ECML 2007 PKDD
WARSAW POLAND

The 18th European Conference on Machine Learning
and
the 11th European Conference on Principles and Practice
of Knowledge Discovery in Databases

Proceedings of the

# Third International Workshop on Mining Complex Data

## MCD 2007

**September 17 and 21, 2007**

**Warsaw, Poland**

**Editors:**
*Zbigniew W. Raś*
University of North Carolina at Charlotte, USA
*Djamel Zighed*
Universite Lyon II, France
*Shusaku Tsumoto*
Shimane Medical University, Japan

# Preface

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 has been given and knowledge has now to be extracted from data getting 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, like 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 analyzes, 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 information 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 can not simply consist of the concatenation of the partial information obtained from each part of the object. The aim would rather 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 are associated the complex data). 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.

Zbigniew W. Raś
Djamel Zighed
Shusaku Tsumoto

# MCD 2007 Workshop Committee

## Workshop Chairs:

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

## Organizing Committee:

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

## Program Committee:

Aijun An (York Univ., Canada)
Elisa Bertino (Purdue Univ., USA)
Ivan Bratko (Univ. of Ljubljana, Slovenia)
Michelangelo Ceci (Univ. Bari, Italy)
Juan-Carlos Cubero (Univ of Granada, Spain)
Tapio Elomaa (Tampere Univ. of Technology, Finland)
Floriana Esposito (Univ. Bari, Italy)
Mirsad Hadzikadic (UNC-Charlotte, USA)
Howard Hamilton (Univ. Regina, Canada)
Shoji Hirano (Shimane Univ., Japan)
Mieczyslaw Klopotek (ICS PAS, Poland)
Bożena Kostek (Technical Univ. of Gdansk, Poland)
Nada Lavrac (Jozef Stefan Institute, Slovenia)
Tsau Young Lin (San Jose State Univ., USA)
Jiming Liu (Univ. of Windsor, Canada)
Hiroshi Motoda (AFOSR/AOARD & Osaka Univ., Japan)
James Peters (Univ. of Manitoba, Canada)
Jean-Marc Petit (LIRIS, INSA Lyon, France)
Vijay Raghavan (Univ. of Louisiana, USA)
Jan Rauch (Univ. of Economics, Prague, Czech Republic)
Henryk Rybiński (Warsaw Univ. of Technology, Poland)
Dominik Slezak (Infobright, Canada)
Roman Slowiński (Poznan Univ. of Technology, Poland)
Jurek Stefanowski (Poznan Univ. of Technology, Poland)
Juan Vargas (Microsoft, USA)
Alicja Wieczorkowska (PJIIT, Poland)
Xindong Wu (Univ. of Vermont, USA)
Yiyu Yao (Univ. Regina, Canada)
Ning Zhong (Maebashi Inst. of Tech., Japan)

# Table of Contents

Using Text Mining and Link Analysis for Software Mining . . . . . . . . . . . . . . . . . . 1
*Miha Grcar, Marko Grobelnik, and Dunja Mladenic*

Generalization-based Similarity for Conceptual Clustering . . . . . . . . . . . . . . . . . . . 13
*S. Ferilli, T.M.A. Basile, N. Di Mauro, M. Biba, and F. Esposito*

Finding Composite Episodes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
*Ronnie Bathoorn and Arno Siebes*

Using Secondary Knowledge to Support Decision Tree Classification of Retrospective Clinical Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
*Dympna O'Sullivan, William Elazmeh, Szymon Wilk, Ken Farion, Stan Matwin, Wojtek Michalowski, Morvarid Sehatkar*

Evaluating a Trading Rule Mining Method based on Temporal Pattern Extraction . 49
*Hidenao Abe, Satoru Hirabayashi, Miho Ohsaki, Takahira Yamaguchi*

Discriminant Feature Analysis for Music Timbre Recognition . . . . . . . . . . . . . . . . 59
*Xin Zhang, Zbigniew W. Raś*

Discovery of Frequent Graph Patterns that Consist of the Vertices with the Complex Structures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 71
*Tsubasa Yamamoto, Tomonobu Ozaki, Takenao Okawa*

Learning to Order Basic Components of Structured Complex Objects . . . . . . . . . . 83
*Donato Malerba, Michelangelo Ceci*

ARoGS: Action Rules Discovery based on Grabbing Strategy and LERS . . . . . . . 95
*Zbigniew W. Raś, Elżbieta Wyrzykowska*

Discovering Word Meanings Based on Frequent Termsets . . . . . . . . . . . . . . . . . . . 106
*Henryk Rybinski, Marzena Kryszkiewicz, Grzegorz Protaziuk, Aleksandra Kontkiewicz, Katarzyna Marcinkowska, Alexandre Delteil*

Feature Selection: Near Set Approach . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116
*James F. Peters, Sheela Ramanna*

Contextual Adaptive Clustering with Personalization . . . . . . . . . . . . . . . . . . . . . . . 128
*Krzysztof Ciesielski, Mieczysław A. Kłopotek, Sławomir Wierzchoń*

Unsupervised Grouping of Trajectory Data on Laboratory Examinations for Finding Exacerbating Cases in Chronic Diseases . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 139
*Shoji Hirano, Shusaku Tsumoto*

Improving Boosting by Exploiting Former Assumptions . . . . . . . . . . . . . . . . . . . . . 151
*Emna Bahri, Nicolas Nicoloyannis, Mondher Maddouri*

Ordinal Classification with Decision Rules . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 163
*Krzysztof Dembczyński, Wojciech Kotłowski, Roman Słowiński*

Quality of Musical Instrument Sound Identification for Various Levels of Accompanying Sounds . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 175
*Alicja Wieczorkowska, Elżbieta Kolczyńska*

Estimating Semantic Distance Between Concepts for Semantic Heterogeneous Information Retrieval . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 185
*Ahmad El Sayed, Hakim Hacid, Djamel Zighed*

Clustering Individuals in Ontologies: a Distance-based Evolutionary Approach . . 197
*Nicola Fanizzi, Claudia d'Amato, Floriana Esposito*

Data Mining of Multi-categorized Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 209
*Akinori Abe, Norihiro Hagita, Michiko Furutani, Yoshiyuki Furutani, and Rumiko Matsuoka*

POM Centric Multiaspect Data Analysis for Investigating Human Problem Solving Function . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 221
*Shinichi Motomura, Akinori Hara, Ning Zhong, Shengfu Lu*

**Author Index** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 233

# Using Text Mining and Link Analysis for Software Mining

Miha Grcar, Marko Grobelnik, and Dunja Mladenic

Jozef Stefan Institute, Dept. of Knowledge Technologies, Jamova 39, 1000 Ljubljana, Slovenia
{miha.grcar, marko.grobelnik, dunja.mladenic}@ijs.si

**Abstract.** Many data mining techniques are these days in use for ontology learning - text mining, Web mining, graph mining, link analysis, relational data mining, and so on. In the current state-of-the-art bundle there is a lack of ”software mining” techniques. This term denotes the process of extracting knowledge out of source code. In this paper we approach the software mining task with a combination of text mining and link analysis techniques. We discuss how each instance (i.e. a programming construct such as a class or a method) can be converted into a feature vector that combines the information about how the instance is interlinked with other instances, and the information about its (textual) content. The so-obtained feature vectors serve as the basis for the construction of the domain ontology with OntoGen, an existing system for semi-automatic data-driven ontology construction.

**Keywords:** software mining, text mining, link analysis, graph and network theory, feature vectors, ontologies, OntoGen, machine learning

Full article in PDF

# Generalization-based Similarity for Conceptual Clustering

S. Ferilli, T.M.A. Basile, N. Di Mauro, M. Biba, and F. Esposito

Dipartimento di Informatica
Università di Bari
via E. Orabona, 4 - 70125 Bari - Italia
{ferilli, basile, ndm, biba, esposito}@di.uniba.it

**Abstract.** Knowledge extraction represents an important issue that concerns the ability to identify valid, potentially useful and understandable patterns from large data collections. Such a task becomes more difficult if the domain of application cannot be represented by means of an attribute-value representation. Thus, a more powerful representation language, such as First-Order Logic, is necessary. Due to the complexity of handling First-Order Logic formulæ, where the presence of relations causes various portions of one description to be possibly mapped in different ways onto another description, few works presenting techniques for comparing descriptions are available in the literature for this kind of representations. Nevertheless, the ability to assess similarity between first-order descriptions has many applications, ranging from description selection to flexible matching, from instance-based learning to clustering.

This paper tackles the case of Conceptual Clustering, where a new approach to similarity evaluation, based on both syntactic and semantic features, is exploited to support the task of grouping together similar items according to their relational description. After presenting a framework for Horn Clauses (including criteria, a function and composition techniques for similarity assessment), classical clustering algorithms are exploited to carry out the grouping task. Experimental results on real-world datasets prove the effectiveness of the proposal.

Full article in PDF

# Finding Composite Episodes

Ronnie Bathoorn and Arno Siebes

Institute of Information & Computing Sciences
Utrecht University
P.O. Box 80.089, 3508TB Utrecht, The Netherlands
{ronnie,arno}@cs.uu.nl

**Abstract.** Mining frequent patterns is a major topic in data mining research, resulting in many seminal papers and algorithms on item set and episode discovery. The combination of these, called composite episodes, has attracted far less attention in literature, however. The main reason is that the well-known frequent pattern explosion is far worse for composite episodes than it is for item sets or episodes. Yet, there are many applications where composite episodes are required, e.g., in developmental biology were sequences containing gene activity sets over time are analyzed.

This paper introduces an effective algorithm for the discovery of a small, descriptive set of composite episodes. It builds on our earlier work employing MDL for finding such sets for item sets and episodes. This combination yields an optimization problem. For the best results the components descriptive power has to be balanced. Again, this problem is solved using MDL.

**keywords**: composite episodes, MDL

Full article in PDF

# Using Secondary Knowledge to Support Decision Tree Classification of Retrospective Clinical Data*

Dympna O'Sullivan[1], William Elazmeh[3], Szymon Wilk[1], Ken Farion[4], Stan Matwin[1,2], Wojtek Michalowski[1], and Morvarid Sehatkar[1]

[1] University of Ottawa, Ottawa, Canada
dympna,wilk,wojtek@telfer.uottawa.ca, mseha092@site.uottawa.ca
[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
stan@site.uottawa.ca
[3] University of Bristol, Bristol, United Kingdom
elazmah@cs.bris.ac.uk
[4] Faculty of Medicine, University of Ottawa, Ottawa, Canada
farion@cheo.on.ca

**Abstract.** Retrospective clinical data presents many challenges for data mining and machine learning. The transcription of patient records from paper charts and subsequent manipulation of data often results in high volumes of noise as well as a loss of other important information. In addition, such datasets often fail to represent expert medical knowledge and reasoning in any explicit manner. In this research we describe applying data mining methods to retrospective clinical data to build a prediction model for asthma exacerbation severity for pediatric patients in the emergency department. Difficulties in building such a model forced us to investigate alternative strategies for analyzing and processing a retrospective data. This paper describes this process together with an approach to mining retrospective clinical data by incorporating formalized external expert knowledge (*secondary knowledge sources*) into the classification task. This knowledge is used to partition the data into a number of coherent sets, where each set is explicitly described in terms of the secondary knowledge source. Instances from each set are then classified in a manner appropriate for the characteristics of the particular set. We present our methodology and outline a set of experiential results that demonstrate some advantages and some limitations of our approach.

Full article in PDF

* The support of the Natural Sciences and Engineering Research Council of Canada, the Canadian Institutes of Health Research and the Ontario Centres of Excellence is gratefully acknowledged.

# Evaluating a Trading Rule Mining Method based on Temporal Pattern Extraction

Hidenao Abe[1], Satoru Hirabayashi[2], Miho Ohsaki[3], and Takahira Yamaguchi[4]

[1] Department of Medical Informatics, Shimane University, School of Medicine
abe@med.shimane-u.ac.jp
[2] Graduate School of Science and Technology, Keio University
and_joy@ae.keio.ac.jp
[3] Faculty of Engineering, Doshisha University
mohsaki@mail.doshisha.ac.jp
[4] Faculty of Science and Technology, Keio University
yamaguti@ae.keio.ac.jp

**Abstract.** In this paper, we present an evaluation of the integrated temporal data mining environment for trading dataset from the Japanese stock market. Temporal data mining is one of key issues to get useful knowledge from databases. However, users often face difficulties during such temporal data mining process for data pre-processing method selection/construction, mining algorithm selection, and post-processing to refine the data mining process as shown in other data mining processes. To get more valuable rules for experts from a temporal data mining process, we have designed an environment which integrates temporal pattern extraction methods, rule induction methods and rule evaluation methods with visual human-system interface. After implementing this environment, we have done a case study to mine temporal rules from a Japanese stock market database for trading. The result shows the availability to find out useful trading rules based on temporal pattern extraction.

Full article in PDF

# Discriminant Feature Analysis for Music Timbre Recognition

Xin Zhang[1] and Zbigniew W. Raś[1,2]

[1] Computer Science Department, University of North Carolina, Charlotte, N.C., USA
[2] Polish-Japanese Institute of Information Technology,
02-008 Warsaw, Poland
{xinzhang, ras}@uncc.edu

**Abstract.** The high volume of digital music recordings in the internet repositories has brought a tremendous need for automatic recommendation system based on content data to help users to find their favorite music items. Music instrument identification is one of the important subtasks of content-based automatic indexing, for which the authors have developed novel new temporal features and implemented a high dimensional sound feature database with all the low-level MPEG7 descriptors as well as popular features in the literature. This paper presents development details of these new features and evaluates them among other 300 features in the database by a logistic discriminant analysis for improving music instrument identification efficiency of rule-based classifiers.

**Keywords**: Automatic Indexing, Music Information Retrieval, MPEG7, Timbre Estimation, Logistic Discriminant Analysis, Feature Selection, and Machine Learning.

Full article in PDF

# Discovery of Frequent Graph Patterns that Consist of the Vertices with the Complex Structures

Tsubasa Yamamoto[1], Tomonobu Ozaki[2], and Takenao Okawa[1]

[1] Graduate School of Engineering, Kobe University
[2] Organization of Advanced Science and Technology, Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, JAPAN
{yamamoto@cs25.scitec., tozaki@cs., ohkawa@}kobe-u.ac.jp

**Abstract.** In some applications, the data can be represented naturally in a special kind of graphs that consist of vertices holding a set of (structured) data such as item sets, sequences and so on. One of the typical examples is metabolic pathway. Metabolic pathway is represented in a graph structured data in which each vertex corresponds to an enzyme described by a set of various kinds of properties such as amino acid sequence, label and so on. We call this kind of complex graphs *multi-structured graphs*. In this paper, we propose an algorithm named FMG for mining frequent patterns in multi-structured graphs. In FMG, the external structure will be expanded by general graph mining algorithm, while the internal structure will be enumerated by some algorithm suitable for its structure. In addition, a pruning technique is introduced to exclude uninteresting patterns. The preliminary experimental results with real data show the effectiveness of the proposed algorithm.

Full article in PDF

# Learning to Order Basic Components of Structured Complex Objects

Donato Malerba and Michelangelo Ceci

Dipartimento di Informatica, Università degli Studi di Bari
via Orabona, 4 - 70126 Bari - Italy
{malerba, ceci}@di.uniba.it

**Abstract.** Determining the ordering of basic components of structured complex objects can be a crucial problem for several applications. In this paper, we investigate the problem of discovering partial or total orders among basic components by resorting to a data mining approach which acquires the domain specific knowledge from a set of training examples. The input of the learning method is the description of user-defined "chains" of basic components. The output is a logical theory that defines two predicates, $first/1$ and $succ/2$, useful for consistently reconstructing all chains in new structured complex objects. The proposed method resorts to an ILP approach in order to exploit possible relations among basic components. We describe an application of the proposed method to learning the reading order of layout components extracted from document images. Determining the reading order enables the reconstruction of a single textual element from texts associated to multiple layout components and makes both information extraction and content-based retrieval of documents more effective. Experimental results show the effectiveness of the proposed method.

Full article in PDF

# ARoGS: Action Rules Discovery based on Grabbing Strategy and LERS

Zbigniew W. Raś[1,3] and Elżbieta Wyrzykowska[2]

[1] Univ. of North Carolina, Dept. of Comp. Science, Charlotte, N.C. 28223, USA;
e-mail: ras@uncc.edu
[2] Univ. of Information Technology and Management, ul. Newelska, Warsaw, Poland;
ewyrzyko@wit.edu.pl
[3] Polish-Japanese Institute of Information Technology, ul. Koszykowa 86, 02-008 Warsaw, Poland; ras@pjwstk.edu.pl

**Abstract.** Action rules can be seen as logical terms describing knowledge about possible actions associated with objects which is hidden in a decision system. Classical strategy for discovering them from a database requires prior extraction of classification rules which next are evaluated pair by pair with a goal to build a strategy of action based on condition features in order to get a desired effect on a decision feature. An actionable strategy is represented as a term $r = [(\omega) \wedge (\alpha \rightarrow \beta)] \Rightarrow [\phi \rightarrow \psi]$, where $\omega$, $\alpha$, $\beta$, $\phi$, and $\psi$ are descriptions of objects or events. The term $r$ states that when the fixed condition $\omega$ is satisfied and the changeable behavior $(\alpha \rightarrow \beta)$ occurs in objects represented as tuples from a database so does the expectation $(\phi \rightarrow \psi)$. This paper proposes a new strategy, called *ARoGS*, for constructing action rules with the main module which resembles *LERS* [**?**]. *ARoGS* system is more simple than *DEAR* and its time complexity is also lower.

Full article in PDF

# Discovering Word Meanings Based on Frequent Termsets

Henryk Rybinski[1], Marzena Kryszkiewicz[1], Grzegorz Protaziuk[1], Aleksandra Kontkiewicz[1], Katarzyna Marcinkowska, Alexandre Delteil[2]

[1] Warsaw University of Technology,
{hrb,mkr,gprotazi}@ii.pw.edu.pl, {akontkie,
kmarcink}@elka.pw.edu.pl
[2] France Telecom R&D
alexandre.delteil@orange-ft.com

**Abstract.** Word meaning ambiguity has always been an important problem in information retrieval and extraction, as well as, text mining (documents clustering and classification). Knowledge discovery tasks such as automatic ontology building and maintenance would also profit from simple and efficient methods for discovering word meanings. The paper presents a novel text mining approach to discovering word meanings. The offered measures of their context are expressed by means of frequent termsets. The presented methods have been implemented with efficient data mining techniques. The approach is domain- and language-independent, although it requires applying part of speech tagger. The paper includes sample results obtained with the presented methods.

**Keywords:** association rules, frequent termsets, homonyms, polysemy

Full article in PDF

# Feature Selection: Near Set Approach

James F. Peters[1], Sheela Ramanna[2]*

[1]Department of Electrical and Computer Engineering,
University of Manitoba
Winnipeg, Manitoba R3T 5V6 Canada
jfpeters@ee.umanitoba.ca
[2] Department of Applied Computer Science,
University of Winnipeg,
Winnipeg, Manitoba R3B 2E9 Canada
s.ramanna@uwinnipeg.ca

**Abstract.** The problem considered in this paper is the description of objects that are, in some sense, qualitatively near each other and the selection of features useful in classifying near objects. The term *qualitatively near* is used here to mean closeness of descriptions or distinctive characteristics of objects. The solution to this twofold problem is inspired by the work of Zdzisław Pawlak during the early 1980s on the classification of objects. In working toward a solution of the problem of the classification of perceptual objects, this article introduces a near set approach to feature selection. Consideration of the nearness of objects has recently led to the introduction of what are known as near sets, an optimist's view of the approximation of sets of objects that are more or less near each other. Near set theory started with the introduction of collections of partitions (families of neighbourhoods), which provide a basis for a feature selection method based on the information content of the partitions of a set of sample objects. A byproduct of the proposed approach is a feature filtering method that eliminates features that are less useful in the classification of objects. This contribution of this article is the introduction of a near set approach to feature selection.

**Keywords**: Description, entropy, feature selection, filter, information content, nearness, near set, perception, probe function.

Full article in PDF

* We gratefully acknowledge the very helpful comments, insights and corrections concerning this paper by Andrzej Skowron and the anonymous reviewers. This research has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grants 185986 and 194376.

# Contextual Adaptive Clustering with Personalization

Krzysztof Ciesielski, Mieczysław A. Kłopotek, Sławomir Wierzchoń

Institute of Computer Science, Polish Academy of Sciences,
ul. Ordona 21, 01-237 Warszawa, Poland
kciesiel,klopotek,stw@ipipan.waw.pl

**Abstract.** We present a new method of modeling of cluster structure of a document collection and outline an approach to integrate additional knowledge we have about the document collection like prior categorization of some documents or user defined / deduced preferences in the process of personalized document map creation.

Full article in PDF

# Unsupervised Grouping of Trajectory Data on Laboratory Examinations for Finding Exacerbating Cases in Chronic Diseases

Shoji Hirano and Shusaku Tsumoto

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

**Abstract.** In this paper we present a method for finding exacerbating cases in chronic diseases based on the cluster analysis technique. Cluster analysis of time series hospital examination data is still a challenging task as it requires comparison of data involving temporal irregulariry and multidimensionalty. Our method first maps a set of time series containing different types of laboratory tests into a directed trajectory representing the time course of patient status. Then the trajectories for individual patients are compared in multiscale and grouped into similar cases. Experimental results on synthetic digit-stroke data showed that our method could yield low error rates (0.016±0.014 for classification and 0.118±0.057 for cluster rebuild). Results on the chronic hepatitis dataset demonstrated that the method could discover the groups of excacerbating cases based on the similarity of ALB-PLT trajectories.

Full article in PDF

# Improving Boosting by Exploiting Former Assumptions

Emna Bahri, Nicolas Nicoloyannis, and Mondher Maddouri

Laboratoire Eric, University Lyon 2.
5 avenue Pierre Mendes France, 69676 Bron Cedex
{e.bahri,nicolas.nicoloyannis}@univ-lyon2.fr
mondher.maddouri@fst.rnu.tn
http://eric.univ-lyon2.fr

**Abstract.** The error reduction in generalization is one of the principal motivations of research in machine learning. Thus, a great number of work is carried out on the classifiers aggregation methods in order to improve generally, by voting techniques, the performance of a single classifier. Among these methods of aggregation, we find the Boosting which is most practical thanks to the adaptive update of the distribution of the examples aiming at increasing in an exponential way the weight of the badly classified examples. However, this method is blamed because of overfitting, and the convergence speed especially with noise. In this study, we propose a new approach and modifications carried out on the algorithm of AdaBoost. We will demonstrate that it is possible to improve the performance of the Boosting, by exploiting assumptions generated with the former iterations to correct the weights of the examples. An experimental study shows the interest of this new approach, called hybrid approach.

**Keywords:** Machine learning, Data mining, Classification, Boosting, Recall, convergence

Full article in PDF

# Ordinal Classification with Decision Rules

Krzysztof Dembczyński[1], Wojciech Kotłowski[1], and Roman Słowiński[1,2]

[1] Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
{kdembczynski, wkotlowski, rslowinski}@cs.put.poznan.pl
[2] Institute for Systems Research, Polish Academy of Sciences, 01-447 Warsaw, Poland

**Abstract.** We consider the problem of ordinal classification, in which a value set of the decision attribute (output, dependent variable) is finite and ordered. This problem shares some characteristics of multi-class classification and regression, however, in contrast to the former, the order between class labels cannot be neglected, and, in the contrast to the latter, the scale of the decision attribute is not cardinal. In the paper, following the theoretical framework for ordinal classification, we introduce two algorithms based on gradient descent approach for learning ensemble of base classifiers being decision rules. The learning is performed by greedy minimization of so-called threshold loss, using a forward stagewise additive modeling. Experimental results are given that demonstrate the usefulness of the approach.

Full article in PDF

# Quality of Musical Instrument Sound Identification for Various Levels of Accompanying Sounds

Alicja Wieczorkowska[1] and Elżbieta Kolczyńska[2]

[1] Polish-Japanese Institute of Information Technology,
Koszykowa 86, 02-008 Warsaw, Poland
alicja@pjwstk.edu.pl,
[2] Agricultural University in Lublin
Akademicka 13, 20-950 Lublin, Poland
elzbieta.kolczynska@ar.lublin.pl

**Abstract.** Research on automatic identification of musical instrument sounds has already been performed through last years, but mainly for monophonic singular sounds. In this paper we work on identification of musical instrument in polyphonic environment, with added accompanying orchestral sounds for the training purposes, and using mixes of 2 instrument sounds for testing. Four instruments of definite pitch has been used. For training purposes, these sounds were mixed with orchestral recordings of various levels, diminished with respect to the original recording level. The experiments have been performed using WEKA classification software.

Full article in PDF

# Estimating Semantic Distance Between Concepts for Semantic Heterogeneous Information Retrieval

Ahmad El Sayed, Hakim Hacid, Djamel Zighed

University of Lyon 2
ERIC Laboratory- 5, avenue Pierre Mendès-France
69676 Bron cedex - France
{asayed, hhacid, dzighed}@eric.univ-lyon2.fr

**Abstract.** This paper brings two contributions in relation with the semantic heterogeneous (documents composed of texts and images) information retrieval: (1) A new context-based semantic distance measure for textual data, and (2) an IR system providing a conceptual and an automatic indexing of documents by considering their heterogeneous content using a domain specific ontology. The proposed semantic distance measure is used in order to automatically fuzzify our domain ontology. The two proposals are evaluated and very interesting results were obtained. Using our semantic distance measure, we obtained a correlation ratio of 0.89 with human judgments on a set of words pairs which led our measure to outperform all the other measures. Preliminary combination results obtained on a specialized corpus of web pages are also reported.

Full article in PDF

# Clustering Individuals in Ontologies: a Distance-based Evolutionary Approach

Nicola Fanizzi, Claudia d'Amato, Floriana Esposito

LACAM – Dipartimento di Informatica, Università degli Studi di Bari
Campus Universitario, Via Orabona 4 – 70125 Bari, Italy
{fanizzi|claudia.damato|esposito}@di.uniba.it

**Abstract.** A clustering method is presented which can be applied to semantically annotated resources in the context of ontological knowledge bases. This method can be used to discover interesting groupings of structured objects through expressed in the standard languages employed for modeling concepts in the Semantic Web. The method exploits an effective and language-independent semi-distance measure over the space of resources, that is based on their semantics w.r.t. a number of dimensions corresponding to a committee of features represented by a group of concept descriptions (discriminating features). A maximally discriminating group of features can be constructed through a feature construction method based on genetic programming. The evolutionary clustering algorithm employed is based on the notion of medoids applied to relational representations. It is able to induce a set of clusters by means of a proper fitness function based on a discernibility criterion. An experimentation with some ontologies proves the feasibility of our method.

Full article in PDF

# Data Mining of Multi-categorized Data

Akinori Abe[1),2)], Norihiro Hagita[1),3)], Michiko Furutani[1)], Yoshiyuki Furutani[1)], and Rumiko Matsuoka[1)]

1) International Research and Educational Institute for Integrated Medical Science (IREIIMS),
Tokyo Women's Medical University
8-1 Kawada-cho, Shinjuku-ku, Tokyo 162-8666 JAPAN
2) ATR Knowledge Science Laboratories
3) ATR Intelligent Robotics and Communication Laboratories
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 JAPAN
ave@ultimaVI.arc.net.my, hagita@atr.jp, {michi,yoshi,rumiko}@imcir.twmu.ac.jp

**Abstract.** At the International Research and Educational Institute for Integrated Medical Sciences (IREIIMS) project, we are collecting complete medical data sets to determine relationships between medical data and health status. Since the data include many items which will be categorized differently, it is not easy to generate useful rule sets. Sometimes rare rule combinations are ignored and thus we cannot determine the health status correctly. In this paper, we analyze the features of such complex data, point out the merit of categorized data mining and propose categorized rule generation and health status determination by using combined rule sets.

Full article in PDF

# POM Centric Multiaspect Data Analysis for Investigating Human Problem Solving Function

Shinichi Motomura[1], Akinori Hara[1], Ning Zhong[2,3], and Shengfu Lu[3]

[1] Graduate School, Maebashi Institute of Technology, Japan
[2] Department of Life Science and Informatics, Maebashi Institute of Technology, Japan
[3] The International WIC Institute, Beijing University of Technology, China
motomura@maebashi-it.org

**Abstract.** In the paper, we propose an approach of POM (peculiarity oriented mining) centric multiaspect data analysis for investigating human problem solving related functions, in which computation tasks are used as an example. The proposed approach is based on Brain Informatics (BI) methodology, which supports studies of human information processing mechanism systematically from both macro and micro points of view by combining experimental cognitive neuroscience with advanced information technology. We describe how to design systematically cognitive experiments to obtain multi-ERP data and analyze spatiotemporal peculiarity of such data. Preliminary results show the usefulness of our approach.

Full article in PDF

# Author Index

Abe, Akinori, 209
Abe, Hidenao, 49

Bahri, Emna, 151
Basile, T.M.A., 13
Bathoorn, Ronnie, 25
Biba, M., 13

Ceci, Michelangelo, 83
Ciesielski, Krzysztof, 128

d'Amato, Claudia, 197
Delteil, Alexandre, 106
Dembczyński, Krzysztof, 163
Di Mauro, N., 13

El Sayed, Ahmad, 185
Elazmeh, William, 37
Esposito, Floriana, 13, 197

Fanizzi, Nicola, 197
Farion, Ken, 37
Ferilli, S., 13
Furutani, Michiko, 209
Furutani, Yoshiyuki, 209

Grcar, Miha, 1
Grobelnik, Marko, 1

Hacid, Hakim, 185
Hagita, Norihiro, 209
Hara, Akinori, 221
Hirabayashi, Satoru, 49
Hirano, Shoji, 139

Kolczyńska, Elżbieta, 175
Kontkiewicz, Aleksandra, 106
Kotłowski, Wojciech, 163
Kryszkiewicz, Marzena, 106
Kłopotek, Mieczysław A., 128

Lu, Shengfu, 221

Maddouri, Mondher, 151
Malerba, Donato, 83
Marcinkowska, Katarzyna, 106
Matsuoka, Rumiko, 209
Matwin, Stan, 37
Michalowski, Wojtek, 37
Mladenic, Dunja, 1
Motomura, Shinichi, 221

Nicoloyannis, Nicolas, 151

O'Sullivan, Dympna, 37
Ohsaki, Miho, 49
Okawa, Takenao, 71
Ozaki, Tomonobu, 71

Peters, James F., 116
Protaziuk, Grzegorz, 106

Raś, Zbigniew W., 59, 95
Ramanna, Sheela, 116
Rybinski, Henryk, 106

Sehatkar, Morvarid, 37
Siebes, Arno, 25
Słowiński, Roman, 163

Tsumoto, Shusaku, 139

Wieczorkowska, Alicja, 175
Wierzchoń, Sławomir, 128
Wilk, Szymon, 37
Wyrzykowska, Elżbieta, 95

Yamaguchi, Takahira, 49
Yamamoto, Tsubasa, 71

Zhang, Xin, 59
Zhong, Ning, 221
Zighed, Djamel, 185