THE 18TH EUROPEAN CONFERENCE ON MACHINE LEARNING
AND
THE 11TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE
OF KNOWLEDGE DISCOVERY IN DATABASES

PROCEEDINGS OF THE INTERNATIONAL WORKSHOP ON

ROUGH SETS IN KNOWLEDGE DISCOVERY: FOUNDATIONS AND APPLICATIONS

RSKD'07

September 21, 2007

Warsaw, Poland

Editors:
Piotr Synak
Polish-Japanese Institute of Information Technology
Jakub Wróblewski
Polish-Japanese Institute of Information Technology

Workshop Organization

RSKD Workshop Chairs

Piotr Synak (Polish-Japanese Institute of Information Technology)
Jakub Wróblewski (Polish-Japanese Institute of Information Technology)

ECML/PKDD Workshop Chair

Marzena Kryszkiewicz (Warsaw University of Technology)

Workshop Program Committee

Aijun An
Jan Bazan
Shoji Hirano
Jan Komorowski
Lech Polkowski
Andrzej Skowron
Guoyin Wang
Hui Wang

Table of Contents

Classifiers from Granulated Data Sets: Concept Dependent and Layered Granulation

Piotr Artiemjew

Department of Mathematics and Computer Science
University of Warmia and Mazury
Olsztyn, Poland
artem@matman.uwm.edu.pl

Abstract. Granulation of data and the idea of a granular data set were proposed by L. Polkowski on the basis of the assumption that is at heart of all data mining techniques, i.e., that given a plausible similarity measure on objects in a data set, objects which are similar in a satisfactory degree would have also similar or even equal decision (class) values. This assumption underlies reasoning by analogy, nearest neighbors methodology, case based reasoning and rough set methods as well.
This assumption taken to an extreme implies that once a granular data system has been induced from a real data set, and decision/classification rules have been computed from it, these rules when applied to the original data should produce classification results close satisfactorily to classification results obtained on the real data with rules induced from the original non–granulated data. A number of tests performed borne out this hypothesis.
We present in this work results of experiments with real data sets and we work with the two extensions of granulation techniques presented by us in the literature so far, i.e., concept dependent granulation and layered granulation.

Full article in PDF

Indiscernibility-based Clustering of Non-Euclidean Relational Data

Shoji Hirano and Shusaku Tsumoto

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

Abstract. In this paper, we present a new method for clustering non-Euclidean relational data based on the combination of indiscernibility level and linkage algorithm. The indiscernibility level quantifies the level of global agreement for classifying two objects into the same category as indiscernible objects. Single-linkage grouping is then used to merge objects according to the indiscernibility level from bottom to top and construct the dendrogram. This scheme enables users to examine the hierarchy of data granularity and obtain the set of indiscernible objects that meets the given level of granularity. Additionally, since indiscernibility level is derived based on the binary classifications assigned independently to each object, it can be applied to non-Euclidean, asymmetric relational data. Through the clustering experiments on a synthetic dataset we demonstrate the property and usefulness of our method.

An Integrated Web-Query Classification Approach Based on Rough Sets

Ernestina Menasalvas[1], Santiago Eibe[1], Maria Valencia[1], and Pedro Sousa[2]

[1] Facultad de Informatica
Universidad Politecnica
Madrid, Spain
{emenasalvas,seibe}@fi.upm.es, mvalencia@zipi.fi.upm.es
[2] Universidad Nova de Lisboa
Lisboa, Portugal
pas@uninova.pt

Abstract. Categorization of web search queries is of increasing interest not only due to increasing the effectiveness and efficiency of returned results but also for the potential revenue in coupled applications such us targeted advertising or reformulation of the query for better results.
Nevertheless, categorization of the queries only based on the features of the query so far is challenging due to the small number of words in queries and the dynamism of sites and users requests.
We define features of the queries based on visibility and decay of the terms they contain to be integrated with a taxonomy of concepts based on the site structure. The proposed method is basically composed of two steps: firstly some queries are categorized according to the results they return and a fast classifier is built based on these results. In a second stage, the method exploits properties of the visibility of the terms in the queries and in the front page along a period, to obtain a set of attributes that are used to further cluster queries and enrich classification. Altogether they integrate an online categorization of queries to help the site decision for targeted advertising. The classifier builder is based on Rough Set techniques integrated with traditional K-means and decision trees. Results of the classification on a site serving news and using GSA as search engine is shown.

Full article in PDF

A Hierarchy Concept in Modeling of Concurrent Systems Described by Information Systems

Krzysztof Pancerz[1,2] and Zofia Matusiewicz[3]

[1] Chair of Computer Science Foundations
University of Information Technology and Management
Sucharskiego Str. 2, 35-225 Rzeszów, Poland
kpancerz@wsiz.rzeszow.pl
[2] Chair of Computer Science and Knowledge Engineering
College of Management and Public Administration
Akademicka Str. 4, 22-400 Zamość, Poland
[3] Chair of Mathematics
University of Information Technology and Management
Sucharskiego Str. 2, 35-225 Rzeszów, Poland
zmatusiewicz@wsiz.rzeszow.pl

Abstract. The paper provides a brief outline of the methodology for building suitable hierarchical models of concurrent systems described by information systems. The models have the form of hierarchical colored Petri nets. An introduction of a hierarchy concept in modeling of large systems seems to be necessary for simplification of legibility of the obtained models. Therefore, the main purpose of the hierarchy constructs is to break down the complexity of the large nets by dividing them into a number of subnets. In the proposed approach, the starting point for building a hierarchical net is an information system describing a given concurrent system. The hierarchy construct starts as early as on the level of description by building the so-called generalized information system. Such a system arises from combining selected real processes of a concurrent system into some generalized processes.

Full article in PDF

Towards Granular Computing: Classifiers Induced From Granular Structures

Lech Polkowski[1,2] and Piotr Artiemjew[2]

[1] Polish–Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
polkow@pjwstk.edu.pl
[2] Department of Mathematics and Computer Science
University of Warmia and Mazury
Olsztyn, Poland
artem@matman.uwm.edu.pl

Abstract. Granular computing as a paradigm is an area frequently studied within the Approximate Reasoning paradigm. Proposed by L.A. Zadeh granular computing has been studied within fuzzy as well as rough set approaches to uncertainty. It is manifest that both theories are immanently related to granulation as fuzzy set theory begins with fuzzy membership functions whose inverse images are prototype granules whereas rough set theory starts with indiscernibility relations whose classes are prototype, or, elementary granules.
Many authors have devoted their works to analysis of granulation of knowledge, definitions of granules, methods for combining (fusing) granules into larger objects, applications of granular structures, see, quoted in references works by A. Skowron, T.Y. Lin, Y.Y. Yao, L. Polkowski and others.
In this work, the emphasis is laid on granular decision (data) systems: they are introduced, methods of their construction with examples are pointed to, and applications are exhibited; those applications are founded on the basic although often implicit principle of data mining, viz., once a plausible for given data similarity measure is found, objects satisfactorily similar should reveal sufficiently close (or, for that matter identical) class values.
In this work, this principle is applied to granules, following the idea presented by L. Polkowski at 2005, 2006 IEEE GrC conferences, that granules built on basis of a similarity relation from a given decision system should consists of objects similar to such a degree that averaging them would lead to new objects which together would constitute a new decision system preserving to a high degree knowledge represented by the original decision system. As knowledge in rough set theory is meant as the classification ability, it seems reasonable to test knowledge content with classifiers as classifier accuracy.
This informal idea is tested in this work with some specific tools for granule construction, granular system building, and some well–tested classifiers known in literature for a few data sets from the UCI repository.
In the following sections we outline: basic ideas of rough computing, granulation of knowledge, the idea of a granular decision system and we include the results of exemplary tests with real data.

Full article in PDF

Improving Rule-Based Classifiers Induced by MODLEM by Selective Pre-processing of Imbalanced Data

Jerzy Stefanowski[1] and Szymon Wilk[1,2]

[1] Institute of Computing Science
Poznań University of Technology
ul. Piotrowo 2, 60–965 Poznań, Poland
{jerzy.stefanowski,szymon.wilk}@cs.put.poznan.pl
[2] Telfer School of Management
University of Ottawa
136 Jean-Jacques Lussier Str., K1N 6N5 Ottawa, Canada
wilk@telfer.uottawa.ca

Abstract. In the paper we discuss inducing rule-based classifiers from imbalanced data, where one class (a minority class) is under-represented in comparison to the remaining classes (majority classes). To improve the ability of a classifier to recognize this class, we propose a new selective pre-processing approach that is applied to data before inducing a rule-based classifier. The approach combines selective filtering of the majority classes with focused over-sampling of the minority class. Results of a comparative experimental study show that our approach improves sensitivity for the minority class while preserving the ability of a classifier to recognize examples from the majority classes.

Full article in PDF

Author Index