Lecture Notes in Artificial Intelligence 4944

Edited by J.G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Zbigniew W. Raś Shusaku Tsumoto
Djamel Zighed (Eds.)

Mining Complex Data

ECML/PKDD 2007 Third International Workshop, MCD 2007
Warsaw, Poland, September 17-21, 2007
Revised Selected Papers

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Zbigniew W. Raś
University of North Carolina
Department of Computer Science
Charlotte, NC 28223, USA
and
Polish-Japanese Institute of Information Technology
Dept. of Intelligent Systems
Koszykowa 86, 02-008 Warsaw, Poland
E-mail: ras@uncc.edu

Shusaku Tsumoto
Shimane Medical University
School of Medicine, Department of Medical Informatics
89-1 Enya-cho, Izumo 693-8501, Japan
E-mail: tsumoto@computer.org

Djamel Zighed
Université Lumière Lyon 2
5 avenue Pierre Mendès-France, 69676 Bron Cedex, France
E-mail: abdelkader.zighed@univ-lyon2.fr

Library of Congress Control Number: 2008926861

CR Subject Classification (1998): H.2.5, H.2.8, H.3.3

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-68415-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-68415-2 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12273379 06/3180 5 4 3 2 1 0

Preface

This volume contains 20 papers selected for presentation at the Third International Workshop on Mining Complex Data–*MCD 2007*–held in Warsaw, Poland, September 17–21, 2007. MCD is a workshop series that started in conjunction with the 5th IEEE International Conference on Data Mining (ICDM) in Houston, Texas, November 27–30, 2005. The second MCD workshop was held again in conjunction with the ICDM Conference in Hong Kong, December 18–22, 2006.

Data mining and knowledge discovery, as stated in their early definition, can today be considered as stable fields with numerous efficient methods and studies that have been proposed to extract knowledge from data. Nevertheless, the famous golden nugget is still challenging. Actually, the context evolved since the first definition of the KDD process, and knowledge now has to be extracted from data becoming more and more complex.

In the framework of data mining, many software solutions were developed for the extraction of knowledge from tabular data (which are typically obtained from relational databases). Methodological extensions were proposed to deal with data initially obtained from other sources, e.g., in the context of natural language (text mining) and image (image mining). KDD has thus evolved following a unimodal scheme instantiated according to the type of the underlying data (tabular data, text, images, etc.), which, at the end, always leads to working on the classical double entry tabular format.

However, in a large number of application domains, this unimodal approach appears to be too restrictive. Consider for instance a corpus of medical files. Each file can contain tabular data such as results of biological analyses, textual data coming from clinical reports, image data such as radiographies, echograms, or electrocardiograms. In a decision-making framework, treating each type of information separately has serious drawbacks. It appears therefore more and more necessary to consider these different data simultaneously, thereby encompassing all their complexity.

Hence, a natural question arises: how could one combine data of different nature and associate them with a same semantic unit, which is for instance the patient? On a methodological level, one could also wonder how to compare such complex units via similarity measures. The classical approach consists in aggregating partial dissimilarities computed on components of the same type. However, this approach tends to make superposed layers of information. It considers that the whole entity is the sum of its components. By analogy with the analysis of complex systems, it appears that knowledge discovery in complex data cannot simply consist of the concatenation of the partial information obtained from each part of the object. The aim, rather, would be to discover more global knowledge giving a meaning to the components and associating them

with the semantic unit. This fundamental information cannot be extracted by the currently considered approaches and the available tools.

The new data mining strategies shall take into account the specificities of complex objects (units with which the complex data are associated). These specificities are summarized hereafter:

Different kind. The data associated to an object are of different types. Besides classical numerical, categorical or symbolic descriptors, text, image or audio/video data are often available.

Diversity of the sources. The data come from different sources. As shown in the context of medical files, the collected data can come from surveys filled in by doctors, textual reports, measures acquired from medical equipment, radiographies, echograms, etc.

Evolving and distributed. It often happens that the same object is described according to the same characteristics at different times or different places. For instance, a patient may often consult several doctors, each one of them producing specific information. These different data are associated with the same subject.

Linked to expert knowledge. Intelligent data mining should also take into account external information, also called expert knowledge, which could be taken into account by means of ontology. In the framework of oncology, for instance, the expert knowledge is organized under the form of decision trees and is made available under the form of best practice guides called standard option recommendations (SOR).

Dimensionality of the data. The association of different data sources at different moments multiplies the points of view and therefore the number of potential descriptors. The resulting high dimensionality is the cause of both algorithmic and methodological difficulties.

The difficulty of knowledge discovery in complex data lies in all these specificities.

We wish to express our gratitude to all members of the Program Committee and the Organizing Committee. Hakim Hacid (Chair of the Organizing Committee) did a terrific job of putting together and maintaining the home page for the workshop as well as helping us to prepare the workshop proceedings. Also, our thanks are due to Alfred Hofmann of Springer for his support.

December 2007

Zbigniew W. Raś
Shusaku Tsumoto
Djamel Zighed

Organization

MCD 2007 Workshop Committee

Workshop Chairs

Zbigniew W. Raś (University of North Carolina, Charlotte)
Shusaku Tsumoto (Shimane Medical University, Japan)
Djamel Zighed (University Lyon II, France)

Organizing Committee

Hakim Hacid (University Lyon II, France)(Chair)
Rory Lewis (University of North Carolina, Charlotte)
Xin Zhang (University of North Carolina, Charlotte)

Program Committee

Aijun An (York University, Canada)
Elisa Bertino (Purdue University, USA)
Ivan Bratko (University of Ljubljana, Slovenia)
Michelangelo Ceci (University of Bari, Italy)
Juan-Carlos Cubero (University of Granada, Spain)
Agnieszka Dardzińska (Białystok Technical University, Poland)
Tapio Elomaa (Tampere University of Technology, Finland)
Floriana Esposito (University of Bari, Italy)
Mirsad Hadzikadic (UNC-Charlotte, USA)
Howard Hamilton (University Regina, Canada)
Shoji Hirano (Shimane University, Japan)
Mieczyslaw Kłopotek (ICS PAS, Poland)
Bożena Kostek (Technical University of Gdańsk, Poland)
Nada Lavrac (Jozef Stefan Institute, Slovenia)
Tsau Young Lin (San Jose State University, USA)
Jiming Liu (University of Windsor, Canada)
Hiroshi Motoda (AFOSR/AOARD and Osaka University, Japan)
James Peters (University of Manitoba, Canada)
Jean-Marc Petit (LIRIS, INSA Lyon, France)
Vijay Raghavan (University of Louisiana, USA)
Jan Rauch (University of Economics, Prague, Czech Republic)
Henryk Rybiński (Warsaw University of Technology, Poland)
Dominik Ślezak (Infobright, Canada)

Roman Słowiński (Poznań University of Technology, Poland)
Jurek Stefanowski (Poznań University of Technology, Poland)
Alicja Wieczorkowska (PJIIT, Poland)
Xindong Wu (University of Vermont, USA)
Yiyu Yao (University Regina, Canada)
Ning Zhong (Maebashi Inst. of Tech., Japan)

Table of Contents

Session A1

Session A2

Session A3

Session A4

Session B1

Session B2