F1 measure**, is used as follows:

$$F1 = \frac{2recall * precision}{recall + precision}$$

The overall performance on all categories can be computed either by the micro-averaging method or by the macro-averaging method. In micro-averaging, the MicF1 score is computed globally over all the binary decisions. In macro-averaging, the MacF1 score is computed for the binary decisions on each individual category first and then averaged over the categories. The micro-averaged score tends to be dominated by the classifier's performance on common categories, while the macro-averaged score is more influenced by the performance on rare categories.

Table 1. With Centroid classification algorithm, system performance on three datasets with different similarity metrics

Similarity Metric	20 Newsgroups		Reuters52c		Sector	
	MicF1	MacF1	MicF1	MacF1	MicF1	MacF1
Dw-cosine	84.3916	83.6943	90.3816	71.4916	89.0864	89.1664
Cosine	81.6286	80.8969	89.0187	71.7848	87.3402	87.6430
Jaccard	74.5218	73.7537	69.1978	46.2682	76.3018	77.5525

3.4 Results and Discussion

As mentioned in section 2, we conduct extensive experiments with different variants of distance weighted cosine metrics, and find the Dw-Cosine metric performs best. Here we present only the results of this proposed metric from this family. We also experiment with other metrics, e.g Dice and Jensen-Shannon Divergence based metric, but Dice performs similar to Jaccard, and Jensen-Shannon Divergence poorer than Jaccard. So we do not report these metrics' performance here.

Table 1 shows the system performance on three datasets with Centroid classification algorithm. Cosine performs much better than Jaccard on three datasets, and Dw-cosine can increase Cosine's MicF1 with 1.3% to 2.7%. As to the MacF1, Dw-cosine achieves equally good result as Cosine on the Reuters52c dataset, but beats Cosine on the other two datasets with around 2.8% (20 Newsgroups) and 1.5% (Sector), respectively. The improvement is impressive, and it can be seen that choosing a suitable similarity metric is quite important.

To better understand the performance of the three metrics, we depict a column graph in figure 2. Dw-cosine wins the competitions on 5 of 6 tracks, and gains the most on the 20 newsgroups dataset.

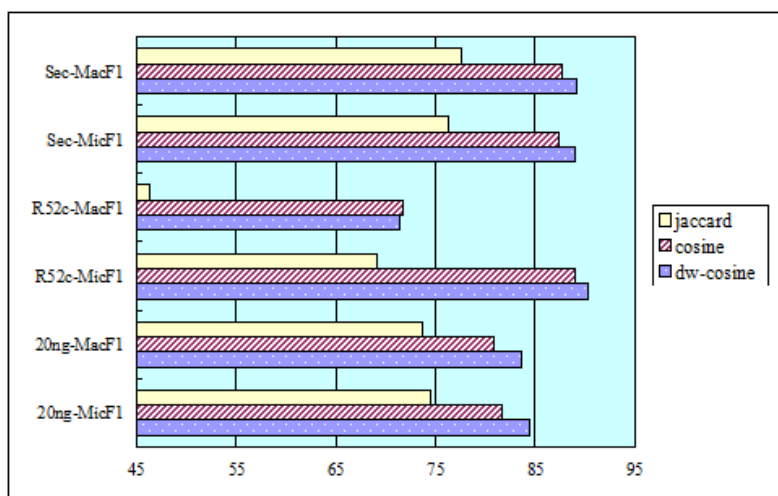


Fig. 2. Visualized system performance on three datasets with Centroid algorithm

Table 2. With k Nearest Neighbour classification algorithm, system performance on three datasets with different similarity metrics, and the values are average of 40 runs (k from 1 to 40)

Similarity Metric	20 Newsgroups		Reuters52c		Sector	
	MicF1	MacF1	MicF1	MacF1	MicF1	MacF1
Dw-cosine	77.1018	76.5166	88.4940	68.2012	84.8566	84.6625
Cosine	76.1351	75.5760	86.6715	68.1401	82.6458	82.4599
Jaccard	71.4865	71.0726	89.5113	63.3028	75.5862	75.3666

Table 2 gives the three metrics' performance with k-Nearest Neighbor algorithm. We vary k from 1 to 40, and then the values in the table are average of these 40 runs. Similar to table 1, Cosine beats Jaccard on 5 tracks, but obtains much poorer MicF1 value on the Reuters52c dataset than Jaccard, which also indicate how important a suitable similarity metric is for a specific text mining problem. Dw-cosine exhibits consistent advantages over Cosine similarity with k-Nearest Neighbor algorithm. It can boost the MicF1 and MacF1 of Cosine from 1% to 2% except for MacF1 on Reuters52c.

T-tests over the 40 runs show that the performance differences in table 2 are all significant but for MacF1 scores of Dw-cosine and Cosine on the Reuters52c dataset.

Similarly, we present a column graph to visualize how these three metrics perform with k-Nearest Neighbor algorithm in figure 3. Dw-cosine and Cosine achieve almost the same MacF1 scores on the Reuters52c dataset, while Dw-cosine shows its biggest advantages over Cosine on the Sector dataset.

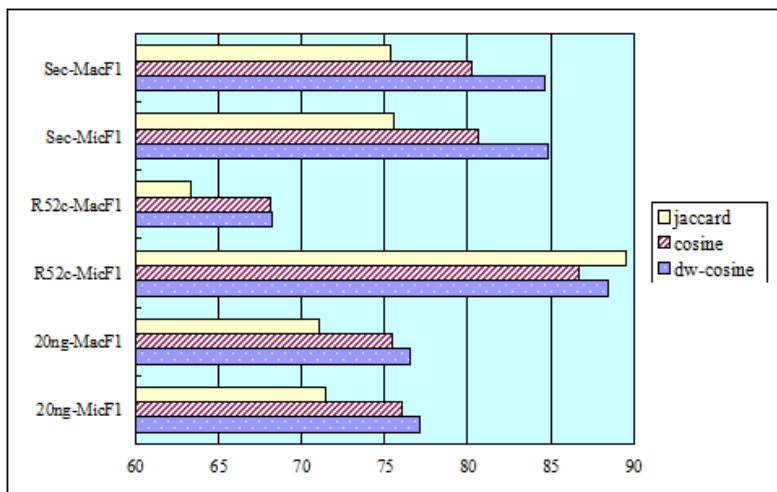


Fig. 3. Visualized average system performance on three datasets with kNN algorithm

4 Conclusions and Future Work

Cosine similarity is widely used in information retrieval, natural language processing, and text mining. It's quite effective, but not perfect. We demonstrate with a few synthetic examples that Cosine similarity tends to be biased by features of higher values and not to pay enough attention to how many features two vectors share. It is this observation that motivates this study. A family of distance weighted cosine metrics is explored and a specific one of this family, i.e. Dw-cosine, achieves consistently better results than the original cosine similarity on three text classification problems. Our experiments also demonstrate that similarity metric may have critical impacts on system performance.

In the future, we would like to experiment with more datasets and apply the proposed distance weighted cosine similarity into other text mining applications, e.g. information retrieval, word sense disambiguation, and so on. We also plan to explore how to detect dataset's properties and suggest suitable similarity metric automatically.

Acknowledgments. This work is supported by the Henan Provincial Research Program on Fundamental and Cutting-Edge Technologies (No. 112300410007), and the High-level Talent Foundation of Henan University of Technology (No. 2012BS027).

References

1. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. (1989) Addison-Wesley Longman Publishing, Boston, MA.

2. Li M., Chen X., Li X., Ma B., and Vitanyi P. M.B.: The Similarity Metric. *IEEE Transactions on Information Theory*, 50(12): 3250-3264 (2004)
3. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1 (2002), 1-47.
4. Cardoso-Cachopo, A.: Improving Methods for Single-label Text Categorization. PhD Thesis, Instituto Superior Técnico, Portugal (2007)
5. Yang Y. and Pedersen J. O.: A comparative study on feature selection in text categorization. In *Proceedings of Fourteenth International Conference on Machine Learning*. (1997): 412-420.
6. Manning C. D. and Schutze H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
7. Yang Y. and Liu X.: A re-examination of text categorization methods. In *Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1999): 42-49.
8. Li B., Lu Q., and Yu S.: An adaptive k-nearest neighbor text categorization strategy. *ACM Transactions on Asian Language Information Processing (TALIP)* 3.4 (2004): 215-226.

Author Index

- Ahmad, Khurshid 497
Ali Pitchay, Sakinah 294
Alvindia, Sweedy K. 234
Arora, Ishan 194
- Bäck, Thomas 505
Bai, Ruibin 417
Bhatia, Arjun 194
Bontempi, Gianluca 24
Bootkrajang, Jakramate 569
- Caelen, Olivier 24
Cai, Bingqi 218
Cai, Manjun 85
Cai, Yaxiong 401
Camacho, D. 334
Cao, Huanhuan 49
Cao, Lina 186
Carro-Calvo, L. 310, 334
Casanova-Mateo, C. 318
Ceberio, Josu 479
Chen, Enhong 49
Chen, Huaping 401, 471
Chen, Lihong 40
Chen, Meirong 161
Cheng, Guojian 67
Cheng, Jian 392
Cheng, Shi 417
Chrószcz, Aleksander 142
Costa, José Alfredo Ferreira 242
Cuadra, L. 334
Cui, Guangzhe 40
Cui, Xiaohui 76
- Dal Pozzolo, Andrea 24
da Silva, Leonardo Enzo Brito 242
Ding, Jian 67
Djeffal, Marwa 594
Dong, Hongbin 40, 76, 427
Dong, Jianqiang 342, 409
Dong, Yuxin 76
Douib, Ameer 585
Drias, Habiba 585, 594
Duan, Qiqi 577
Dupont, Yoann 226
- Emms, Martin 286
Eshkol, Iris 226
- Fan-jun, Li 1
Feng, Xiaodong 628
Fernando, Tim 497
Fu, FangMei 32
- Gallo-Marazuela, D. 318, 326
Gao, Yang 168, 603
Gee, Sen Bong 270, 553
Gorawska, Anna 142
Gorawski, Marcin 142
Graening, Lars 445
Guan, Qiang 101
Guo, Li 32
Guo, Yi-nan 161, 392
- Hammer, Barbara 302
Han, Liping 611
Han, Qilong 40
Hernando, Leticia 479
Hirèche, Célia 585
Hu, Jun 553
Hu, Shanqing 545
Huang, Binbin 252
- Ince, Kemal 521
Isemann, Daniel 497
Itti, Laurent 178
- Jia, Jiong 32
Jia, Yusheng 168
Jin, Yaochu 366
Josan, Gurpreet Singh 513
- Kabán, Ata 294, 569
Kashyap, Dinesh Kumar 513
Klawonn, Frank 521, 561
- Labajo-Salazar, A. 318
Lasek, Piotr 126
Lee, Minho 118
Li, Baoli 611
Li, Bin 545
Li, Na 101

- Li, Xueping 401
 Liang, Jing 577
 Liao, Yung Siang 553
 Lin, Rongheng 252
 Lipinski, Dariusz 16
 Liu, Hai-Lin 463
 Liu, Jing 202, 210, 218
 Liu, Jinpeng 350
 Liu, Quan 110
 Liu, Yonghui 153
 Lozano, Jose A. 479

 Ma, Ailong 278
 Magdalena-Saiz, J. 310
 Majewski, Maciej 16
 McCall, John 537
 Mendiburu, Alexander 479
 Miao, Zhigao 376
 Min, Weidong 153
 Moubayed, Noura Al 537
 Munoz, Douglas P. 178

 Narayanan, Vivek 194
 Nie, Hairong 85
 Niu, Ben 577

 Olhofer, Markus 505
 Ouyang, Yicun 262

 Pan, Huaxian 67
 Paolozza, Angelina 178
 Pastor-Sánchez, A. 318, 326
 Paukštė, Andrius 529
 Peng, Kai 252
 Petrovski, Andrei 134, 537
 Petrovski, Sergei 134
 Portilla-Figueras, A. 310, 318, 326, 334

 Qian, Liqiang 603
 Qin, Quande 417
 Qin, Xiongpai 619
 Qiu, Xin 270, 553
 Quan, Min 628

 Ramsay, Thomas 445
 Rao, Jinghai 49
 Rattadilok, Prapa 134
 Reehuis, Edgar 505
 Reynolds, James N. 178

 Saavedra-Moreno, B. 310
 Salcedo-Sanz, S. 310, 318, 326, 334

 Schleif, Frank-Michael 302
 Sendhoff, Bernhard 505
 Shao, Hao 384, 401, 471
 Shi, Huaqi 8
 Shi, Wei 93
 Shi, Yinghuan 168, 603
 Shi, Yuhui 417
 Song, Shenming 455

 Tan, Kay Chen 270, 553
 Tan, Yanzhi 8
 Tang, Ke 350, 358, 366, 376
 Tang, Xijin 186
 Tao, Feng 384
 Tellier, Isabelle 226
 Tian, Jilei 49
 Tian, Sisi 110
 Tschumitschew, Katharina 561
 Tseng, Po-He 178

 Vidar, Ephrime A. 234
 Vogel, Carl 497

 Wan, Lunjun 358
 Wang, Chun 161
 Wang, Fangxiao 437
 Wang, Fei 342, 409
 Wang, Ilaine 226
 Wang, Lixia 202
 Wang, Rui 358
 Wang, Yingjie 427
 Waterschoot, Serge 24
 Weise, Thomas 93
 Wu, Sen 628
 Wu, Shengli 8

 Xiao, Mingming 487
 Xu, Rui 384, 401, 471
 Xu, Wenjun 110

 Yan, Junwei 110
 Yang, Fan 32
 Yang, Fangchun 252
 Yang, Wanqi 168
 Yang, Zhibo 58
 Yin, Guisheng 76, 427
 Yin, Hujun 168, 262, 603
 Ying, Li 1
 Yu, Kuifei 49
 Yu, MuDan 32

Yu, Zhibin 118
 Yuan, Bo 58, 342, 409

 Zhang, Baoxian 49
 Zhang, HaiXin 32
 Zhang, Hu 455
 Zhang, Jianguo 427
 Zhang, Liangpei 278
 Zhang, Miao 487
 Zhang, Qingbin 85
 Zhang, Xiao 487
 Zhao, Yuan 463

Zheng, ZhongLong 32
 Zhong, Yanfei 278
 Zhou, Aimin 455
 Zhou, Fajun 85
 Zhou, Xiaoyun 619
 Zhu, Chen 210
 Zhu, Hengshu 49
 Zhu, Xiaolong 8
 Zhu, Xibin 302
 Zhu, Zexuan 437
 Zhuang, Lili 366
 Zou, Hua 252